Unsupervised Skill Discovery with Empowerment

By

Andrew Levy

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Computer Science at Brown University

PROVIDENCE, RHODE ISLAND

October 2025

© Copyright 2025 Andrew Levy

This dissertation by Andrew Levy is accepted in its present form by the Department of Computer Science as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

George Konidaris, Advisor

Recommended to the Graduate Council

Date _____ _____

George Konidaris, Reader

Date _____ _____

Michael Littman, Reader

Date _____ _____

Peter Stone, Reader

Approved by the Graduate Council

Date _____ _____

Janet A. Blume, Dean of the Graduate School

This work is dedicated to my family.
To my parents, who taught me to value curiosity and hard work.
To my wife, Ali, whose love and sacrifice made this journey possible.
And to my children, Theo, Esme, and Ellis, may your lives be a joyful and unending search for truth.

Acknowledgments

This thesis would not have been possible without the support of several key individuals. I am deeply grateful to my advisor, George Konidaris, for granting me tremendous freedom over many years to fully explore the niche idea of empowerment. I am also indebted to my mother and father-in-law, Julie and Tom Garmisa. Your steadfast presence and support have been a pillar of strength for our family, making this achievement possible.

Abstract of *Unsupervised Skill Discovery with Empowerment*, by Andrew Levy, Ph.D., Brown University, Date: August 27, 2025.

Unsupervised skill discovery aims to help agents learn large skillsets at low cost. To realize this potential, an algorithm must address two challenges without human supervision: (i) how agents can learn diverse sets of policies and (ii) how agents can operate in learned internal states with low state uncertainty. Overcoming this second challenge requires addressing the two sources of state uncertainty: belief state aliasing (i.e., poor representation learning) and partial observability. Yet, existing unsupervised skill discovery approaches have not adequately addressed either of these two challenges.

The purpose of this thesis is to demonstrate that an unsupervised skill discovery algorithm based on Empowerment can address both challenges, enabling agents to build larger skillsets without human supervision. We first show that computing the empowerment of various internal states addresses the first problem of learning diverse sets of policies. We then show that maximizing empowerment addresses the second challenge of minimizing state uncertainty. Specifically, we show that (a) maximizing empowerment with respect to a representation learning function (e.g., a recurrent neural network (RNN)) can reduce state uncertainty caused by belief state aliasing, and (b) maximizing empowerment with respect to a behavior policy can reduce state uncertainty caused by partial observability.

Contents

List of Figures

CHAPTER 1

Introduction

General-purpose agents have the potential to expand human freedom. Generalist agents that are a cheap and competent source of labor can lower the costs of services we currently pay others to do, as well as automate some of the tasks we complete ourselves. Both of these outcomes would enable humans to spend less time on repetitive work and more time exploring more meaningful pursuits. But to achieve this vision, general-purpose agents must be a cheap source of labor, which means the cost of training them to develop large skillsets needs to be low.

The problem is that a low-cost framework for helping agents build large skillsets has not yet been developed. The dominant frameworks for learning skills, Behavior Cloning (BC) (Pomerleau, 1988) and Reinforcement Learning (RL) (Sutton and Barto, 1998), can help agents learn skills but are not cheap as they require extensive human supervision. For instance, behavior cloning can require labor-intensive procedures such as robot teleoperation (Levine et al., 2018; Mandlekar et al., 2018; Brohan et al., 2023; Zhao et al., 2023; Finn et al., 2017) or the collection of egocentric videos that require paying humans to wear cameras as they perform tasks (Giusti et al., 2016; Ng et al., 2020; Zheng et al., 2025b; Grauman et al., 2021). Similarly, traditional reinforcement learning, in which an agent learns a skill by maximizing some task-specific reward, can require that

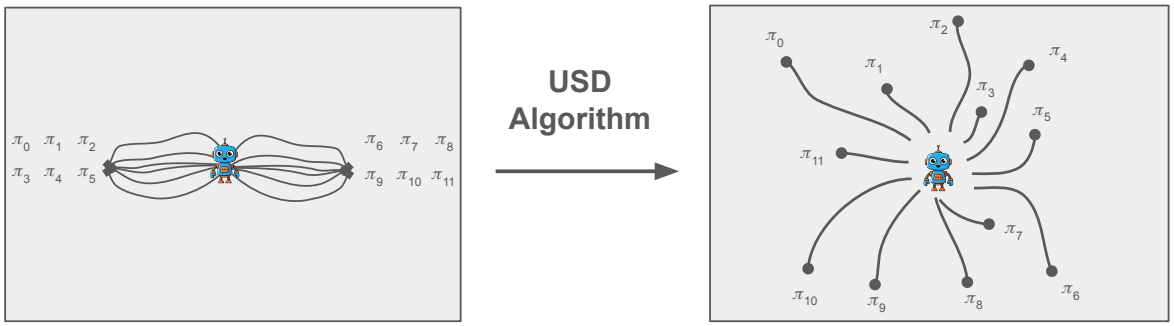
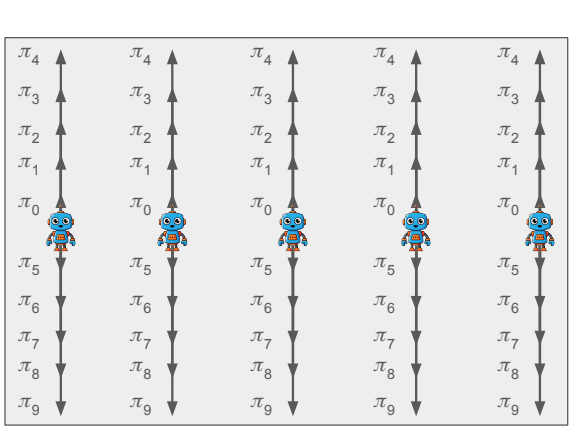


Figure 1.1: One problem unsupervised skill learning algorithms must resolve without human supervision is skillset diversity. Agents need to be able to learn the large sets of differentiated policies that can be executed from various starting representations. For instance, an agent may start with a redundant skillset such as the one shown on the left side of the figure where the 12 policies $\pi_0, \ldots, \pi_{11}$ in the skillset only target two observations. The USD algorithm should encourage the agent to learn a diverse set of skills where each skill targets a unique observation, such as the skillset on the right side of the figure.

a human both design a reward function and continually transmit rewards or some other feedback (Knox and Stone, 2009; Warnell et al., 2018; Argall et al., 2008; Bajcsy et al., 2017; Luo et al., 2025).

A potentially more practical learning framework that has emerged over the last decade is Unsupervised Skill Discovery (USD) (Gregor et al., 2016; Eysenbach et al., 2019; Sharma et al., 2020). This class of algorithms aims to lower the cost of skill discovery by removing the need for expensive human supervision. Specifically, USD algorithms aim to take an agent's reward-free and label-free history of actions and observations and extract diverse sets of skills from this cheap interaction data. Yet, to achieve the goal of learning large skillsets at lower cost, unsupervised skill discovery algorithms must address the following two problems.

The first problem is policy diversity. From various starting representations, agents must be able to learn a large set of policies, or functions that map representations to actions, such that each policy targets distinct observations not targeted by other policies in the skillset. For instance, consider the agent in the simple 2D world in Figure 1.1. At



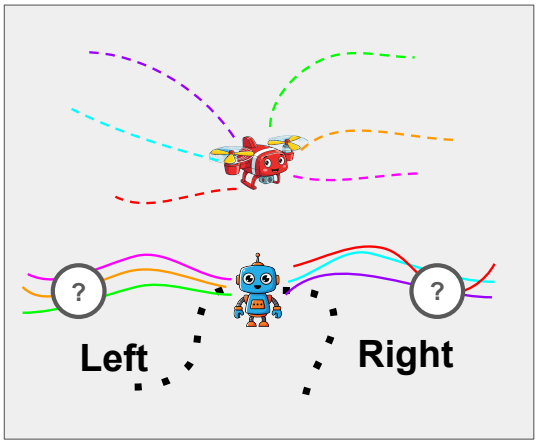


Figure 1.2: Illustration of belief state aliasing and its effect on skillset size. Belief state aliasing occurs when agents map histories that represent different underlying distributions of states (i.e., different belief states) to the same encoding. (Left) An agent in 2D world maps histories with different x positions but same y position to the same encoding. That is, the agent operates in a representation where it is blind to its x position. As a result of this aliasing, the agent's skillset is smaller as it cannot reliably target particular x positions. (Right) An agent maps two histories (dotted black lines) in which agent is told the button that turns the drone on is in different locations ("left" or "right") to the same encoding. If the agent is in this aliased representation, it will be unsure of the location of the "On" button. As a result of the aliasing, the agent in the drone setting cannot reliably target (x, y) positions for the drone as shown by the dotted trajectories.

the start of training, the agent may have a redundant skillset where various policies target similar observations. An unsupervised skill discovery algorithm must encourage the agent to discover diverse sets of policies that target distinct observations.

The second problem USD algorithms must address is state uncertainty. Agents that operate in representations with higher state uncertainty tend to have less diverse skillsets. This is because greater uncertainty produces more stochastic dynamics, which broadens the distribution of observations that each skill targets. This, in turn, may result in redundant skills that target overlapping distributions of observations. To minimize state uncertainty and thereby boost skillset size, agents need to overcome the two sources of state uncertainty: belief state aliasing and partial observability.

Belief state aliasing results from poor representation learning when agents encode his-

tories representing different underlying distributions of states (i.e., belief states (Kaelbling et al., 1998; Littman, 2009)) to the same or similar encodings. This aliasing increases state uncertainty because when an agent is in an aliased representation, the agent may be in any state in the mixture of belief state distributions. Figure 1.2 (Left) shows a simple case of belief state aliasing in the simple 2D domain described previously and its effect on the agent's skillset. In this example, the agent maps histories in which the agent has the same y position but different x positions to the same representation. As a result, when the agent is in some representation, its true state can be in one of many x positions. This reduces the number of distinct skills an agent can execute because while the agent can still reliably target specific y positions, it cannot reliably target particular (x, y) positions as its starting position can be any x coordinate. For example, a skill that moves the agent a little to the right can achieve observations that range from slightly off the west edge of the room to the eastern edge of the room.

Belief state aliasing is more common in partially observable domains where representation learning is needed. This is because agents may need the information contained in their full history of actions and observations, but using the full history as a representation is often impractical. Consider the simple partially observable setting in Figure 1.2 (Right), which consists of an agent, a drone, and two buttons. The agent can move freely around the 2D world as well as direct a drone provided that it first presses the "On" button before pressing the "Off" button. The location of the "On" button is randomly chosen each episode. The setting is partially observable as the agent's observation will only sometimes include a boolean variable indicating which button is the "On" button. In this setting, belief state aliasing can occur when the agent maps two or more histories with different signals about the location of the "On" button to the same encoding. From this aliased representation, the agent's skillset is smaller because while the agent can still move itself to various (x, y) locations, it cannot with certainty direct the drone to a particular location as shown in 1.2 (Right).

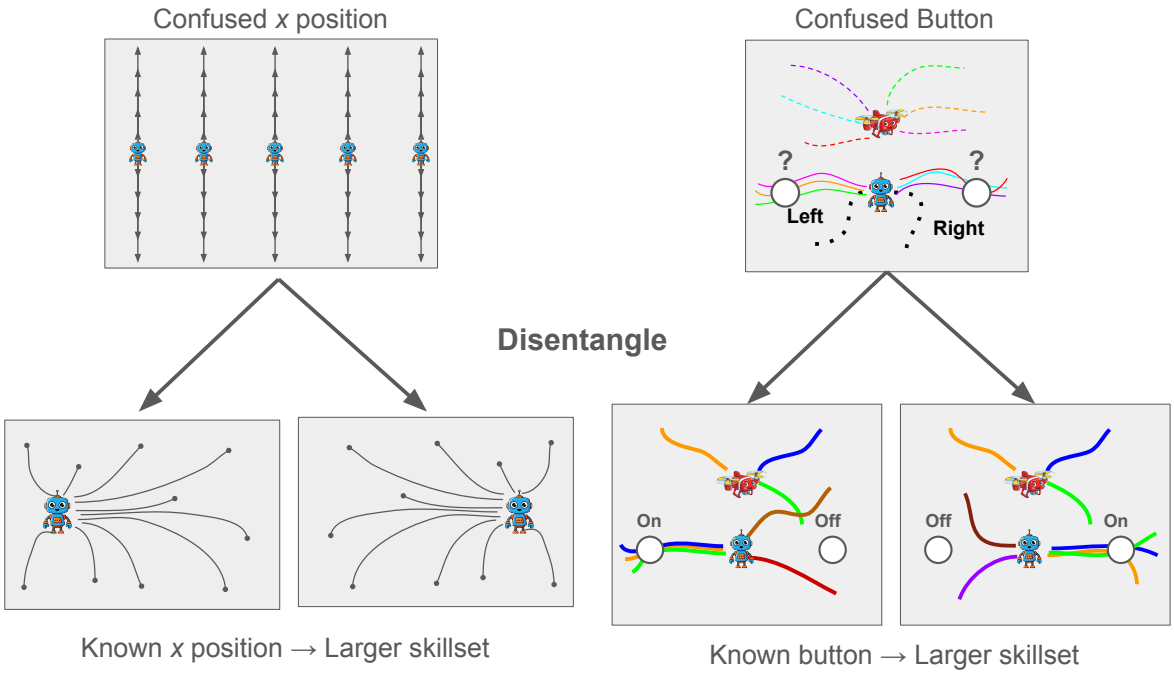


Figure 1.3: State uncertainty caused by belief state aliasing can be resolved through better representation learning. In the 2D world setting, agents can gain information about the x position of the agent by assigning different representations to histories that end in different (x, y) positions. In the drone setting, agents can gain information about the location of the button by assigning different representations to histories that signal different "On" button locations. Further, the figure illustrates how reducing state uncertainty can expand the size of an agent's skillset. The agent in the 2D world can now target particular (x, y) locations and the agent in the drone setting can now reliably move the drone to particular (x, y) locations.

The uncertainty arising from belief state aliasing can be resolved through better representation learning. Specifically, agents can assign different representations to histories (i.e., disentangle histories) that correspond to different belief states. Figure 1.3 illustrates how representation learning can be used to resolve the belief state aliasing in the settings discussed so far. In the 2D world setting, histories that end in different (x, y) positions for the agent can be assigned different representations. Similarly, in the drone setting, histories that provide different signals of the location of the "On" button can be assigned different representations. In both settings, resolving the uncertainty caused by belief state aliasing can significantly expand the agent's skillset as the agent in the 2D world can now target different x positions and the agent in the drone setting can direct the drone to

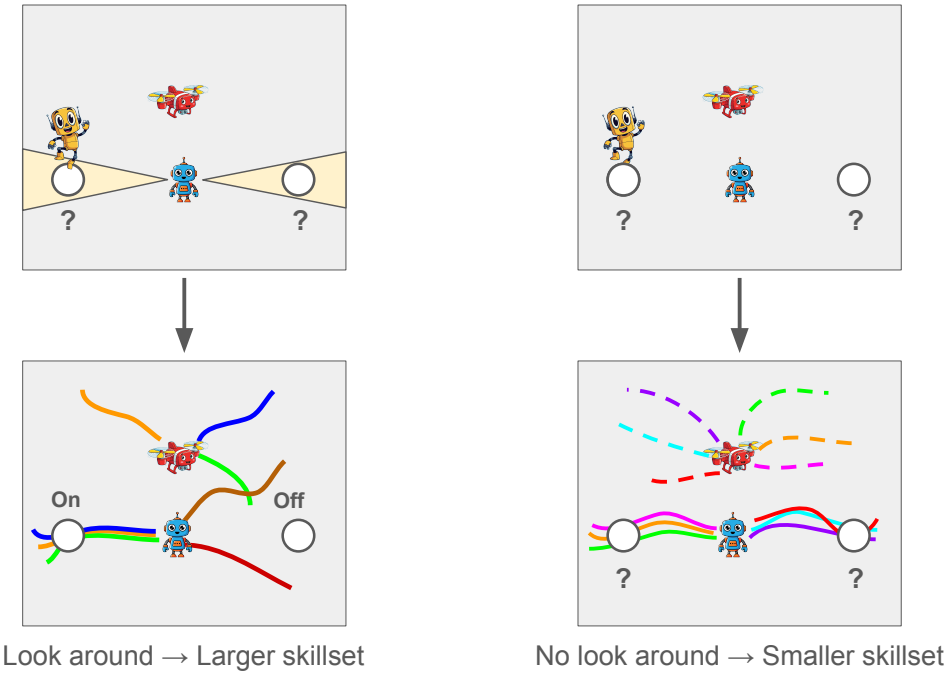


Figure 1.4: When uncertainty is caused by partial observability, agents can gain information and thereby increase their skillsets by executing information-seeking actions in the environment. In this illustration, the left side of the figure shows an agent that is at first unsure which button is the "On" button. The agent then executes the information-seeking action of looking around and sees the other agent on the Left button, which means the Left button is the "On" button. This can put the agent in a representation where there is no uncertainty about the underlying state, resulting in a larger skillset. On the other hand, the right side of the figure shows an agent choosing not to look around. As a result, the agent will continue to be uncertain about the underlying state and thus have a smaller skillset.

particular (x, y) positions with high reliability.

The second cause of state uncertainty is partial observability. Even if the agent has learned sufficient statistic representations that provide as much information about the underlying state as the agent's full history (e.g., a belief state) and thereby resolve belief state aliasing, the agent may still be uncertain about certain features of the state. For instance, consider the drone setting where the agent has not received any signal about the "On" button location. Given this history, the agent would still be uncertain about the location of the "On" button even if the agent had learned to generate sufficient statistic representations. To reduce uncertainty caused by partial observation, agents

can execute information-seeking actions, which are actions that elicit observations from the environment that provide clues about the underlying state of the environment. For instance, consider the modified drone setting in Figure 1.4 where agents now have an extra action that enables the agent to "look around". In addition, sometimes when the agent looks around, it sees another agent and this agent is always on the "On" button. In this setting, if the agent has not been given the signal of the "On" button location, the agent can potentially gain information about the underlying state and thereby expand its skillset by looking around. On the other hand, if the agent does not look around, it will remain uncertain about the "On" button and its skillset will be more limited.

Despite the impact both policy diversity and state uncertainty can have on the size of an agent's skillset, existing unsupervised skill discovery algorithms have not adequately addressed either of these two challenges. Regarding the first challenge of policy diversity, existing approaches have reported metrics related to skillset size that are often orders of magnitude smaller than the maximum possible skillset in the environment (Florensa et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018; Sharma et al., 2020; Choi et al., 2021; Strouse et al., 2022; Park et al., 2024b; Zheng et al., 2025a; Hu et al., 2024). Regarding the second challenge of state uncertainty, no unsupervised skill discovery algorithm prior to this thesis has attempted to perform representation learning in partially observable settings, making it unclear whether existing work provides a mechanism for addressing belief state aliasing in these settings. In addition, prior work has not provided a mechanism for encouraging information-seeking actions and thus existing work does not address state uncertainty arising from partial observability.

This thesis aims to develop a more complete unsupervised skill discovery algorithm that addresses both challenges in unsupervised skill discovery so that agents can build large skillsets. The foundation of our approach is an objective function, defined shortly, that measures the number of distinct skills associated with a particular tuple. The first component of the tuple is some representation from which skills will be executed. The

second component is a skillset, or a distribution of policies, that are executed from the starting representation. The third component of the tuple is a set of representation learning parameters such as an RNN. The final component is a world model that provides the distribution of observations at the next time step given the prior history of actions and observations through the current time step. The objective at the center of our approach measures the number of distinct skills in the tuple's skillset when the skillset is executed from the tuple's starting representation. In practice, skillset size is measured by sampling a large batch of policies from the skillset and concurrently executing the batch of skills from the starting representation using the agent's world model. Larger skillsets are those where skills tend to terminate in more distinct observations.

To address the two key challenges in unsupervised skill discovery, this thesis develops a simple approach. To address the first challenge of learning diverse skillsets from some starting representation, we simply maximize the objective with respect to the set of policies. This encourages more differentiated skillsets where each skill specializes in targeting specific observations. To address the part of the second challenge related to state uncertainty arising from belief state aliasing, we maximize the objective with respect to the parameters of the RNN or other representation learning function. If there is belief state aliasing depressing the size of an agent's skillset, this should encourage the RNN to disentangle those histories representing different belief states. Then to address the problem of state uncertainty arising from partial observability, we treat this objective as a reward and train a behavior policy to maximize this reward. If there are features of the state that the agent is uncertain about and are depressing the size of an agent's skillset, this objective should encourage policies that elicit information from the environment about these features.

The objective we use at the foundation of our approach is the mutual information between skills and observations. This objective is principled as it provides a way to measure the number of distinct skills that can be executed from some representation.

The objective is also unsupervised because measuring the mutual information does not require expensive supervision in the form of rewards or action labels. Maximizing mutual information in practice does require a world model, but this world model can be trained in an unsupervised manner from cheap interaction data. Note, however, that the experiments in this thesis will make the assumption that a world model has been provided to the agent. This is not a scalable assumption, and Chapter 6 discusses how different types of world models can be learned.

Our approach to unsupervised skill discovery can also be expressed in terms of empowerment. Empowerment is the maximum mutual information between skills and observations with respect to the agent's set of skills and thereby represents the maximum number of skills that can be executed from some representation (Klyubin et al., 2005; Salge et al., 2013; Jung et al., 2012; Mohamed and Rezende, 2015; Gregor et al., 2016). In terms of empowerment, our approach addresses the first challenge of learning diverse sets of policies by computing the empowerment of different representations and the second challenge of state uncertainty by maximizing empowerment with respect to the agent's representation learning function and behavior policy.

1.1 Contributions

This section details the three contributions of the thesis and describes how the contributions relate to prior work.

1. **Tighter variational lower bound on the mutual information between skills and observations**

   We introduce a tighter variational lower bound on the mutual information between skills and observations relative to prior empowerment-based skill discovery methods (Gregor et al., 2016; Florensa et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018). The tighter bound results from adding a conditioning variable to the variational

posterior distribution in the mutual information term that represents the skillset under consideration. The tighter variational lower bound enables agents to better measure the diversity of a candidate skillset, which in turn makes it easier to discover more diverse skillsets. To tractably maximize this mutual information lower bound with respect to the agent's skillset, we also introduce a new actor-critic architecture.

We provide a theorem proving that our variational mutual information objective provides a tighter bound than prior work. We also provide empirical support that our improvements to mutual information optimization enable agents to learn skillsets that are often orders of magnitude larger than prior work.

Context: Maximizing the mutual information between skills and observations to address the first challenge of learning diverse skillsets has been an approach that has been tried previously. Indeed, the original unsupervised skill discovery algorithms (Gregor et al., 2016; Florensa et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018; Sharma et al., 2020) all tried to learn skills by maximizing some mutual information between skills and observations. After these algorithms were unable to learn large skillsets, another class of algorithms emerged (Zheng et al., 2025a; Strouse et al., 2022; He et al., 2025; Hu et al., 2024, 2025; Park et al., 2022, 2023b, 2024b) arguing that changes needed to be made to the mutual information objective to boost exploration such as (i) adding regularization terms (Strouse et al., 2022; He et al., 2025; Zheng et al., 2025a), (ii) applying skill factorization where each dimension of the skill makes changes to a particular state dimension (Hu et al., 2024, 2025; Carr et al., 2025), or (iii) swapping the entropy terms in the mutual information objective for other terms (e.g., contrastive losses) (Laskin et al., 2022; Zheng et al., 2025a; Warde-Farley et al., 2019; Achiam et al., 2018; Zhang et al., 2021; Choi et al., 2021; Campos et al., 2020). Others argued that mutual information is not capable of learning large skillsets in continuous settings and thus some other diversity metric was needed (Park et al., 2022, 2023b, 2024b). Yet, recent work has

demonstrated that many of these changes do not produce significantly larger skillsets (Levy et al., 2025). Our first contribution provides strong evidence that the main problem with the original mutual information objective was that prior algorithms were maximizing a loose lower bound on the objective, meaning that agents were often significantly underestimating the diversity of skillsets when evaluating different sets of skills, in turn making it difficult to learn diverse skillsets.

2. **Maximizing empowerment with respect to a representation learning function discourages belief state aliasing**

   The second contribution is to show that maximizing empowerment with respect to an RNN (or other representation learning function) encourages the RNN to provide more information about the underlying state. That is, maximizing empowerment encourages the RNN to resolve belief state aliasing and disentangle histories that correspond to different underlying distributions of states.

   We support this claim with both theoretical and empirical results. We provide two theorems to support this contribution. The first theorem shows that the average empowerment produced by an RNN is upper bounded by the average empowerment produced by a representation learning function that outputs sufficient statistic representations. Given that sufficient statistic representations provide as much information about the underlying state as the agent's history and thereby have no belief state aliasing, this theorem provides evidence that empowerment can help overcome the problem of state uncertainty caused by poor representation learning. Yet this theorem does not provide evidence that empowerment can provide a dense reward signal that continually encourages agents to disentangle histories associated with different belief states. In practice, it may require numerous updates to disentangle histories and thus a dense reward signal is important. The purpose of the second theorem is to show that empowerment is indeed a dense reward signal for discouraging belief state aliasing. The theorem shows that if the agent is considering

two RNNs and one RNN provides more information about the underlying state than the other (i.e., one RNN reduces state uncertainty on average more than the other), the RNN providing more information will produce an average empowerment at least as large as the other RNN (and larger under some easily achievable conditions) and thus be encouraged by the empowerment objective.

We also support this claim with empirical results in (a) low-dimensional fully observable environments, (b) high-dimensional fully observable environments, and (c) partially observable environments. In all of these settings, we demonstrate that our approach can jointly learn both large sets of skills and the representations the skills are conditioned on. Indeed, our approach is the first unsupervised skill discovery approach to learn both skills and representations in partially observable environments.

Context: This idea of using empowerment for representation learning builds on the work of Klyubin et al. (2008) and Capdepuy (2011) who also implemented similar ideas but in more limited contexts. Both used a simpler version of empowerment that maximizes the mutual information between a single primitive action and the succeeding observation. This version of mutual information does not enable agents to learn longer-horizon skills, which are important for downstream tasks. Also, the single primitive action version of empowerment is difficult to use as a signal for representation learning because often different states do not have significantly different one-step empowerment but may have different longer-horizon empowerment. In addition, both works applied their approaches to fully observable grid-world or simpler settings, whereas we apply our approach to continuous and partially observable settings. We also provide the theorems mentioned above that prove that empowerment encourages information seeking by representation learning. Other works that used empowerment for representation learning include Bharadhwaj et al. (2022), which also only employed single-step empowerment to fully observable

settings. Their approach also incorporates task rewards into the objective so it is not fully unsupervised representation learning. In addition, related to our work are multi-step inverse methods (Lamb et al., 2023; Rudolph et al., 2024) that use an objective similar to the mutual information between single-step actions and downstream observations for representation learning. However, these results also do not learn skills and are only applied to fully observable settings.

3. **Maximizing empowerment with respect to a behavior policy encourages information seeking by environment interaction**

   The third contribution of the thesis is to show that maximizing empowerment with respect to a behavior policy encourages agents to seek information through environment interaction. That is, maximizing an empowerment reward can be a solution to the problem of state uncertainty caused by partial observability. If the following are true: (a) there are some features of the state that the agent is uncertain about, (b) this uncertainty is reducing the size of an agent's skillset, and (c) the agent can take actions in the environment to elicit information about the unknown features, then maximizing an empowerment reward will encourage these information-seeking actions and the agent may be able to reduce some uncertainty caused by partial observability.

   We support this contribution with both theoretical and empirical results. We provide one theorem that shows that if the agent is considering two behavior policies and one policy terminates in a representation that provides more information about the underlying state than the terminating representation achieved by the other policy, then the policy that provides more information will produce larger empowerment on average and thus will be encouraged by the empowerment objective. We also provide empirical results in a few simple domains showing that agents prefer to execute information-seeking policies.

Context: To our knowledge, this thesis provides the first results showing empowerment encourages information-seeking actions. Many prior works have also performed empowerment maximization but for different applications such as stabilization (Jung et al., 2012; Karl et al., 2017; Zhao et al., 2020) or predator avoidance (Mohamed and Rezende, 2015). These applications are critical for agents but information seeking is notable because it can be a long-term driver of skillset growth. In realistic settings, there will often be features of the state that the agent is not only uncertain about but that also depress the size of an agent's skillset. Our result that empowerment encourages information seeking means an empowerment-maximizing agent will continually be encouraged to learn about these unknown state features and thereby build its skillset over time.

Also related to our contributions on information seeking are algorithms that directly encourage information seeking such as those that minimize belief state uncertainty (Rhinehart et al., 2021; Aoki et al., 2011; Williams, 2007; Kreucher et al., 2005). The key advantage of using empowerment for information seeking instead of trying to directly minimize state uncertainty is that there is not always a strong correlation between information and empowerment. Sometimes small bits of information can have large impacts on an agent's skillset. For example, consider the agent in the drone setting that is not sure which one of two buttons is the "On" button. Conversely, large bits of information can have no impact on an agent's skillset (e.g., the number of blades of grass in some area or the exact pattern of a QR code). An agent that is trying to directly maximize information may exert effort trying to extract these bits of information and thus be very inefficient at growing its skillset. On the other hand, agents that maximize empowerment will be encouraged to focus on the bits of information that most impact the size of their skillset.

CHAPTER 2

Background and Related Work

This chapter provides the relevant background and related work information needed to understand the contributions of the thesis. The background sections describe how environments and agents are modeled in this thesis, define empowerment, and review relevant properties of mutual information, among other topics. The related work sections discuss prior work in the main areas within artificial intelligence research covered by this thesis: unsupervised reinforcement learning, representation learning in partially observable settings, and information seeking.

2.1 Modeling the Environment

In this thesis, we will model the environment using a discrete-time Controlled Hidden Markov Process (CHMP), which is a probabilistic graphical model defined by the tuple $(\mathcal{S}, \mathcal{A}, \Omega, p(s_0), p(s_{t+1}|s_t, a_t), p(o_t|s_t))$. $\mathcal{S}$ is the state space with states $s \in \mathcal{S}$; $\mathcal{A}$ is the action space with actions $a \in \mathcal{A}$; Ω is the observation space with observations $o \in \Omega$; $p(s_0)$ 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 distribution over observations given the current state. A CHMP setting is thus equivalent to a Partially Observed Markov Decision Process (POMDP) (Åström, 1965; Kaelbling

et al., 1998) without a reward function. In this thesis, we will consider CHMPs with both (a) Markov and (b) non-Markov observations. In the Markov variant, the distribution over underlying states given the latest observation is conditionally independent of the prior history of actions and observations: $p(s_t|o_t) = p(s_t|a_0, o_1, \ldots, a_{t-1}, o_t)$. On the other hand, in the non-Markov variant, there is some history of actions and observations in which $p(s_t|o_t) \neq p(s_t|a_0, o_1, \ldots, a_{t-1}, o_t)$. Note that in this thesis, we will use the environment characterizations of "non-Markov" and "partially observable" interchangeably.

2.2 Modeling the Agent

The goal of the thesis is to train an agent to learn three particular functions. The first function is a recurrent neural network $f_\eta : \mathcal{C} \times \mathcal{A} \times \mathcal{O} \to \mathcal{C}$ that maps learned representations $c_t \in \mathcal{C}$, actions a_t, and observations o_{t+1} to the next learned representation c_{t+1}. The goal of this representation learning is to learn sufficient statistic representations $x_t \in \mathcal{X}$, which are representations that provide distributions over the underlying state that are conditionally independent of the prior history of actions and observations: $p(s_t|x_t) = p(s_t|a_0, o_1, \ldots, a_{t-1}, o_t)$.

The second function is a skillset function $f_\lambda : \mathcal{C} \to p(\pi)$ that maps a learned representation c_t to a distribution over policies $p(\pi)$, in which each policy $\pi : \mathcal{C} \to \mathcal{A}$ maps learned representations to primitive actions. Specifically, in this thesis, the skillset function will output a skill-conditioned policy $\pi_z : \mathcal{C} \times \mathcal{Z} \to \mathcal{A}$, which maps learned representations c_t and skills $z \in \mathcal{Z}$ to primitive actions. For instance, in our implementation, f_λ will output the parameters (i.e., the weights and biases) of a skill-conditioned policy neural network. This skill-conditioned policy represents a distribution over policies $p(\pi)$ because the combination of (i) a skill z sampled from some fixed distribution over skills $p(z)$ and (ii) the skill-conditioned policy π_z, produces some policy π mapping learned representations to actions. This skillset function will be trained to learn all the distinct skills an agent could execute from some learned representation c_t. A distinct skill in the context of

empowerment will mean a policy that terminates in observations that are not targeted by other policies in the $p(\pi)$ skillset. Note that the skillset function will be trained in simulation using the agent's world model.

The third function is a behavior policy $\pi_b : \mathcal{C} \rightarrow \mathcal{A}$ that maps learned representations c_t to primitive actions a_t. In contrast to the skill-conditioned policies the agent learns, the agent will use this behavior policy to actually interact in the environment. Specifically, this policy will be encouraged to seek information, meaning that the agent will be encouraged to take actions to achieve learned representations with less uncertainty about the underlying state.

2.3 Mutual Information and Empowerment

Given two random variables, A and B, defined by two distributions including (i) a source distribution $p(a)$ and (ii) a channel distribution $p(b|a)$, the mutual information between A and B, $I(A;B)$, measures the rate at which samples of A that can be sent across the channel $p(b|a)$ producing a sample b that can then be decoded back to the original sample of a with arbitrarily low error for each use of the channel. That is, mutual information captures the number of distinct values of $a \sim A$ that can be reliably sent across the channel. $I(A;B)$ is defined

$$I(A;B) = H(B) - H(B|A) \tag{2.1}$$

$$= H(A) - H(A|B) \tag{2.2}$$

$$= \mathbb{E}_{a\sim p(a), b\sim p(b|a)}[\log p(a) - \log p(a|b)], \tag{2.3}$$

where $H(C)$ of some random variable C is the Shannon entropy (Shannon, 1948): $H(C) = \mathbb{E}_{c\sim p(c)}[-\log p(c)]$. Per line 2.2, mutual information is higher when (i) the entropy $H(A)$ is higher, which can occur when there are more distinct samples of a sent across the channel, and/or (ii) $H(A|B)$ is lower, which means the samples of a are more distinct in terms of

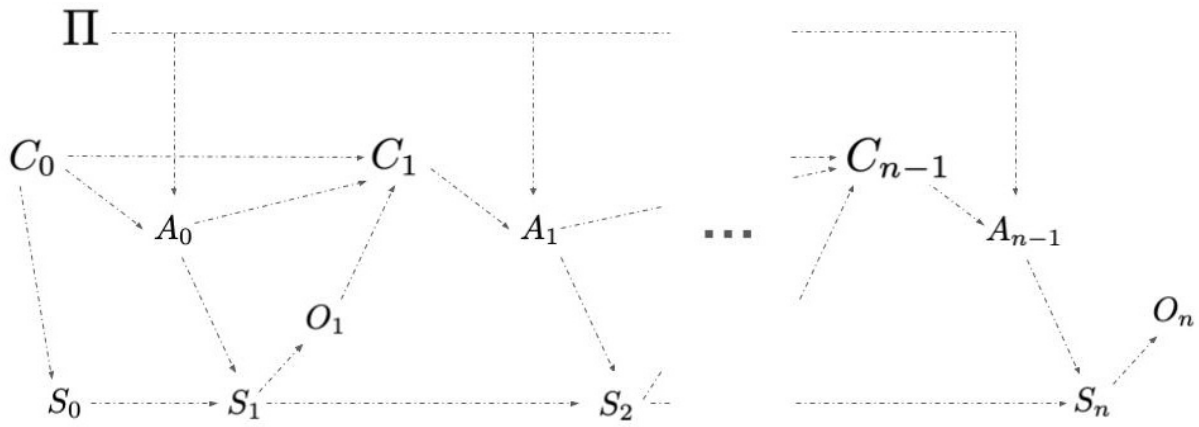


Figure 2.1: Probabilistic graphical model for executing skills π from some starting learned representation c_0. The initial state random variable S_0 is sampled from the distribution of underlying states $p(s_0|c_0)$. Future states are sampled from the state transition dynamics $p(s_{t+1}|s_t, a_t)$. Actions are sampled from the policy π. Learned representations are sampled from some distribution $p(c_{t+1}|c_t, a_t, o_{t+1})$. Observations are sampled from the observation distribution $p(o_t|s_t)$.

the samples b that they target. As a result of the entropy terms, mutual information is measured in logarithmic units (e.g., bits or nats).

To measure the number of distinct skills contained in some skillset defined by $p(\pi)$ when the skillset is executed from some learned representation c_0, the mutual information $I(\Pi; O_n|c_0)$ between a skill random variable Π and skill-terminating observation random variable O_n can be used. The source distribution is the distribution over policies $p(\pi|c_0)$. The channel distribution $p(o_n|c_0, \pi)$ samples skill-terminating observations from the probabilistic graphical model shown in Figure 2.1, where (i) the initial state $p(s_0|c_0)$ is sampled from the underlying distribution over states, (ii) future states are sampled from the state transition dynamics $p(s_{t+1}|s_t, a_t)$, (iii) primitive actions are generated from the policy π, (iv) learned representations are generated by some distribution $p(c_{t+1}|c_t, a_t, o_{t+1})$, and (v) observations are sampled from the observation distribution $p(o_t|s_t)$. This mutual

information is defined

$$I(\Pi; O_n|c_0) = H(O_n|c_0) - H(O_n|c_0, \Pi) \tag{2.4}$$

$$= H(\Pi|c_0) - H(\Pi|c_0, O_n) \tag{2.5}$$

$$= \mathbb{E}_{\pi\sim p(\pi|c_0), o_n\sim p(o_n|c_0,\pi)}[\log p(\pi|c_0, o_n) - \log p(\pi|c_0)]. \tag{2.6}$$

Thus, per line 2.4, the number of distinct skills grows when the entropy $H(O_n|c_0)$ increases, which can occur when the number of observations o_n a skillset targets increases, and/or the conditional entropy $H(O_n|c_0, \Pi)$ decreases, which can occur when different skills target different groupings of skill-terminating observations o_n. Similarly, per line 2.5, the number of distinct skills grows when $H(\Pi|c_0)$ grows, which can occur when a more entropic distribution over policies is executed, and/or $H(\Pi|c_0, O_n)$ decreases, which can occur when different skills target different groupings of skill-terminating observations.

The empowerment of a representation measures the maximum mutual information between skills and observations with respect to the skillset $p(\pi|c_0)$:

$$\mathcal{E}(c_0) = \max_{p(\pi|c_0)} I(\Pi; O_n|c_0). \tag{2.7}$$

That is, empowerment is the largest number of distinct skills that can be executed from some representation c_0.

2.4 Relevant Properties of Mutual Information

We will use the following four properties of mutual information in this thesis:

1. Data-Processing Inequality:

$$I(A; B) \geq I(A; C) \text{ if } c = f(b) \tag{2.8}$$

2. Convexity w.r.t. Channel Distribution:

$$I(A;C) \leq \mathbb{E}_{b\sim p(b)}[I(A;C|b)] \text{ if } p(c|a) = \mathbb{E}_{b\sim p(b)}[p(c|a,b)] \tag{2.9}$$

3. Variational Lower Bound:

$$I(A;B) \geq \mathbb{E}_{a\sim p(a),b\sim p(b|a)}[\log q_\psi(a|b) - \log p(a)] \tag{2.10}$$

4. Chain Rule:

$$I(A;B_0,B_1,\ldots,B_{n-1}) = I(A;B_0) + I(A;B_1,\ldots,B_{n-1}|B_0) \tag{2.11}$$

The Data Processing Inequality states that the mutual information between a random variable A and another random variable B is at least as large as the mutual information between A and a third random variable C that is some function of B (see Ch.2 in Cover and Thomas (2006) for proof). For instance, one implication of this property is that the mutual information between skills and observations is upper bounded by the mutual information between skills and states, assuming observations o_t are generated using some function of states s_t. In other words, by using observations instead of states, the agent may learn fewer skills because multiple states s_t may map to the same observation o_t.

The second property is the convexity of mutual information with respect to the channel distribution. That is, if the channel $p(c|a)$, between two random variables A and C is a weighted average of channels $p(c|a,b)$ that involve a third random variable B (i.e., $p(c|a) = \mathbb{E}_{b\sim p(b)}[p(c|a,b)]$), then the mutual information between A and C, $I(A;C)$, is less than or equal to the weighted average of the mutual information between A and C when conditioned on values of b (see chapter 2 in Cover and Thomas (2006)). The intuition here is that if the channel distribution is actually a noisy mixture of channels, on average more distinct samples of a can be sent over the channel if the exact channel was known. We

will use this property in Chapters 4 and 5 to prove that using empowerment as a reward function encourages information seeking through representation learning and environment interaction.

The third property shows that if the posterior $p(a|b)$ is replaced with a different variational distribution $q_\psi(a|b)$ with variational parameters ψ, then the variational mutual information term $I^V(A;B) = \mathbb{E}_{a\sim p(a),b\sim p(b|a}[\log q_\psi(a|b) - \log p(a)]$ is a lower bound of the original mutual information $I(A;B)$ (see Barber and Agakov (2003) for proof). The looseness of the bound $I(A;B) - I^V(A;B) = \mathbb{E}_{b\sim p(b)}[D_{KL}(p(a|b)||q_\psi(a|b))]$ depends on an average KL divergence between the true and variational posteriors. That is, if $q_\psi(a|b) \approx p(a|b)$, the bound will be tight but otherwise will be loose. Both prior work and the work in this thesis will use a variational lower bound on the mutual information between skills and observations because the true mutual information in continuous domains is typically intractable to compute.

The fourth property is the chain rule for mutual information, which states that the mutual information between a random variable A and a sequence of n random variables $B_0, \ldots, B_{n-1}$ can be broken down into the sum of (i) the mutual information between A and the first B_0 in the sequence, $I(A;B_0)$, and (ii) the mutual information between A and the remaining B_i terms conditioned on B_0, $I(A;B_1,\ldots,B_{n-1}|B_0)$. We will use this recursive property of mutual information in Chapter 5 to compute a longer horizon version of empowerment using only shorter horizon mutual information terms.

2.5 Markov Decision Processes

This thesis will also involve solving a finite horizon Markov Decision Process (MDP) (Puterman, 1994), which is a problem setting defined by the tuple $(\mathcal{X}, \mathcal{A}, p(x_{t+1}|x_t, a_t), R)$, in which $x_t \in \mathcal{X}$ are Markov representations (not necessarily the underlying state s_t), $a_t \in \mathcal{A}$ are primitive actions, $p(x_{t+1}|x_t, a_t)$ are the Markov transition dynamics, and

$R : \mathcal{X} \to \mathbb{R}$ is the reward for achieving a particular representation x_t. The objective in a finite horizon MDP is to learn a policy $\pi : \mathcal{X} \to \mathcal{A}$ that maximizes expected, cumulative reward: $\max_\pi \mathbb{E}_\pi[\sum_{t=1}^{T} R(x_t)]$. Finite horizon MDPs can be solved using dynamic programming (Bertsekas, 2001) and reinforcement learning methods (Sutton and Barto, 1998). In this thesis, we will implement a finite horizon MDP to maximize empowerment with respect to a behavior policy to encourage information-seeking actions.

2.6 Related Work: Unsupervised Reinforcement Learning

The algorithms contained in this thesis are a part of a broader class of algorithms known as Unsupervised Reinforcement Learning (URL). To decrease the cost of designing and transmitting rewards, URL algorithms implement a pretraining phase where agents, without rewards, try to learn components that can be useful for a downstream RL task when a reward is introduced (Agarwal et al., 2025a). Most often, these components are either (i) skills that target diverse observations (i.e., unsupervised skill discovery algorithms), (ii) representations that serve as input to policies, and (iii) statistics of policies that describe future state visitation distributions (e.g., the successor measure).

2.6.1 Skills (Unsupervised Skill Discovery)

Unsupervised skill discovery algorithms generally fall into two categories: (i) Goal-Conditioned Reinforcement Learning (GCRL) and (ii) Mutual Information-based Skill Learning.

GCRL algorithms learn policies that target particular observations (Kaelbling, 1993; Andrychowicz et al., 2017; Levy et al., 2019; Durugkar et al., 2021; Ma et al., 2022; Chuck et al., 2025; Sikchi et al., 2024; Agarwal et al., 2023; Park et al., 2023a; McClinton et al., 2021; Nair et al., 2018; Nasiriany et al., 2019; Eysenbach et al., 2021). During pretraining,

all observations can be used as potential goals for the agent to learn to achieve. Then downstream GCRL tasks, in which the agent is asked to achieve a particular subset of goals, can potentially be completed in a zero-shot manner. Two problems with unsupervised GCRL algorithms are that (i) they may require that the agent know the distribution of reachable goal observations and (ii) goal-conditioned policies can be difficult to learn in stochastic settings, where specific observations cannot be achieved with high probability. In this thesis, we present unsupervised skill discovery algorithms that can overcome both of these limitations.

Mutual information-based skill learning aims to learn skillsets that maximize some mutual information between skills and observations (Florensa et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018; Sharma et al., 2020; Choi et al., 2021; Strouse et al., 2022; Park et al., 2024b; Zheng et al., 2025a; Hu et al., 2024; Baumli et al., 2021; Campos et al., 2020; Laskin et al., 2022; Tiomkin et al., 2025). As noted in section 2.3, maximizing mutual information with respect to a skill-conditioned policy encourages each of the skills in the skill-conditioned policy to target a distinct skill-terminating observation. Learning skills by maximizing mutual information can enable zero-shot performance in downstream tasks. Given some goal observation o_n, the posterior distribution that many algorithms learn, $q(z|o_n)$, can be sampled to determine a skill to achieve the provided goal observation in a zero-shot manner.

The work in this thesis improves on many existing mutual information-based skill learning algorithms in three ways. First, we maximize a tighter variational lower bound on mutual information relative to prior work (Gregor et al., 2016; Eysenbach et al., 2019), making it easier to discover more diverse skillsets. Second, by maximizing mutual information with respect to a representation learning function (e.g., an RNN), this thesis provides algorithms for not only learning skills but also representations. This enables our approach to address the issue of belief aliasing and thereby learn larger skillsets in partially observable settings. Third, by maximizing empowerment with respect to a behavior policy,

this thesis provides algorithms that encourage agents to *achieve* learned representations with less underlying state uncertainty (i.e., encourages information seeking actions). This additional reduction in state uncertainty enables agents to attain larger skillsets than agents that only maximize mutual information with respect to a skill-conditioned policy.

2.6.2 Representations

Another segment of Unsupervised RL algorithms uses a pretraining stage to learn compressed representations of observations that seek to only contain the information needed for control (Ghosh et al., 2019; Ma et al., 2023; Islam et al., 2022; Lamb et al., 2023; Levine et al., 2024; Rudolph et al., 2024). Some of these algorithms learn representations by training inverse models that try to predict some initial primitive action given some future observation (Islam et al., 2022; Lamb et al., 2023; Levine et al., 2024; Rudolph et al., 2024). This approach is similar to a variant of our approach of maximizing mutual information with respect to representation learning parameters. However, in this approach the mutual information is between a single primitive action and a future observation so these approaches do not learn temporally-extended, closed-loop skills, which can be helpful for downstream tasks. In addition, the experiments for these methods have been limited to Markov settings, whereas we show results where agents learn representations in partially observable settings.

2.6.3 Policy Statistics

Another segment of unsupervised reinforcement learning uses a pretraining stage to compute particular statistics of policies. In successor feature approaches, these statistics are cumulative state features (Dayan, 1993; Barreto et al., 2017; Park et al., 2024a; Zhu et al., 2025). In other approaches, the statistic is the successor measure, which provides the discounted state visitation distribution achieved by a policy (Touati and Ollivier, 2021; Touati et al., 2023; Agarwal et al., 2025b). These statistics can then be used to search

for a policy that maximizes reward in a zero-shot manner. A key challenge with these approaches is that they assume Markov observations. This thesis provides an unsupervised RL algorithm that can learn Markov representations in partially observable settings as well as skills conditioned on those representations.

2.7 Related Work: Representation Learning in Partially Observable Settings

Also related to the work in this thesis is a class of algorithms that learn representations in partially observable settings (Littman and Sutton, 2001; Lin and Mitchell, 1993; Schmidhuber, 1991, 1990; Bakker, 2001; Hausknecht and Stone, 2015; Ni et al., 2022; Wierstra et al., 2007; Heess et al., 2015; Hafner et al., 2019b; Allen et al., 2024). A key difference from our work is that many of these approaches rely on manually-specified, task-specific rewards to provide a signal for learning representations, which in practice can be costly to implement and/or overly sparse. Our approach uses the mutual information between skills and observations, which is unsupervised and task-invariant. Also, in contrast to the algorithms presented in this thesis, these approaches do not jointly learn large skillsets while performing representation learning.

2.8 Related Work: Information Seeking

There have been many prior works that have developed algorithms to encourage information seeking (Aoki et al., 2011; Friston, 2009; Friston et al., 2016; Rhinehart et al., 2021; Kreucher et al., 2005; Williams, 2007; Berseth et al., 2021; Fountas et al., 2020), which is sometimes referred to as active state estimation in the sensor management literature and minimizing surprise in the intrinsic motivation literature. The typical objective these works use for decision making is to minimize the entropy of the agent's state uncertainty $H(S_{t+1}|O_{\leq t+1}, A_{\leq t})$ or maximize information gain $I(O_{t+1}; S_{t+1}|O_{\leq t}, A_{\leq t})$,

which is the difference between the uncertainty with respect to the underlying state before (i.e., $H(S_{t+1}|O_{\leq t}, A_{\leq t})$) and after ($H(S_{t+1}|O_{\leq t+1}, A_{\leq t})$) the next observation. These objectives directly reward the agent for lower uncertainty. Another popular objective is to minimize the entropy of the state visitation distribution $H(d^\pi(s))$, where the state visitation distribution is $d^\pi(s) = \frac{1}{T}(\sum_{t=0}^{T-1} p_\pi(s_t = s))$, in which $p_\pi(s_t = s)$ is the probability of the agent using policy π achieving state s at time t, and T is the number of time steps in an episode (Berseth et al., 2021; Rhinehart et al., 2021). This objective not only encourages information seeking like minimizing the entropy of the state, but also encourages the agent to prevent changes in the state.

The problem with these objectives for building skillsets by information seeking is that more information does not mean larger skillsets. There can be settings where gaining a small number of bits of information can have a large impact on the size of an agent's skillset. For instance, consider a scenario where (a) a robot is unsure which of two buttons starts a microwave and (b) if the agent presses the wrong button, the microwave is disabled. With respect to the feature of the state representing the correct start button for the microwave, the uncertainty is not large as there is only one bit of information to be gained. However, this single bit can have a large effect on the number of observations an agent can reliably target, because by knowing how to start the microwave, the agent can reliably cook a variety of foods. On the other hand, gaining many bits of information about the underlying state can have a little to no impact on the size of an agent's skillset. For instance, a robot can count the number of blades of grass outside the building it is in, ask strangers their middle names, or stare at randomly sampled QR codes and gain many bits of information. But gaining information about these features of the state space likely will not expand the number of observations they can reliably target. Given that realistic settings are filled with state features that have a negligible impact on skillset size, trying to maximize information gain likely will not be an efficient way to maximize skillset size. On the other hand, maximizing empowerment can offer a more practical

approach to information seeking as empowerment-maximizing agents will focus on the bits of information that most affect the size of their skillsets.

CHAPTER 3

Learning Diverse Skillsets with Empowerment

Empowerment has been a popular objective for addressing the challenge of learning diverse skillsets in an unsupervised manner, and with good reason. The mutual information between skills and observations does provide a principled way to measure how many distinct skills are in some set of policies. Thus, maximizing the mutual information with respect to different skillsets should enable agents to learn large sets of distinct skills from some starting representation. Yet, the initial unsupervised skill learning approaches that used pure empowerment objectives (Gregor et al., 2016; Florensa et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018) were not able to learn large skillsets. For instance, Florensa et al. (2017), Eysenbach et al. (2019), and Achiam et al. (2018) only report learning at most 20 skills in popular MuJoCo domains like Swimmer and Half Cheetah, which have relatively large observation spaces and where agents were able to execute skills for hundreds of steps. The empowerment, or the maximum number of skills an agent can execute in these domains, should be orders of magnitude larger.

After the initial set of empowerment-based algorithms was unable to generate large skillsets, another wave of unsupervised skill discovery approaches argued that changes

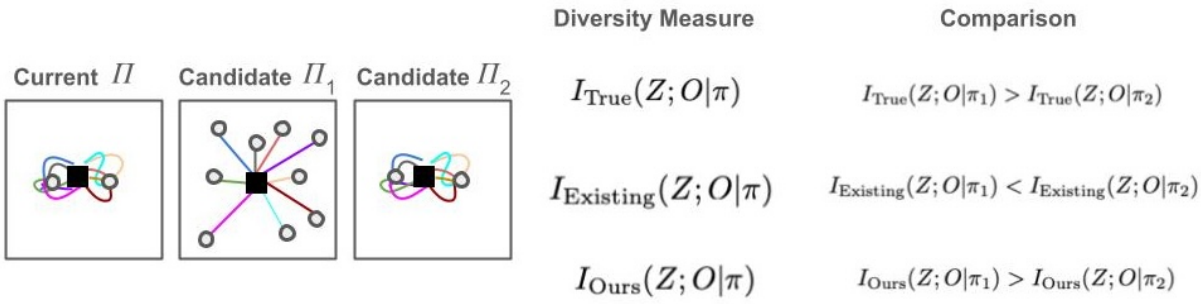


Figure 3.1: Existing empowerment-based skill discovery approaches maximize a loose lower bound on mutual information that can significantly underestimate the diversity of skillsets that differ from the current skillset. For instance, candidate skillset Π_1 is more diverse and has larger mutual information (i.e., $I_{\text{True}}(Z;O)$) than candidate skillset Π_2 as each policy in Π_1 targets a unique observation. Yet because the loose mutual information bound employed by existing approach can penalize skillsets that are different from the current skillset, existing approaches may assign a larger diversity score to the redundant skillset. Our variational mutual information objective that provides a tighter bound on mutual information fixes this issue and would assign a larger diversity score to candidate skillset Π_1.

needed to be made to the mutual information objective so that more skills could be learned (Zheng et al., 2025a; Strouse et al., 2022; He et al., 2025; Hu et al., 2024, 2025; Park et al., 2022, 2023b, 2024b). Some argued that regularization terms (Strouse et al., 2022; He et al., 2025; Zheng et al., 2025a) should be added to boost exploration. Other works argued that the skills should be factorized such that each dimension of the skill makes changes to a particular state dimension (Hu et al., 2024, 2025; Carr et al., 2025). Others discussed how the individual entropy terms in the mutual information needed to be swapped for other terms (Laskin et al., 2022; Zheng et al., 2025a; Warde-Farley et al., 2019; Achiam et al., 2018; Zhang et al., 2021). Others argued that mutual information is not capable of learning large skillsets in continuous settings (Park et al., 2022, 2023b, 2024b). Yet, the results of these works still showed skillset sizes that were orders of magnitude smaller than the true empowerment of representations in these environments.

This chapter argues that the primary reason for the limited skillsets produced by the original empowerment-based skill discovery algorithms was that they were maximizing a loose lower bound on the mutual information between skills and observations. The loose

bound was particularly poor for skillsets that deviated from the current greedy skillset. In other words, when agents were evaluating the diversity of candidate skillsets, they were often significantly underestimating the diversity of those skillsets, particularly those that differed from the current skillset, even if they were truly more diverse than the current one. Figure 3.1 provides an illustration of this idea. The result was often stagnant skillsets that did not grow over time (Campos et al., 2020; Park et al., 2022; Levy et al., 2023).

The cause of the loose mutual information lower bound used by prior work is the structure and the training of variational posterior distribution. When a variational mutual information estimate is used to evaluate the diversity of some candidate skillset, (i) the variational posterior should condition on some variable representing this candidate skillset and (ii) this variational posterior should be trained to match the true posterior of the candidate skillset, not the current skillset. We fix this issue with a new variational lower bound on mutual information. In this objective, (i) the variational posterior is conditioned on a variable representing the skillset under consideration and (ii) the variational posterior is trained to match the true posterior of the candidate skillset.

Yet, maximizing this new variational mutual information objective with respect to an agent's skillset is difficult. The optimization problem is equivalent to a bandit problem in which actions are skillsets and the reward is the mutual information of executing the skillset action from some representation (i.e., the reward is how diverse the skillset is). Optimizing this objective with a traditional actor-critic architecture is not practical. The actor could be implemented as a mapping from an agent's representation to an action vector that represents an agent's skillset. Specifically, the action vector could represent the parameters of a skill-conditioned policy neural network, which means the skillset action could be thousands of dimensions long. The problem with this implementation, though, lies with the critic, which would need to approximate how mutual information changes from small, simultaneous adjustments to numerous parameters in the skill-conditioned policy.

To make maximizing our new variational mutual information objective tractable, we introduce a new actor-critic architecture. The actor takes the form described above. But instead of training a single traditional critic, we show that we can simulate the gradients of this critic using parameter-specific critics. That is, for each parameter in the skill-conditioned policy action, we train in parallel a different parameter-specific critic to approximate the mutual information for skillsets that involve small changes to only one of the parameters of the skill-conditioned policy while the other parameters remain constant.

We support our claim that the combination of (a) the tighter variational lower bound on mutual information and (b) the new actor-critic architecture can produce large skillsets with a variety of empirical results. Because our approach is compute intensive, all environments we trained on have low-dimensional underlying state spaces. Yet many have other challenging features such as stochastic transition dynamics and/or high-dimensional observations (i.e., images). In all settings, our approach learned vastly larger skillsets than the baselines, which include algorithms from the original set of pure empowerment-based skill discovery algorithms, including VIC (Gregor et al., 2016) and DIAYN Eysenbach et al. (2019), and one leading algorithm from the next wave of empowerment-adjacent algorithms, CSF (Zheng et al., 2025a).

3.1 Deriving a Tighter Bound on Mutual Information

To make computing the empowerment of a representation more tractable, both existing work and our approach start with a lower-bound empowerment objective (in equations 3.1 and 3.2) that limits the search space of skillsets to those formed by (i) skill-conditioned policies $\pi_z : \mathcal{C} \times \mathcal{Z} \to \mathcal{A}$, which map representations and skills to actions, and (ii) fixed

distributions over skills $p(z)$, rather than all possible skillsets:

$$\max_{p(\pi)} I(\Pi; O_n|c) \geq \max_{\pi_z} I(Z; O_n|c, \pi_z) \tag{3.1}$$

$$= \max_{\pi_z} \mathbb{E}_{z\sim p(z), o_n \sim p(o_n|c,\pi_z,z)}[\log p(z|c, \pi_z, o_n) - \log p(z))]. \tag{3.2}$$

The lower bound results from the combination of the Data-Processing Inequality and the smaller search space.

Our approach differs from prior work, though, in the precise variational mutual information objective we use to replace the true mutual information. Computing the true mutual information in equations 3.1 and 3.2 for different candidate skill-conditioned policies π_z is not tractable in continuous settings as a result of the posterior distribution $p(z|c, \pi_z, o_n)$, which would require integrating over the intermediate variables (e.g., $a_0, o_1, c_1, \ldots, a_{n-1}, o_n$) in a trajectory. Instead, it is common practice to replace the original posterior with a variational posterior distribution (e.g., $q_\psi(z|c, \pi_z, o_n)$ parameterized by ψ (Mohamed and Rezende, 2015):

$$I^V(Z; O_n|c, \pi_z) = \mathbb{E}_{z\sim p(z), o_n \sim p(o_n|c,\pi_z,z)}[\log q_\psi(z|c, \pi_z, o_n) - \log p(z))]. \tag{3.3}$$

This produces a lower bound on the original mutual information (i.e., $I^V(Z; O_n|c, \pi_z) \leq I(Z; O_n|c, \pi_z)$) and the gap is an average KL divergence between the true and variational posteriors (Barber and Agakov, 2003; Poole et al., 2019):

$$I(Z; O_n|c, \pi_z) - I^V(Z; O_n|c, \pi_z) = \mathbb{E}_{o_n\sim p(o_n|c,\pi_z)}[D_{KL}(p(z|c, \pi_z, o_n)||q_\psi(z|c, \pi_z, o_n))] \tag{3.4}$$

That is, the variational lower bound on mutual information will be tight if the variational posterior is close to the true posterior.

There are two issues with the way prior work (Gregor et al., 2016; Eysenbach et al., 2019; Achiam et al., 2018) handles the variational mutual information objective when

evaluating the diversity of some candidate skillset π_z, which together can result in a loose lower bound. First, instead of using the variational posterior $q_\psi(z|c, \pi_z, o_n)$, they remove the π_z conditioning variable so that the variational posterior is $q_\psi(z|c, o_n)$. Second, they train $q_\psi(z|c, o_n)$ to only match the posterior of the *current* skillset $p(z|c, \pi_z^{\text{Current}}, o_n)$ using the KL divergence objective $D_{KL}(p(z|c, \pi_z^{\text{Current}}, o_n)||q_\psi(z|c, o_n))$. This means that for some candidate skillset π_z, prior work will be measuring the diversity of that skillset with a variational mutual information term I^V that uses a variational posterior trained to match the true posterior of the current skillset π_z^{Current}, not the true posterior of the candidate skillset. As a result, if the true posterior of the candidate skillset differs significantly from the true posterior of the current skillset (e.g., same skills target different observations), then the variational mutual information will form a loose lower bound and thus significantly underestimate the diversity of the candidate skillset, even when the candidate skillset is more diverse. Note, however, that this lower bound will be tighter when candidate skillsets are more similar to the current skillset as the KL divergence between the candidate and current posteriors will not be too different. Thus, the variational mutual information objective used by prior work can discourage skillsets that are both more diverse and different from the current skillset in favor of skillsets that are more similar to the current greedy skillset.

We address this issue and produce a tighter variational lower bound on mutual information by (i) including a variable representing π_z in the posterior (i.e., $q_\psi(z|c, \pi_z, o_n)$) and then (ii) training the variational posterior to match the true posterior of the candidate skillset π_z: $\min_\psi D_{KL}(p(z|c, \pi_z, o_n)||q_\psi(z|c, \pi_z, o_n))$. Thus, we also define variational mutual information using equation 3.3. These changes enable agents to better measure the diversity of candidate skillsets, which in turn makes it more likely they can learn diverse skillsets. Note that the next section will detail how we can represent an entire skillset π_z as a scalar value input to the variational posterior. Theorem 1 proves that this procedure for the variational posterior distribution produces a tighter lower bound on

mutual information.

Theorem 1. *Let $q_{\psi_0}(z|c, \pi_z, o_n)$ be a variational posterior distribution that (i) includes a conditioning variable representing the candidate skillset π_z and (ii) is trained to minimize the KL divergence: $D_{KL}(p(z|c, \pi_z, o_n)||q_\psi(z|c, \pi_z, o_n))$. Let $q_{\psi_1}(z|c, o_n)$ be a second variational posterior distribution that (i) does not include a conditioning variable representing the candidate skillset π_z and (ii) is trained to minimize the KL divergence: $D_{KL}(p(z|c, \pi_z^{Current}, o_n)||q_\psi(z|c, o_n))$. Then, the variational mutual information for candidate skillset π_z produced by ψ_0, $I^{V_0}(Z; O_n|c, \pi_z)$, will form a tighter bound on the true mutual information $I(Z; O_n|c, \pi_z)$ than the variational mutual information produced by ψ_1, $I^{V_1}(Z; O_n|c, \pi_z)$:*

$$I^{V_1}(Z; O_n|c, \pi_z) \leq I^{V_0}(Z; O_n|c, \pi_z) \leq I(Z; O_n|c, \pi_z).$$

Proof. To start, $I(Z; O_n|c, \pi_z)$ is larger than both $I^{V_0}(Z; O_n|c, \pi_z)$ and $I^{V_1}(Z; O_n|c, \pi_z)$ because their differences depend on average KL divergences between the true posterior of the candidate skillset $p(z|c, \pi_z, o_n)$ and the variational posteriors $q_{\psi_0}(z|c, \pi_z, o_n)$ and $q_{\psi_1}(z|c, o_n)$, and KL divergences are nonnegative (Cover and Thomas, 2006). Next, given that ψ_0 can take the form of either (a) ψ_1 if that is the optimal set of parameters for minimizing the average KL divergence between the true posterior and the variational posterior $q_{\psi_0}(z|c, \pi_z, o_n)$ or (b) any other set of parameters, then the average KL divergence between the true posterior and the variational posterior formed by ψ_0 must be no larger than the average KL divergence between the true posterior and the variational posterior formed by ψ_1: $\mathbb{E}_{o_n \sim p(o_n|c,\pi_z)}[D_{KL}(p(z|c, \pi_z, o_n)||q_{\psi_0}(z|c, \pi_z, o_n))] \leq \mathbb{E}_{o_n \sim p(o_n|c,\pi_z)}[D_{KL}(p(z|c, \pi_z, o_n)||q_{\psi_1}(z|c, o_n))]$. Given that the average KL divergence is the difference between the true and variational mutual information terms (Barber and Agakov, 2003), then $I(Z; O_n|c, \pi_z) - I^{V_0}(Z; O_n|c, \pi_z) \leq I(Z; O_n|c, \pi_z) - I^{V_1}(Z; O_n|c, \pi_z)$ or $I^{V_0}(Z; O_n|c, \pi_z) \geq I^{V_1}(Z; O_n|c, \pi_z)$. □

3.2 Maximizing the Mutual Information Lower Bound

While the variational mutual information in equation 3.3 provides a tighter bound on the true mutual information and thereby enables agents to more accurately measure the diversity of a skillset, it is challenging to maximize with respect to skillsets formed from skill-conditioned policies. That is, it is difficult to compute the following variational empowerment objective:

$$\mathcal{E}^V(c) = \max_{\pi_z} I^V(Z; O_n|c, \pi_z) \tag{3.5}$$

$$= \max_{\pi_z} \mathbb{E}_{z\sim p(z), o_n \sim p(o_n|c,\pi_z,z)}[\log q_\psi(z|c, \pi_z, o_n) - \log p(z)] \tag{3.6}$$

This objective takes the form of a bandit problem with a potentially massive action space. Here the bandit policy to be learned is a mapping from the starting representation c to an action vector that represents the skill-conditioned policy π_z. Specifically, this vector would contain the potentially thousands of parameters (i.e., the weights and biases) that make up the skill-conditioned policy neural network. The reward is the variational mutual information $I^V(Z; O_n|c, \pi_z)$ for executing that skillset action π_z in representation c. That is, the reward is how diverse the skillset π_z is when it is executed from c.

The problem is that the traditional actor-critic would not be practical as a result of the large action space. An actor $f_\lambda : \mathcal{C} \to \pi_z$ that maps representations to skill-conditioned policy parameters could be implemented. However, a critic $Q_\alpha(c, \pi_z)$ would be challenging to train because the critic would need to approximate $I^V(Z; O_n|c, \pi_z)$ for many skill-conditioned policies π_z that contain small changes to numerous parameters within π_z, which is a difficult function to learn.

Instead, we will use a different actor-critic approach that "simulates" the gradient from the traditional actor-critic method. The key insight is that in the traditional architecture, the gradient of the critic with respect to any parameter λ_j in the actor f_λ is $\frac{dQ}{\lambda_j} = \sum_{i=0}^{|\pi_z|-1} \frac{dQ}{d\pi_z^i} \frac{d\pi_z^i}{d\lambda_j}$, where π_z^i is the i-th entry of the π_z vector. That is, we only need

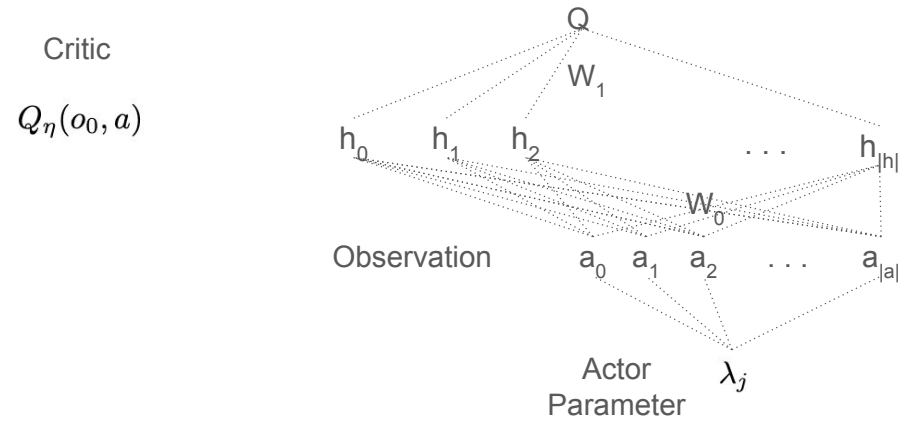


Figure 3.2: Figure visualizes the function form of a 1 hidden layer critic. We use this visual to show that the derivative of Q with respect to a parameter λ_j of the bandit policy actor depends on the derivatives of Q with respect to the individual entries in the skill-conditioned policy vector π_z.

to simulate how the Q function changes in response to small adjustments to each of the parameters in π_z (i.e., $\frac{dQ}{d\pi_z^i}$) while the others remain constant.

Before describing how we will take advantage of this insight when designing our actor-critic, we first show that this is true for a 1-hidden layer MLP critic $Q_\alpha(c, \pi_z = f_\lambda(c))$ with respect to some parameter λ_j in the bandit policy actor $f_\lambda(c)$. The critic will take the following form, which is visualized in Figure 3.2. The output $Q = n(\mathbf{h}W_1)$, in which $n(\cdot)$ is a nonlinear function; $\mathbf{h}$ is the hidden layer vector with $|h|$ dimensions; and $\mathbf{h}W_1$ performs the dot product between vector $\mathbf{h}$ and the weight matrix W_1, which is simply a vector because Q is a scalar. Next, each entry $h_i \in \mathbf{h}$ is defined $h_i = n(\pi_z W_{0,i})$, in which $W_{0,i}$ is the i-th column of weight matrix W_0. Note that in this definition the neural network connections between the observation c and h_i are ignored because c has no dependence on the parameters of the bandit policy actor λ. Lastly, each entry $\pi_z^i \in \pi_z$ is defined $\pi_z^i = f(\lambda_j, c, \lambda_{/j})$. That is, each entry in π_z is some function of the parameter λ_j under consideration, the initial observation c, and the other parameters (excluding λ_j) in λ.

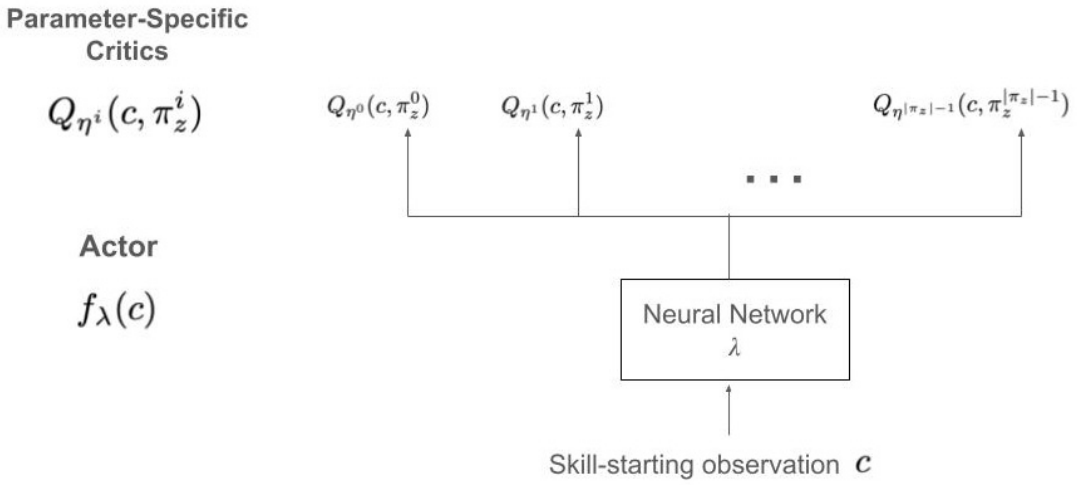


Figure 3.3: Visual of how the parameter-specific critics attach to the actor. In this case, the actor maps a representation c to the parameters of the skill-conditioned policy $\pi_z = [\pi_z^0, \pi_z^1, \ldots, \pi_z^{|\pi_z|-1}]$. For each dimension in π_z, there is a critic $Q_{\alpha^i}(c, \pi_z^i)$ that approximates the variational mutual information of executing the skill-conditioned policy π_z^i from representation c. π_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., $f_\lambda(c)[j]$), while the i-th parameter takes on value π_z^i.

With this functional form,

$$\begin{aligned}
\frac{dQ}{d\lambda_j} =& \frac{dQ}{d(\mathbf{h}W_1)}\Big(\sum_{i=0}^{|h|-1} \frac{d(\mathbf{h}W_1)}{dh_i}\frac{dh_i}{d(\pi_z W_{0,i})}\Big(\sum_{k=0}^{|\pi_z|-1}\frac{d(\pi_z W_{0,i})}{d\pi_z^k}\frac{d\pi_z^k}{d\lambda_j}\Big)\Big) \\
=& \sum_{k=0}^{|\pi_z|-1}\frac{d\pi_z^k}{d\lambda_j}\Big(\sum_{i=0}^{|h|}\frac{dQ}{d\mathbf{h}W_1}\frac{d(\mathbf{h}W_1)}{dh_i}\frac{dh_i}{d(\pi_z W_{0,i})}\frac{d(\pi_z W_{0,i})}{d\pi_z^k}\Big) \\
=& \sum_{k=0}^{|\pi_z|-1}\frac{dQ}{d\pi_z^k}\frac{d\pi_z^k}{d\lambda_j}
\end{aligned} \tag{3.7}$$

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 π_z (i.e., $\frac{dQ}{d\pi_z^k}$ for $k = 0, \ldots, |\pi_z| - 1$).

Given that we can match the gradients from the traditional actor-critic approach if we can accurately estimate $\frac{dQ}{d\pi_z^i}$ (i.e., how mutual information changes from small changes to one parameter of π_z assuming the other parameters are constant), we will use a new actor-critic architecture. In this architecture, we train, in parallel, parameter-specific

critics $Q_{\alpha^i}(c, \pi_z^i)$ to respectively approximate $I^V(Z; O_n|c, \pi_z^i)$ for $i = 0, \ldots, |\pi_z| - 1$, in which π_z^i is a *scalar* representing the skill-conditioned policy in which all entries in π_z take on their greedy values from the actor $f_\lambda(c)$ except the i-th parameter which takes on value π_z^i. We then use the trained critics to update the actor f_λ so that it outputs more diverse skill-conditioned policies π_z using equation 3.7. Figure 3.3 provides a visual of the parameter-specific actor-critic architecture. Note that to train the parameter-specific critics, we also need to train in parallel parameter-specific variational posteriors $q_{\psi^i}(z|c, \pi_z^i, o_n)$ for $i = 0, \ldots, |\pi_z|$ to minimize the KL divergence between the true posterior $p(z|c, \pi_z^i, o_n)$ and the variational posterior $q_{\psi^i}(z|c, \pi_z^i, o_n)$.

Algorithm 1 Actor-Critic Method for Maximizing $I^V(Z; O_n|c, \pi_z)$ w.r.t. π_z

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

Algorithm 1 provides the full algorithm for the actor-critic method for maximizing our variational mutual information lower bound with respect to π_z. The first step is to train, in parallel and until convergence, all parameter-specific variational posteriors $q_{\psi^i}(z|c, \pi_z^i, o_n)$ to match the true posteriors $p(z|c, \pi_z^i, o_n)$ for noisy skillsets π_z^i. The second step is to train all parameter-specific critics $Q_{\alpha^i}(c, \pi_z^i)$, in parallel and until convergence, to approximate variational mutual information for noisy skillsets π_z^i. The final step is to update the actor.

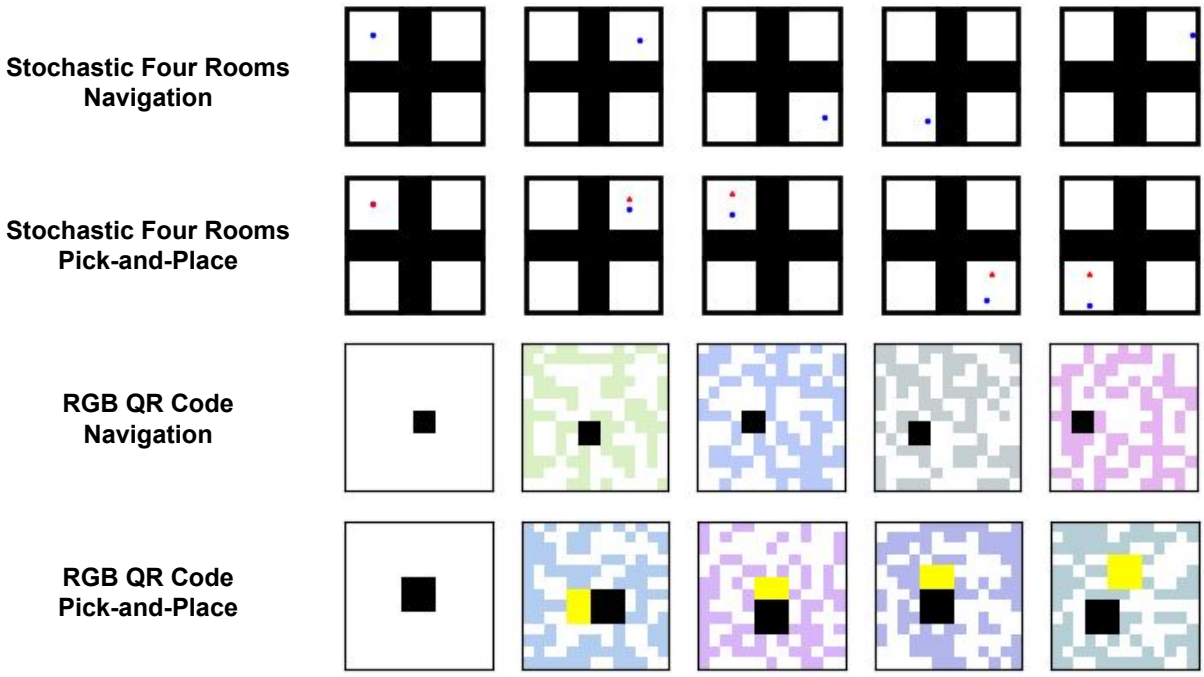


Figure 3.4: Sample state trajectories executed by a random policy in four of the five settings. The blue marker in the stochastic four rooms agent is the agent navigating around. The red marker in the pick-and-place task is the object that can be moved if the agent is within some threshold. The black square in the RGB QR code tasks is the agent, and the yellow square in the pick-and-place task is the object that can be moved if the agent is within some threshold.

3.3 Experiments

We support our hypothesis that maximizing a tighter lower bound on mutual information should have a large impact on the size of skillsets an agent can learn through several experiments.

3.3.1 Environments

We tested our approach and a mix of baselines on five environments. All have fully observable states and low-dimensional underlying state spaces, but they each have features that make them more difficult such as stochastic transition dynamics and/or high-dimensional observations. We need to test on domains with low-dimensional underlying state spaces because maximizing mutual information is a compute intensive optimization

as it requires searching across a large space of skillsets. We provide visualizations for four of the five settings in Figure 3.4.

1. 2D Room

   Our first setting is a 2D room with walls on all four sides. Actions are continuous, 2D vectors and reflect changes in (x, y) position. Observations are continuous, 2D vectors and show the agent's (x, y) position. The primary challenge in this domain relative to the rest is that the skill-starting observation can be anywhere in the room. That is, the agent needs to learn skillsets from all locations in the room and these skillsets will differ significantly depending on where the agent starts. For instance, the agent's set of skills should act differently when the agent starts in the top left corner of the room versus the center of the room. The remainder of the environments will have a single starting state. In this task, skills consist of 8 primitive actions, while in the rest, skills consist of 5 primitive actions.

2. Stochastic Four Rooms Navigation

   The next setting is another room environment but with stochastic transition dynamics. Actions are again 2D and reflect changes in (x, y) position. Observations are 2D and show the agent's (x, y) position. But now the transition dynamics work by first executing the agent's action in the current room and then transporting the agent to the corresponding position in one of the four rooms selected at random. So if the agent finished an action at a position $(\Delta x, \Delta y)$ from the center of the bottom left room, the agent may be transported to a position that is $(\Delta x, \Delta y)$ from the center of the top right room. Note that the observations show the actual position rather than the delta from the center of each room.

3. Stochastic Four Rooms Pick-and-Place

   This manipulation domain takes place in the same stochastic four rooms setting except now there is an object that the agent can manipulate if the agent is within a

certain threshold of the object. Note that in each stochastic transition, the agent and object will travel to the same randomly selected room.

4. 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).

5. RGB QR Code Pick-and-Place

   This environment is the same as the navigation task except there is now an object that can be moved if the object is within reach.

Given that the goal of the thesis is to develop unsupervised skill discovery approaches that can learn large skillsets, we evaluate agents by the variational mutual information of the skillsets they have learned.

3.3.2 Baselines

We compare our approach to a mix of baselines. In the 2D room task we compare to Variational Intrinsic Control (VIC) (Gregor et al., 2016), Diversity Is All You Need (DIAYN) (Eysenbach et al., 2019), Contrastive Successor Features (CSF) (Zheng et al., 2025a), and a variant of our approach where we train the parameter-specific critics to approximate the loose lower bound of mutual information where the parameter-specific variational posteriors are not conditioned on the skillset under consideration. The remaining four environments compare our approach to VIC and a variant of Goal-Conditioned RL (Choi et al., 2021).

We compare to VIC and DIAYN in order to compare to two of the original pure empowerment-based skill discovery methods. VIC is similar to our approach as it maxi-

mizes the mutual information between skills and skill-terminating observations, but uses the loose lower bound on the mutual information by not including the candidate skillset in the variational posterior. DIAYN uses a different mutual information than both VIC and our approach as it is between skills and observations from randomly selected time steps along a trajectory. DIAYN also includes a policy entropy term in the objective that encourages stochastic policies. We compare to CSF as it is one of the leading algorithms among the recent wave of unsupervised skill discovery approaches. CSF provides a mutual information-based version of METRA (Park et al., 2024b), which is another leading approach. We compare to the baseline of our approach that uses the loose lower bound to assess whether the performance of our approach is due to the tighter bound or the bandit style approach to skillset optimization. We also compare to a goal-conditioned RL approach as these have been successful in deterministic settings but are generally not tested in stochastic settings where specific states cannot be achieved with high probability.

3.3.3 Results

Table 3.1: Average (±std) variational mutual information of learned skillsets (nats)

Algorithm	**2D Room**	**4 Rooms Nav**	**4 Rooms Pick**	**QR Nav**	**QR Pick**
Ours	8.0 ± 0.0	5.1 ± 0.3	8.7 ± 0.3	3.5 ± 0.1	6.0 ± 0.2
VIC	4.1 ± 1.3	0.2 ± 0.4	-0.1 ± 0.3	-0.4 ± 0.0	-0.6 ± 0.1
DIAYN	-0.4 ± 0.0	N/A	N/A	N/A	N/A
CSF	-0.4 ± 0.7	N/A	N/A	N/A	N/A
Loose Bound	4.1 ± 0.8	N/A	N/A	N/A	N/A
GCRL	N/A	0.3 ± 0.4	3.9 ± 0.6	-0.4 ± 0.3	-2.6 ± 5.8

Table 4.1 provides the average variational mutual information (5 random seeds) for our approach and the set of baselines. The results are reported in the logarithm unit of nats. For instance, the skillset our approach learned in the 2D room contains 8.0 nats of skills, which is equivalent to around 2980 skills. As the table reports, our approach learned vastly larger skillsets than the baselines in all settings. The baseline that came closest was VIC in the 2D room, but our approach still learned around 50 times more

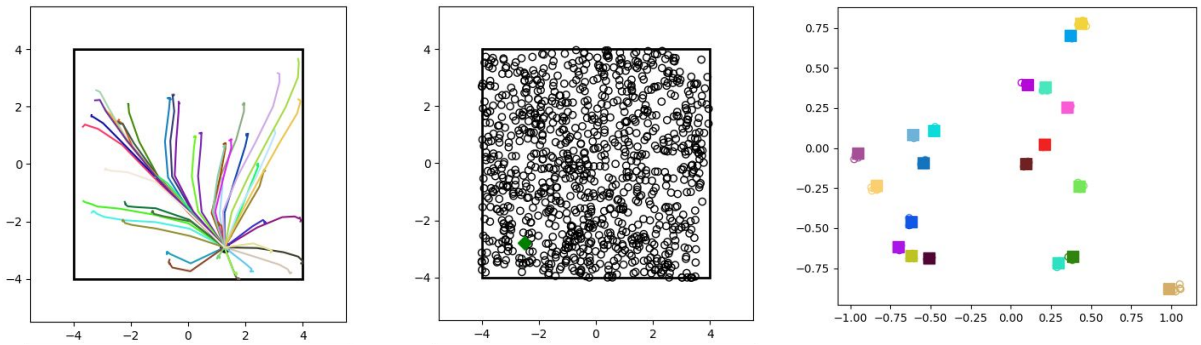


Figure 3.5: Some visualization of the entropy terms in the 2D room domain. Both the left and center images visualize the $H(O_n)$ entropy term, which shows the observation coverage of the skillset. Specifically, the left image shows the trajectories from a single starting observation produced by 45 randomly sampled skills. The center image shows the skill-terminating (x, y) positions from 1000 randomly sampled skills when starting at the green marker near the bottom left corner. The right image visualizes the $H(Z)$ and $H(Z|O_n)$ entropy terms, which sample skills from the fixed skill space $p(z)$ and the trained variational posterior $q_\psi(z|o_0, \pi_z, o_n)$, respectively. The randomly sampled skills are shown by the 20 squares. For each skill, 5 samples from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$ are shown as circles. The large observation coverage and the very tight variational posterior around each skill shows the agents is learning large, diverse skillsets.

skills (our skillset had 3.9 more nats of skills).

We can also assess the performance of our approach by visualizing the four entropy terms in the symmetric version of the mutual information between skills and observations: $H(O_n)$, $H(O_n|Z)$, $H(Z)$, and $H(Z|O_n)$. For instance, Figure 3.5 visualizes the $H(O_n)$, $H(Z)$, and $H(Z|O_n)$ terms for the 2D room domain. As the $H(O_n)$ images show, the skills in the learned skillset target nearly all the possible observations. The $H(Z|O_n)$ image, which visualizes the samples from the tight variational posterior, shows that skills are targeting highly precise (x, y) coordinates. Figures 3.6-3.9 visualize the entropy terms for the remaining settings. These similarly show that our approach is learning large and diverse skillsets.

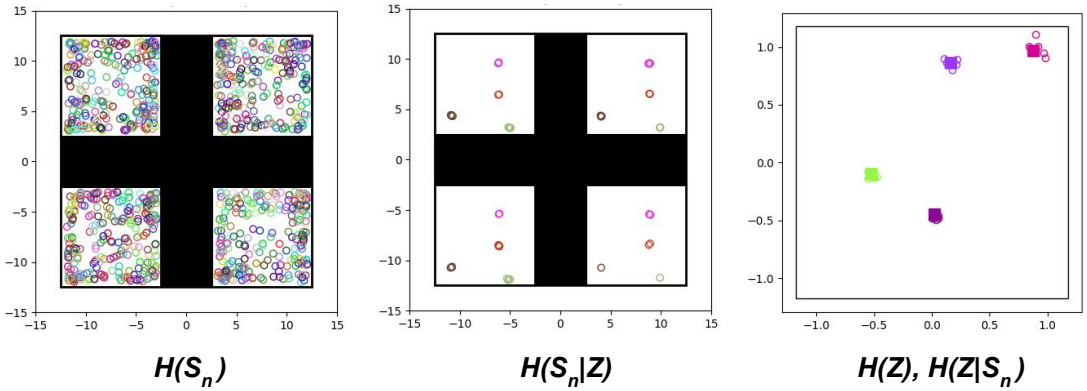


Figure 3.6: Entropy visualizations for the stochastic four rooms domain. Left image visualizes $H(O_n)$ by marking the skill-terminating state from 1000 skills randomly sampled. The center image visualizes $H(O_n|Z)$ by showing 12 samples of skill-terminating states from 4 specific skills randomly sampled. The right image visualizes (i) $H(Z)$ by showing the skill space (black rectangle) and (ii) $H(Z|O_n)$ by showing samples of the variational posterior (empty circles) for four different skills (filled squares)).

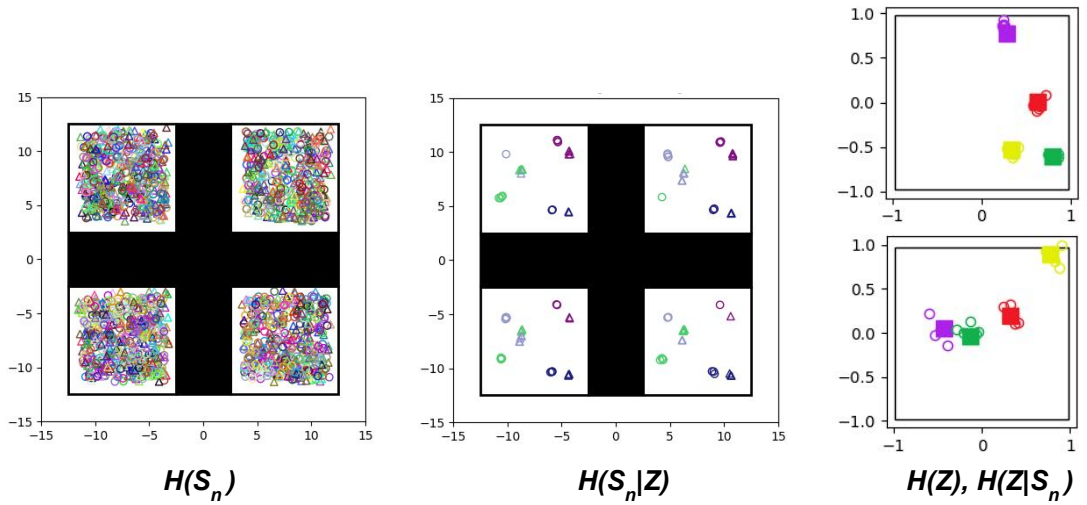


Figure 3.7: Images show the entropy visualizations for the stochastic four rooms pick-and-place domain. The left image shows the skill-terminating observations o_n that result from 1000 skills uniformly sampled from the learned skill space. The near uniform coverage of the observation space shows that $H(O_n)$ 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 observations. Per the image, each skill targets an abstract observation representing an offset from the center of a room for both the agent and object, showing that $H(O_n|Z)$ is low. The right image focuses on four skills and shows 5 samples from the variational posterior $q_\psi(z|o_0, l, \pi_z, o_n)$. Per the image, the samples form a narrow distribution around the executed skill, showing that $H(Z|O_n)$ is low.

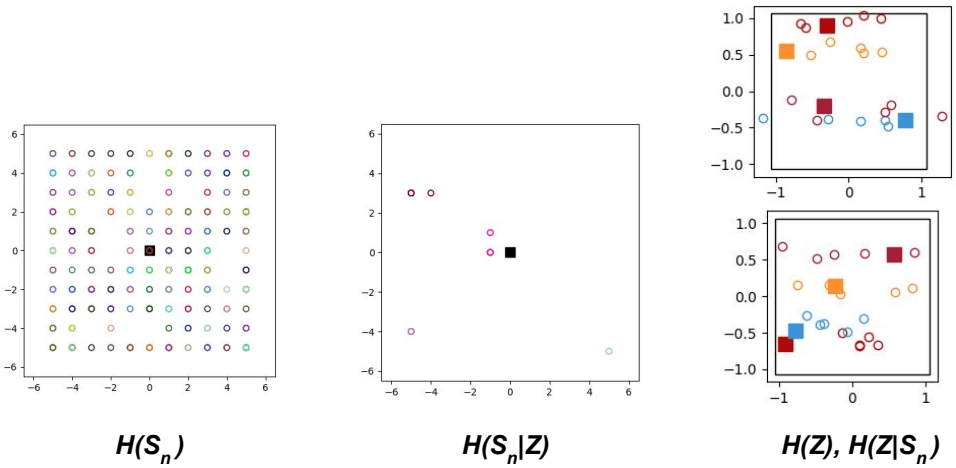


Figure 3.8: Entropy visualizations for the RGB QR code navigation task. Left image visualizes $H(O_n)$ by marking the skill-terminating states o_n produced by executing 1000 samples of skills from the learned skill space. Center image visualizes $H(O_n|Z)$ by executing four skills 12 times each and recording the skill-terminating observations. 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.

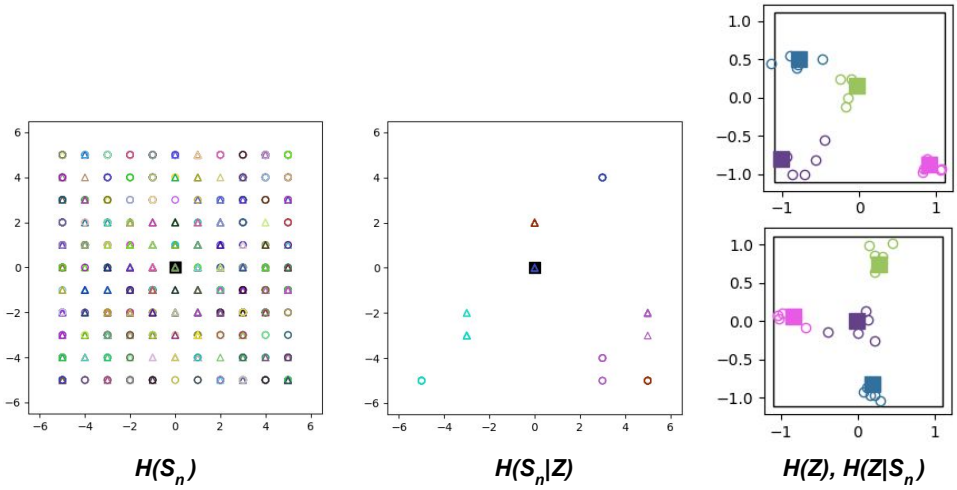


Figure 3.9: Entropy visualizations for the RGB QR code pick-and-place tasks. Left image visualizes $H(O_n)$ by marking the skill-terminating observations o_n produced by executing 1000 samples of skills from the learned skill space. Center image visualizes $H(O_n|Z)$ by executing four skills 12 times 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, the agent learns a diverse skillset as $H(O_n)$ is large while the conditional entropies $H(O_n|Z)$ and $H(Z|O_n)$ are small.

CHAPTER 4

Reducing State Uncertainty from Belief State Aliasing using Empowerment

The previous chapter showed how empowerment addresses the problem of learning diverse sets of policies in an unsupervised manner. Yet, agents will not be able to learn many distinct policies if the agent operates in representations with high uncertainty in the underlying state. This is because more uncertainty produces more randomness in the transition dynamics, which in turn broadens the distributions of observations a skill targets. If this broadening results in skills targeting overlapping distributions of observations, the skills can become redundant and the number of distinct skills in a skillset falls. Thus, a second major challenge to learning large skillsets in an unsupervised manner is minimizing state uncertainty.

In order to manage state uncertainty, agents must address the two sources of state uncertainty. One source is from the inherent partial observability in the environment. That is, for some histories of actions and observations h_t, the probability of the underlying state (i.e., $p(s_t|h_t)$) will not always be 1. For instance, when the agent in the drone setting in

Figure 1.4 has not seen which button is the start button for the drone, the agent will have some uncertainty about the underlying state (i.e., $p(\text{left button} == \text{On} \mid \text{history } h_t) = \frac{1}{2}$). The second source of state uncertainty is from poor representation learning that causes belief state aliasing. Belief state aliasing occurs when an agent's RNN (or other representation learning function such as an observation encoder in fully observable settings) takes two histories of actions and observations with different underlying distributions of state (i.e., different belief states) and encodes them to the same or similar representations. Figure 1.2 depicts this situation in the drone setting where the agent's RNN maps two histories, where one history signals that the left button is the "On" button while the other history signals that the right button is the "On" button, to the same representation. For this aliased representation c, $p(\text{left button} == \text{On} \mid c) = \frac{1}{2}$. As a result of both of these sources of uncertainty, the size of the skillset that can be executed from representation c falls because the agent cannot reliably control the drone.

This chapter shows how maximizing empowerment with respect to the parameters of the representation learning function discourages belief state aliasing. That is, maximizing empowerment with respect to an RNN, for instance, encourages the RNN to disentangle histories representing different underlying distributions of states. This should be an intuitive result because if some belief state aliasing is depressing the size of an agent's skillset, training the RNN to maximize the size of an agent's skillset using empowerment should naturally encourage the RNN to disentangle these aliased histories. To bolster this hypothesis, this chapter provides both theoretical and empirical support.

We provide two theorems showing how empowerment encourages agents to seek information (i.e., reduce uncertainty) about their underlying state through representation learning. First, we prove that the average empowerment produced by an RNN is maximized when the agent's representations are sufficient statistic representations of histories with respect to the underlying state. That is, empowerment is maximized when there is no belief state aliasing in the agent's representations. This result is notable because it

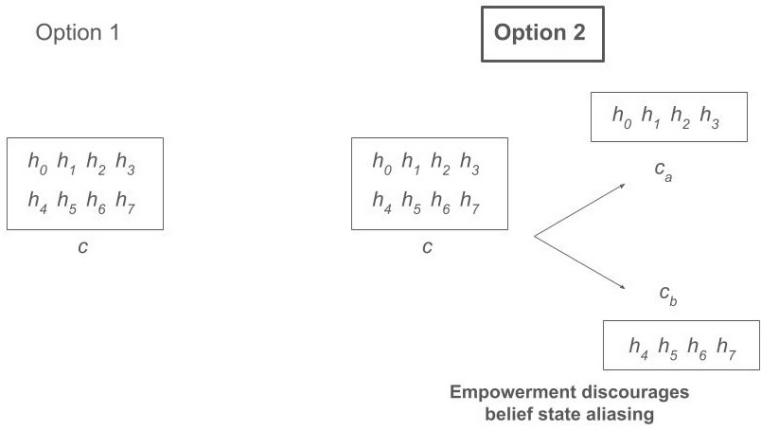


Figure 4.1: Our second theorem shows that Empowerment provides a dense reward for discouraging belief state aliasing. That is, RNNs are rewarded for disentangling histories that represent different underlying distributions of states even if the RNN is not able to fully disentangle histories. For instance, if the RNN has two options: (i) make no change and (ii) disentangle a group of eight histories such that they are broken up into two representations, empowerment will encourage the RNN to perform the latter even though it does not fully disentangle the histories.

means that training an RNN to maximize empowerment can be an objective for learning representations with minimal belief state uncertainty. Yet, this theorem does not provide evidence that the empowerment reward is dense with respect to the RNN. It may take numerous small changes for an RNN to output sufficient statistic representations of histories and the first theorem alone does not explain whether empowerment reward is sufficiently dense to encourage the RNN to make these changes.

Our second theorem proves that empowerment is a dense reward by showing that given any two RNNs, if one RNN provides more information about the underlying state than the other, then the average empowerment produced by the more informative RNN will be at least as large as the average empowerment produced by the other RNN (and larger under some easily achievable conditions), and thus the empowerment objective will prefer the RNN that provides more information. As a result, empowerment is a dense reward as it encourages RNNs to disentangle representations even if the resulting representations are not yet sufficient statistic representations. For instance, as described in Figure 4.1, if an RNN is deciding between two options in which one option involves

no change to its parameters, and the second option involves disentangling a set of eight histories that correspond to different belief states such that the set of eight is split into two representations, empowerment will encourage this disentanglement so long as it increases the agent's skillset. Future RNN updates can then try to further disentangle the remaining histories. We also note that empowerment not only encourages more informative representations, but it will encourage the RNN to disentangle those representations that are most depressing the size of an agent's skillset. In real-world settings there may be numerous histories with different belief states that become entangled, but many of these aliased state features will have little to no impact on the agent's skillset size and thus should not be a priority for disentanglement. By training the RNN to directly maximize skillset size, the RNN should prioritize the particular belief state aliasing that most affects the number of distinct skills in an agent's skillset.

We also support our claim that empowerment discourages belief state aliasing with a series of empirical results. In both fully observable and partially observable settings, we show that empowerment enables agents to simultaneously learn skills and informative representations, ultimately resulting in large skillsets. Our approach vastly outperforms other baselines such as VIC Gregor et al. (2016), DIAYN (Eysenbach et al., 2019), CSF (Zheng et al., 2025a), and variations of our approach.

4.1 Theoretical Analysis

We next provide our two theorems. The purpose of the first theorem is to show that average empowerment produced by an RNN can be no larger than the average empowerment produced by a representation function that outputs sufficient statistic representations. That is, the average empowerment produced by an RNN is maximized when it outputs sufficient statistic representations of histories with respect to the underlying state. This is an important result because it shows empowerment can be an objective for removing belief state aliasing without actually needing access to the underlying state space.

Yet this theorem does not provide evidence that empowerment can be a dense reward for continually encouraging information seeking (i.e., uncertainty reduction). The purpose of the second theorem is to show that empowerment is a dense reward for information seeking. This theorem shows that if there are two RNNs and one RNN provides additional information about the underlying state, it will produce larger average empowerment under certain conditions and thereby be encouraged by the empowerment objective.

Theorem 2. *Let η be the parameters of any RNN $f_\eta : \mathcal{C} \times \mathcal{A} \times \mathcal{O} \to \mathcal{C}$ and let $\eta_x : \mathcal{X} \times \mathcal{A} \times \mathcal{O} \to \mathcal{X}$ be a function that takes as input sufficient statistics x_t, actions a_t, and observations o_{t+1} and outputs sufficient statistic representations x_{t+1}, then the average empowerment produced by η is upper bounded by the average empowerment produced by η_x:* $\mathbb{E}_{c_0 \sim p(c_0|\eta)}[\mathcal{E}(c_0, \eta)] \leq \mathbb{E}_{x_0 \sim p(x_0|\eta_x)}[\mathcal{E}(x_0, \eta_x)]$.

Proof.

$$\mathbb{E}_{c_0 \sim p(c_0|\eta)}[\mathcal{E}(c_0, \eta)] = \mathbb{E}_{c_0 \sim p(c_0|\eta)}[I(Z; O_n | c_0, \eta, \pi_z^*)] \tag{4.1}$$

$$\leq \mathbb{E}_{c_0 \sim p(c_0|\eta), x_0 \sim p(x_0|c_0, \eta)}[I(Z; O_n | c_0, x_0, \eta, \pi_z^*)] \tag{4.2}$$

$$\leq \mathbb{E}_{x_0 \sim p(x_0|\eta_x)}[I(Z; O_n | x_0, \eta_x, \pi_z^x)] \tag{4.3}$$

$$\leq \mathbb{E}_{x_0 \sim p(x_0|\eta_x)}[I(Z; O_n | x_0, \eta_x, \pi_z^{x,*})] \tag{4.4}$$

$$= \mathbb{E}_{x_0 \sim p(x_0|\eta_x)}[\mathcal{E}(x_0, \eta_x)] \tag{4.5}$$

□

Line 4.1 applies the definition of the empowerment of the tuple (c_0, η), in which c_0 is a learned representation and η is a set of RNN parameters. π_z^* represents the mutual information maximizing skill-conditioned policy parameters when the starting representation is c_o and the RNN parameters are η.

The lower bound in line 4.2 applies the convexity property of mutual information with respect to the channel distribution (Cover and Thomas, 2006; Capdepuy, 2011). The

convexity property states that in a mutual information $I(A;B)$, if the channel distribution (i.e., the distribution $p(b|a)$ for all $p(a,b) > 0$) is a weighted mixture of channels $p(b|a,c)$ (i.e., $p(b|a) = \int_c p(c)p(b|a,c)$), then the original mutual information of the mixed channel is upper bounded by the average mutual information of the individual channels in the mixture (i.e., $I(A;B) \leq \mathbb{E}_{c\sim p(c)}[I(A;B|C)]$). In our case, the channel distribution, which provides the distribution of skill-terminating observations given (i) the starting representation c_0, (ii) RNN parameters η, (iii) skill-conditioned policy π_z^*, and (iv) skill z, is $p(o_n|c_0,\eta,\pi_z^*,z)$. And this channel distribution is a weighted mixture of the channels that also include the sufficient statistic representation: $p(o_n|c_0,\eta,\pi_z^*,z) = \int_{x_0} p(x_0|c_0,\eta)p(o_n|c_0,x_0,\eta,\pi_z^*,z)$. Consequently, $I(Z;O_n|c_0,\eta,\pi_z^*) \leq \mathbb{E}_{x_0\sim p(x_0,\eta)}[I(Z;O_n|c_0,x_0,\eta,\pi_z^*)]$.

Line 4.3 removes the dependence on the RNN parameterized by η by (i) using the RNN defined by η_x to produce the representations that serve as the input to the skill-conditioned policy and (ii) replacing π_z^* with a specific skill-conditioned policy π_z^x. As we will show, this will replace each mutual information term, $I(Z;O_n|c_0,x_0,\eta,\pi_z^*)$, with a new mutual information term, $I(Z;O_n|x_0,\eta_x,\pi_z^x)$, that is at least as large. For each context x_0, π_z^x will be constructed as follows. For each x_0, find the tuple (c_0,x_0) with the largest $I(Z;O_n|c_0,x_0,\eta,\pi^*)$ as there can be multiple contexts c_0 associated with the same sufficient statistic x_0. Then, for each x_0 let π_z^x be the skill-conditioned policy distribution $p(a_t|x_0,x_t) = \int_{c_t} p(c_t|c_0,x_0,x_t,s_t)p(a_t|c_0,x_0,x_t,c_t,s_t) = \int_{c_t} p(c_t|c_0,x_0,x_t)p(a_t|c_t)$, in which $p(a_t|c_t)$ is the probability specified by the skill-conditioned policy defined by π_z^*. $p(c_t|c_0,x_0,x_t,s_t) = p(c_t|c_0,x_0,x_t)$ because given the tuple (c_0,x_0), $C_t \to X_t \to S_t$ form a Markov chain. (Note that the skill-conditioned policies in this proof will take as input the step number t as input, or there can n skill-conditioned policies for each of the n actions taken in a skill). Thus, the skill-conditioned policy π_z^x will have the same distribution over actions as executed by π_z^* when conditioned on the contexts x_t from the RNN defined by η_x. Next, we show that for all (c_0,x_0), the original mutual information $I(Z;O_n|c_0,x_0,\eta,\pi_z^*)$ is upper bounded by the new mutual information $I(Z;O_n|x_0,\eta_x,\pi_z^x)$.

We first show that for each mutual information term $I(Z; O_n|c_0, x_0, \eta, \pi_z^*)$ from which π_z^x was constructed in the previous step, $I(Z; O_n|x_0, \eta_x, \pi_z^x) = I(Z; O_n|c_0, x_0, \eta, \pi_z^*)$. That is, we replace the original mutual information term with an equivalent mutual information term. Given that the source distributions $p(z)$ are the same by definition as they are fixed, to show that the mutual information terms are the same, we need to show that the channel distributions $p(o_n|c_0, x_0, \eta, \pi_z^*, z) = p(o_n|x_0, \eta_x, \pi_z^x, z)$ are the same for all (z, o_n) tuples. We show this below by proving by induction that the joint distributions $p(x_{t-1}, s_{t-1}, a_{t-1}, s_t, o_t, x_t|c_0, x_0, \eta, \pi_z^*, z) = p(x_{t-1}, s_{t-1}, a_{t-1}, s_t, o_t, x_t|x_0, \eta_x, \pi_z^x, z)$ for $t = 1, \ldots, n$. Then because the joint distributions $p(x_{n-1}, s_{n-1}, a_{n-1}, s_n, o_n, x_n|c_0, x_0, \eta, \pi_z^*, z) = p(x_{n-1}, s_{n-1}, a_{n-1}, s_n, o_n, x_n|x_0, \eta_x, \pi_z^x, z)$ for the last time step, the channel distributions are equal: $p(o_n|c_0, x_0, \eta, \pi_z^*, z) = p(o_n|x_0, \eta_x, \pi_z^x, z)$.

The proof by induction goes as follows. In the base case at $t = 1$, the distribution $p(x_0|c_0, x_0, \eta, \pi_z^*, z) = p(x_0|x_0, \eta_x, \pi_z^x, z)$ because the same x_0 appears in the conditioning variables. $p(s_0|c_0, x_0, \eta, \pi_z^*, z) = p(s_0|x_0, \eta_x, \pi_z^x, z)$ because $C_0 \to X_0 \to S_0$ form a Markov chain. $p(a_0|c_0, x_0, \eta, \pi_z^*, z, s) = p(a_0|c_0, x_0, \eta, \pi_z^*, z) = p(a_0|x_0, \eta_x, \pi_z^x, z)$ using the definition of π_z^x. $p(s_1, o_1|c_0, x_0, \eta, \pi_z^*, z, s_0, a_0) = p(s_1, o_1|x_0, \eta_x, \pi_z^x, z, s_0, a_0)$ as the next state s_1 and observation o_1 only depend on s_0 and a_0 and is independent of the other variables. Lastly, $p(x_1|c_0, x_0, \eta, \pi_z^*, z, s_0, a_0, o_1) = p(x_1|x_0, \eta_x, \pi_z^x, z, s_0, a_0, o_1)$ because the next context x_1 only depends on x_0, a_0, o_1, which are the same in both cases. Thus, the base case of the induction proof is true as $p(x_0, s_0, a_0, o_1, x_1|c_0, x_0, \eta, \pi^*, z) = p(x_0, s_0, a_0, o_1, x_1|x_0, \eta_x, \pi_z^x, z)$.

Assuming the proof holds through $t = k-1$, then at step $t = k$, $p(x_{k-1}, s_{k-1}|c_0, x_0, \eta, \pi_z^*, z) = p(x_{k-1}, s_{k-1}|x_0, \eta_x, \pi_z^x, z)$ because the joint distributions $p(x_{k-2}, s_{k-2}, a_{k-2}, o_{k-1}, x_{k-1}|c_0, x_0, \eta, \pi_z^*, z) = p(x_{k-2}, s_{k-2}, a_{k-2}, o_{k-1}, x_{k-1}|x_0, \eta_x, \pi_z^b, z)$. $p(a_{k-1}|c_0, x_0, \eta, \pi_z^*, z, x_{k-1}) = p(a_{k-1}|x_0, \eta_x, \pi_z^x, z, x_{k-1})$ using the definition of π_z^x. Again, $p(s_k, o_k, x_k|c_0, x_0, \eta, \pi_z^*, z, x_{k-1}, s_{k-1}, a_{k-1}) = p(s_k, o_k, x_k|x_0, \eta_x, \pi_z^x, z, x_k$ as s_k, o_k only depend on s_{k-1} and a_{k-1} and x_k only depends on x_{k-1}, a_{k-1}, o_k. Thus, the induction proof holds through step $t = k$ as $p(x_{k-1}, s_{k-1}, a_{k-1}, s_k, o_k, x_k|c_0, x_0, \eta, \pi_z^*, z) = p(x_{k-1}, s_{k-1}, a_{k-1}, s_k, o_k, x_k|x_0, \eta_x, \pi_z^x, z)$.

Thus, $I(Z; O_n|x_0, \eta_x, \pi_z^x) = I(Z; O_n|c_0, x_0, \eta, \pi_z^*)$ for those (c_0, x_0) tuples from which π_z^x was constructed. For the other smaller $I(Z; O_n|c_0, x_0, \eta, \pi_z^*)$ terms that were not used to construct π_z^x, these will also be replaced by $I(Z; O_n|x_0, \eta_x, \pi_z^x)$. If in these replacements, the new mutual information $I(Z; O_n|x_0, \eta_x, \pi_z^x) > I(Z; O_n|c_0, x_0, \eta, \pi_z^*)$, the inequality in line 4.3, will be replaced by a strictly less than.

In line 4.4, the lower bound results from replacing the skill-conditioned policy π_z^x with the optimal skill-conditioned policy $\pi_z^{x,*}$ for the specific x_0 context and RNN defined by η_x. If this replacement produces larger mutual information, then the inequality becomes a strictly less than.

The final line 4.5 uses the definition of the empowerment of a sufficient statistic x_0 with representation distribution defined by η_x. This completes the proof that an RNN defined by η produces an average empowerment that is upper bounded by the averaged empowerment of a function η_x that generates sufficient statistic representations.

Theorem 3. *Let η^a and η^b be the parameters of two RNNs such that for any joint distribution $p(h, c_0^a, c_0^b, s)$ in which h is a history, c_0^a and c_0^b are the contexts produced by processing the history through the two respective RNNs a and b, and s be a state sampled from the belief state distribution $p(s|h)$. Assuming (i) $I(C_0^b; S|C_0^a) > 0$ (i.e., η^b provides more information about the underlying state than η^a) and (ii) $C_t^a \to C_t^b \to S_t$ form a Markov chain when conditioned on c_0^a, c_0^b for $t = 0, 1, \ldots, n-1$, then $\mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[\mathcal{E}(c_0^a, \eta^a)] \leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[\mathcal{E}(c_0^b, \eta^b)]$.*

Proof.

$$\mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[\mathcal{E}(c_0^a, \eta^a)] = \mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[I(Z; O_n|c_0^a, \eta^a, \pi_z^{a,*})] \tag{4.6}$$

$$\leq \mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a), c_0^b \sim p(c_0^b|c_0^a, \eta^a)}[I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})] \tag{4.7}$$

$$\leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[I(Z; O_n|c_0^b, \eta^b, \pi_z^b)] \tag{4.8}$$

$$\leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[I(Z; O_n|c_0^b, \eta^b, \pi_z^{b,*})] \tag{4.9}$$

$$= \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[\mathcal{E}(c_0^b, \eta^b)] \tag{4.10}$$

□

Line 4.6 applies the definition of the empowerment of the tuple containing a context representation and RNN parameters. $\pi^{a,*}$ represents the mutual information maximizing skill-conditioned policy parameters for the (context, RNN) tuple of (c_0^a, η^a).

The lower bound in line 4.7 applies the convexity property of mutual information with respect to the channel distribution (Cover and Thomas, 2006; Capdepuy, 2011). In our case, if the RNN defined by η^b provides more information about the underlying state than η^a (i.e., $I(C_b; S|C_a) > 0$), then the channel distribution $p(o_n|c_0^a, \eta^a, \pi_z^{a,*}, z)$ is a weighted mixture of channels $p(o_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z)$ (i.e., $p(o_n|c^a, \eta^a, \pi_z^{a,*}, z) = \int_{c_0^b} p(c_0^b|c_0^a, \eta^a) p(o_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z)$) and so the mutual information of the mixed channel $I(Z; O_n|c_0^a, \eta_a, \pi_z^{a,*})$ is upper bounded by the average mutual information of the individual channels $\mathbb{E}_{c_0^b \sim p(c_0^b|c_0^a)}[I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})]$.

Line 4.8 removes the dependence on the RNN parameterized by η^a by (i) using the RNN defined by η^b to produce the representations that serve as inputs to the skill-conditioned policy and (ii) replacing $\pi_z^{a,*}$ with a specific skill-conditioned policy π_z^b. As we will show, this will replace each mutual information term, $I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$, with a new mutual information term, $I(Z; O_n|c_0^b, \eta_b, \pi_z^b)$, that is at least as large. For each context c_0^b, π_z^b will be constructed as follows. For each c_0^b, find the tuple (c_0^a, c_0^b) with the largest $I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi^{a,*})$ as there can be multiple contexts c_0^a associated

with the same c_0^b. Then, for each c_0^b let π_z^b be the skill-conditioned policy distribution $p(a_t|c_0^b, c_t^b) = \int_{c_t^a} p(c_t^a|c_0^a, c_0^b, c_t^b, s_t)p(a_t|c_0^a, c_0^b, c_t^b, c_t^a, s_t) = \int_{c_t^a} p(c_t^a|c_0^a, c_0^b, c_t^b)p(a_t|c_t^a)$, in which $p(a^t|c_t^a)$ is the probability specified by the skill-conditioned policy defined by $\pi_z^{a,*}$. That is, the skill-conditioned policy π_z^b will have the same distribution over actions as executed by $\pi_z^{a,*}$ when conditioned on the contexts c_t^b from the RNN defined by η^b. Next, we show that for all (c_0^a, c_0^b), the original mutual information $I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$ is upper bounded by the new mutual information $I(Z; O_n|c^b, \eta^b, \pi_z^b)$.

We first show that for each mutual information term $I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$ from which π_z^b was constructed in the previous step, $I(Z; O_n|c_0^b, \eta^b, \pi_z^b) = I(Z; O_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$. That is, we replace the original mutual information term with an equivalent mutual information term. Given that the source distributions $p(z)$ are the same by definition as they are fixed, to show that the mutual information terms are the same, we need to show that the channel distributions $p(o_n|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(o_n|c_0^b, \eta^b, \pi_z^b, z)$ are the same for all (z, o_n) tuples. We show this below by proving by induction that the joint distributions $p(c_{t-1}^b, s_{t-1}, a_{t-1}, s_t, o_t, c_t^b|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_{t-1}^b, s_{t-1}, a_{t-1}, s_t, o_t, c_t^b|c_0^b, \eta^b, \pi_z^b, z)$ for $t = 1, \ldots, n$. Then because the joint distribution $p(c_{n-1}^b, s_{n-1}, a_{n-1}, s_n, o_n, c_n^b|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_{n-1}^b, s_{n-1}, a_{n-1}, s_n, o_n, c_n^b|c_0^b, \eta^b, \pi_z^b, z)$ at the last time step, the channel distributions are equal: $p(o_n|c_0^a, c_0^b, \eta_a, \pi_z^{a,*}, z) = p(o_n|c_0^b, \eta_b, \pi_z^b, z)$.

The proof by induction goes as follows. In the base case at $t = 1$, the distribution $p(c_0^b|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_0^b|c_0^b, \eta^b, \pi_z^b, z)$ because the same c_0^b appears in the conditioning variables. $p(s_0|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(s_0|c_0^b, \eta^b, \pi_z^b, z)$ because $p(s_0|c_0^a, c_0^b, z) = p(s_0|c_0^b, z)$ per the Markov assumption in the theorem statement. $p(a_0|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z, s_0) = p(a_0|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(a_0|c_0^b, \eta^b, \pi_z^b, z)$ using the definition of π_z^b. $p(s_1, o_1|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z, s_0, a_0) = p(s_1, o_1|c_0^b, \eta^b, \pi_z^b, z, s_0, a_0)$ as the next state s_1 and observation o_1 only depends on s_0 and a_0 and are independent of the other variables. Lastly, $p(c_1^b|c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z, s_0, a_0, o_1) = p(c_1^b|c_0^b, \eta^b, \pi_z^b, z, s_0, a_0, o_1)$ because the next context c_1^b only depends on c_0^b, a_0, o_1, which are the same in both cases. Thus, the base case of the induction proof is true as

$p(c_0^b, s_0, a_0, s_1, o_1, c_1^b | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_0^b, s_0, a_0, s_1, o_1, c_1^b | c_0^b, \eta^b, \pi_z^b, z)$.

Assuming the proof holds through $t = k-1$, then at step $t = k$, $p(c_{k-1}^b, s_{k-1} | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_{k-1}^b, s_{k-1} | c_0^b, \eta^b, \pi_z^b, z)$ because the joint distribution $p(c_{k-2}^b, s_{k-2}, a_{k-2}, s_{k-1}, o_{k-1}, c_{k-1}^b | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_{k-2}^b, s_{k-2}, a_{k-2}, s_{k-1}, o_{k-1}, c_{k-1}^b | c_0^b, \eta^b, \pi_z^b, z)$. $p(a_{k-1} | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z, c_{k-1}^b, s_{k-1}) = p(a_{k-1} | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, p(a_{k-1} | c_0^b, \eta^b, \pi_z^b, z, c_{k-1}^b)$ using the definition of π_z^b. Again, $p(s_k, o_k, c_k^b | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z, c_{k-1}^b, s_{k-1}, a_{k-1}) = p(s_k, o_k, c_k^b | c_0^b, \eta^b, \pi_z^b, z, c_{k-1}^b, s_{k-1}, a_{k-1})$ as s_k, o_k only depend on s_{k-1} and a_{k-1} and c_k^b only depends on c_{k-1}^b, a_{k-1}, o_k. Thus, the induction proof holds through step $t = k$ as $p(c_{k-1}^b, s_{k-1}, a_{k-1}, s_k, o_k, c_k^b | c_0^a, c_0^b, \eta^a, \pi_z^{a,*}, z) = p(c_{k-1}^b, s_{k-1}, a_{k-1}, s_k, o_k, c_k^b | c_0^b, \eta^b, \pi_z^b, z)$.

Thus, $I(Z; O_n | c_0^b, \eta^b, \pi_z^b) = I(Z; O_n | c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$ for those (c_0^a, c_0^b) tuples from which π_z^b was constructed. For the other smaller $I(Z; O_n | c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$ terms that were not used to construct π_z^b, these will also be replaced by the $I(Z; O_n | c_0^b, \eta^b, \pi_z^b)$ terms that are at least as large as $I(Z; O_n | c_0^a, c_0^b, \eta^a, \pi_z^{a,*})$. If the replacement mutual information terms are larger, then the inequality in line 4.8 will be replaced by a strictly less than.

In line 4.9, the lower bound results from replacing the skill-conditioned policy π_z^b with the optimal skill-conditioned policy $\pi_z^{b,*}$ for the specific c_0^b context and RNN defined by η^b. If this replacement produces larger mutual information, then the inequality becomes a strictly less than.

The final line 4.10 uses the definition of the empowerment of a context c_0^b with representation distribution defined by η^b. This completes the proof that an RNN defined by η^b providing more information on the underlying state than an RNN defined by η_a, produces average empowerment that is at least as large as the other RNN.

4.2 Maximizing Mutual Information w.r.t. Skills and Representations

Maximizing empowerment with respect to a representation function in practice involves maximizing a mutual information objective with respect to both a skill-conditioned policy and a representation function:

$$\max_{\eta} \mathbb{E}_{c_0 \sim p(c_0|\eta)}[\mathcal{E}(c_0, \eta)] \geq \max_{\eta, f_\lambda} \mathbb{E}_{c_0 \sim p(c_0|\eta)}[I(Z; O_n | c_0, \eta, \pi_z = f_\lambda(c_0))]. \tag{4.11}$$

η is the parameters of the representation function and f_λ is a function that outputs the skill-conditioned policy parameters for a particular context c_0. Thus, the left side of 4.11 shows the objective of maximizing, with respect to the representation learning parameters, the average empowerment of the context representations produced by a candidate set of representation learning parameters. Given that the empowerment of a representation still needs to be learned, we maximize a lower bound of this average empowerment objective, in which the maximum mutual information is replaced with the current mutual information term.

To maximize the mutual information term objective in equation 4.11, we use two actor-critic architectures. The first actor-critic is the same as the one introduced in Chapter 3. The actor $f_\lambda : \mathcal{C} \to \pi_z$ maps context representations c to a vector representing the parameters of a skill-conditioned policy. The parameter-specific critics $Q_{\alpha^i} : \mathcal{C} \times \pi_z^i \to \mathbb{R}$ approximate the mutual information of skillsets defined by the scalar π_z^i, which represents a skill-conditioned policy in which all dimensions $j \neq i$ are set to the greedy values of $f_\lambda(c_0)[j]$, while the i-th component is set to π_z^i. To help the parameter-specific critics approximate variational mutual information, the first actor-critic also includes parameter-specific variational posteriors $q_{\psi^i}(z|c_0, \eta, \pi_z^i, o_n)$ that are trained to match the true posteriors of candidate skillsets.

The second actor-critic is responsible for training the representation learning parameters (e.g., the parameters of an RNN) such that they produce high average empowerment. The actor $f_\gamma : v \to \eta$ maps a fixed vector v to the representation learning parameters η. The parameter-specific critics $Q_{\xi^i} : \eta^i \to \mathbb{R}$ approximate the average mutual information produced by the set of representation learning parameters η^i:

$$Q_{\xi^i}(\eta^i) \approx \mathbb{E}_{c_0 \sim \eta^i}[\mathcal{E}(c_0, \eta^i)]. \tag{4.12}$$

Similar to π_z^i, η^i is a scalar that represents the vector of representation learning parameters in which dimensions $j \neq i$ take on the greedy value from the actor $f_\gamma(v)[j]$, while the i-th dimension assumes value η^i. Also, similar to the first actor-critic, to assist the parameter-specific critics in approximating average mutual information, the second actor-critic will also have parameter-specific variational posteriors $q_{\psi^i}(z|c_0, \eta^i, o_n)$ that are trained to match the true posteriors of candidate skillsets. Note the repeated use of ψ^i notation is for simplicity. Different sets of variational posterior parameters are used in the first and second actor-critics.

Algorithm 2 provides the algorithm for updating both the skill discovery and representation learning actor-critics, which are visualized in Figure 4.2. The algorithm alternates between updating the two actor-critics. For each actor-critic update, the parameter-specific critics are initially updated by first updating the parameter-specific variational posteriors and then the parameter-specific critics. For instance, for the skill discovery actor-critic, the variational posteriors $q_{\psi^i}(z|c_0, \pi_z^i, o_n)$ are first trained to match the true posteriors $p(z|c_0, \pi_z^i, o_n)$ for noisy π_z^i for M iterations ($M = 300$ in our experiments). Then the parameter-specific critics $Q_{\alpha^i}(c_0, \pi_z^i)$ are trained to approximate the variational mutual information $I^V(Z; O_n|c_0, \pi_z^i)$ for noisy π_z^i for M iterations. After the variational posteriors and critics have been updated during each actor-critic update, the actor is then updated once using an objective that sums all the parameter-specific critic objectives. For instance, in the representation learning actor-critic, the actor f_γ is updated using the objective

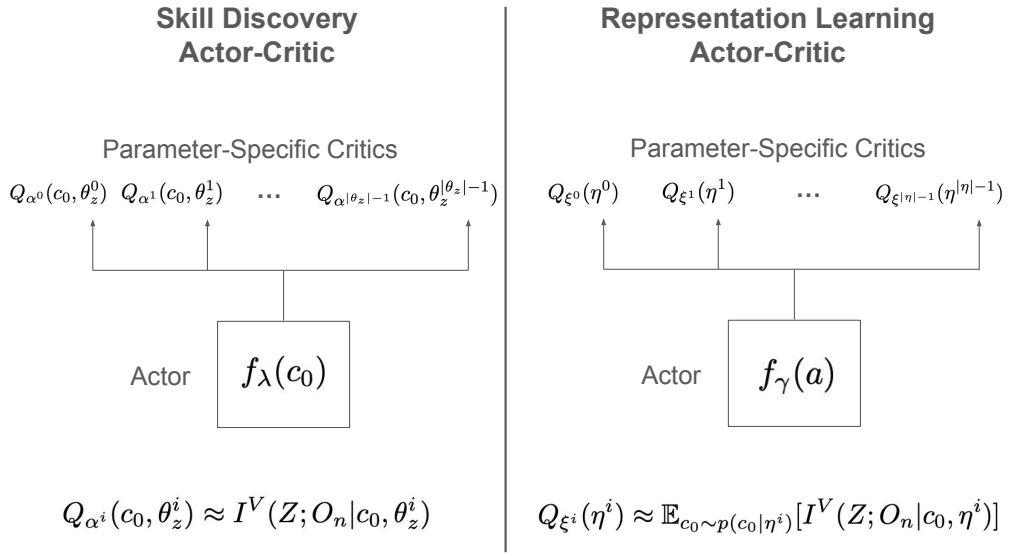


Figure 4.2: We use two actor-critic architectures to maximize our average mutual information with respect the parameters of a skill-conditioned policy and an RNN. The actor-critic on the left is designed to learn diverse skillsets across various contexts c_0. The actor $f_\lambda(c_0)$ maps a context to a skillset θ_z. The parameter-specific critics measure how many distinct skills are in each skillset θ_z^i using variational mutual information. The actor-critic on the right is designed to learn RNNs that output representations producing large average mutual information across all contexts. The actor $f_\lambda(a)$ maps a constant vector a to the parameters, η, of the RNN. Each parameter-specific critic measures the average mutual information produced by the RNN defined by the scalar η^i.

Algorithm 2 Skill Discovery and Representation Learning with Empowerment

▷ Update Skill Discovery Actor-Critic

for all dimensions $i = 0, \ldots, |\pi_z| - 1$ in parallel **do**

  for M iterations **do** ▷ Update Variational Posterior

    Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \epsilon \nabla_{\psi_i}(D_{KL}(p(z|c_0, \pi_z^i, o_n)||q_{\psi^i}(z|c_0, \pi_z^i, o_n)))$ with noisy π_z^i

  end for

  for M iterations **do** ▷ Update Critic

    Update Q_{α^i}: $\alpha^i \leftarrow \alpha^i - \epsilon \nabla_{\alpha^i}((Q_{\alpha^i}(c_0, \pi_z^i) - \text{Target})^2)$ with noisy π_z^i,

    Target $= \mathbb{E}_{z \sim p(z), o_n \sim p(o_n|c_0, \pi_z^i, z)}[\log q_{\psi^i}(z|c_0, \pi_z^i, o_n) - \log p(z)]$

  end for

end for

Update f_λ: $\lambda \leftarrow \lambda + \epsilon \nabla_\lambda (\sum_{i=0}^{|\theta_z|-1} Q_{\alpha^i}(c_0, \pi_z^i = f_\lambda(c_0)[i]))$ ▷ Update Actor

▷ Update Representation Learning Actor-Critic

for all dimensions $i = 0, \ldots, |\eta| - 1$ in parallel **do**

  for M iterations **do** ▷ Update Variational Posterior

    Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \epsilon \nabla_{\psi_i}(D_{KL}(p(z|c_0, \eta^i, o_n)||q_{\psi^i}(z|c_0, \eta^i, o_n)))$ with noisy η^i

  end for

  for M iterations **do** ▷ Update Critic

    Update Q_{ξ^i}: $\xi^i \leftarrow \xi^i - \epsilon \nabla_{\xi^i}((Q_{\xi^i}(\eta^i) - \text{Target})^2)$ with noisy η^i,

    Target $= \mathbb{E}_{c_0 \sim p(c_0|\eta^i), z \sim p(z), o_n \sim p(o_n|c_0, \eta^i, z)}[\log q_{\psi^i}(z|c_0, \eta^i, o_n) - \log p(z)]$

  end for

end for

Update f_γ: $\gamma \leftarrow \gamma + \epsilon \nabla_\gamma (\sum_{i=0}^{|\eta|-1} Q_{\xi^i}(\eta^i = f_\gamma(a)[i]))$ ▷ Update Actor

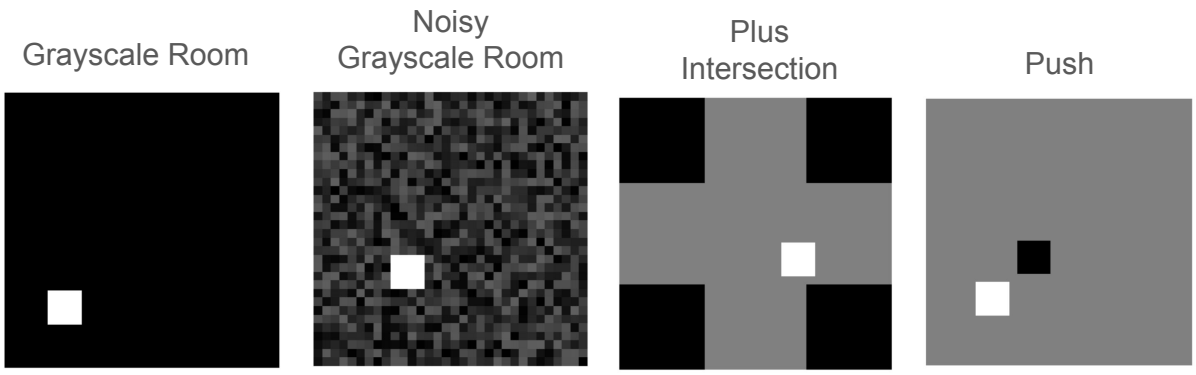


Figure 4.3: Sample image observations from each of the four high-dimensional fully observable settings.

$J(\gamma) = \sum_{i=0}^{|\eta|-1} Q_{\xi^i}(\eta^i = f_\gamma(v)[i])$, where $f_\gamma(v)[i]$ outputs the i-th component of the vector $f_\gamma(v)$.

4.3 Experiments

The central hypothesis of this chapter is that agents can reduce uncertainty and thereby learn large skillsets by jointly maximizing mutual information with respect to both skillsets and representations. We evaluate our approach that implements this hypothesis as well as a collection of baselines in a series of experiments.

4.3.1 Environments

We tested our approach and baselines in both fully observable and partially observable settings.

Fully Observable Domains: We implemented the following six fully observable domains. 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 the agent is white. The second high-dimensional setting is a stochastic version of the previous setting in which darker background pixels are randomly sampled from a range of black to gray colors. The third high-dimensional setting is a "plus" shaped intersection of 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 4.3 shows sample image observations from the high-dimensional settings. Note that these high-dimensional tasks are significantly harder than those in the prior chapter, because the starting observation can be mostly anywhere in the environment so the agent needs to learn large skillsets from most locations in the environment. The number of primitive actions in each skill $n = 7$ for all tasks.

In addition to the experiments described so far where we focus on skill discovery, we also implemented a set of downstream RL experiments in all the high-dimensional settings except the push task. We implemented these downstream RL tasks as Goal-Conditioned RL (GCRL) experiments (Kaelbling, 1993; Schaul et al., 2015; Andrychowicz et al., 2017)

Partially Observable Domains: We evaluate whether our algorithm can learn large skillsets in non-Markov settings using three environments. All environments are small in terms of their observation dimensionality, but all involve continuous observation and action spaces and all are non-Markov. Visuals of all three settings are provided in Figure 4.4.

The first environment is a variant of the T-Maze setting (Bakker, 2001; Allen et al., 2024). In this setting, an agent starts in a thin hallway and at the eastern end of the hallway, a larger hallway perpendicular to the initial hallway opens either to the north or south. In this variant, if the agent tries to enter the larger hallway in the wrong direction (e.g., the agent makes a turn south but the hallway actually opens to the north), the agent becomes stuck for the remainder of the episode. During each episode in evaluation, the agent starts at the western end of the thin hallway and for only the first timestep the

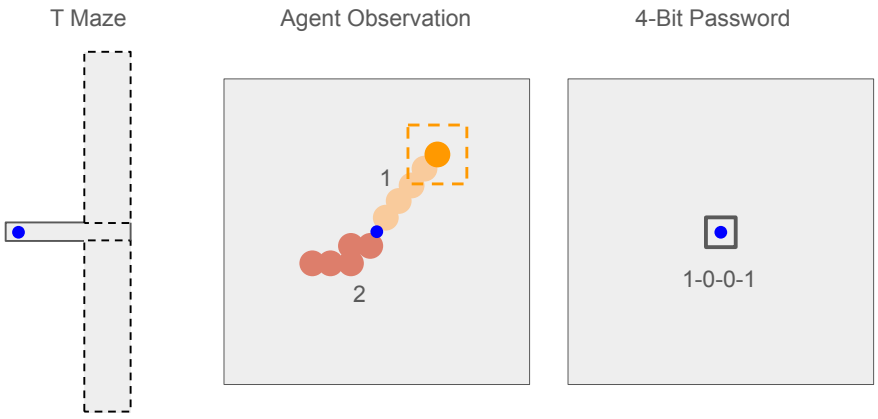


Figure 4.4: Visuals of the three non-Markov environments in which we evaluated our approach.

agent is given a binary signal indicating whether the hallway opens to the north or south. For the next 15 timesteps, the agent remains frozen in place no matter what $(\Delta x, \Delta y)$ action the agent executes. Then for the remainder of the episode, the agent is free to move. The most diverse skillset π_z that maximizes the mutual information is one in which each skill targets a precise region of the (x, y), particularly in the large hallway. To do this, the agent needs to have an RNN that "remembers" the initial binary signal that describes the direction the hallway opens. Otherwise, the only observations an agent's skillset can definitively target are limited to the thin hallway.

In the second environment, Agent Observation, there is a randomly sampled (x, y) goal region that an agent needs to navigate to before time expires, and if the agent fails to do so then the agent is returned to the center for the remainder of the episode. However, unlike a traditional goal-conditioned RL domain where agents are given the goal as part of each observation, in this setting the agent needs to infer the goal from watching two other agents pursue the goal, while the primary agent remains frozen in place. During the first 5 timesteps of this period, the primary agent observes another agent move toward and achieve the goal (shown by the orange circles in Figure 4.4). When the other agent achieves the goal, the primary agent receives a signal that the goal has been achieved. For the following 5 timesteps, the goal-achieving agent is removed and a different "decoy"

agent starts to move randomly in a manner that is unlikely to achieve the goal (red circles in Figure 4.4). After these 10 timesteps, the primary agent then must attempt to move within a threshold of the goal (dashed orange square in Figure 4.4) in the next eight actions. The mutual information maximizing skillset executes skills that first head to the goal region, which was shown by the first observed agent, and then target distinct (x, y) positions. In order to build this skillset, the agent needs an RNN representation that remembers the goal location targeted by the first agent.

The third setting is a 4-bit password setting. During evaluation, the agent starts each episode in the cage and receives a single bit of the password for each of the first four timesteps. During each of the next four timesteps, the agent can output 1 bit of the password. If there are any mistakes, the agent remains stuck in the cage for the remainder of the episode. Note that this setting is more challenging than the first two as the agent needs an RNN that remembers a sequence of observations (i.e., the password) as well as a sequence of actions (i.e., the number of bits of the password the agent has already output). The mutual information maximizing skillset will be one in which most skills start each episode by outputting the correct 4-bit password and then target distinct regions outside the cage.

Note that during the training episodes (i.e., non-evaluation episodes), we also provide the agent with a type of curriculum to make it easier to jointly learn representations and skills conditioned on those representations. These curricula are implemented as a wider distribution of histories that include histories that extend the duration of observations providing information about the underlying state. For instance, in T-Maze, the number of timesteps that include the binary flag indicating the direction of the hallway is randomly sampled from the range $[1, 16]$ during training episodes, where 1 is the same as the evaluation episodes. Similarly, in Agent Observation we randomly sample the duration of the first agent that achieves the goal from the range of $[5, 10]$ timesteps, in which 5 is the same as the evaluation episodes and 10 means no decoy agent is shown. In 4-Bit

Password, we assist the agent by executing the correct bit (even if the agent outputs the wrong bit) and provide the next bit as an additional dimension to the observation for a randomly sampled number of timesteps in the range $[0, 3]$, where 0 means no help is provided and 3 means the agent does not have to correctly execute the first 3 bits of the password and is also provided with a hint of the final bit. The purpose of adding these curricula is that when mutual information is maximized with respect to both the RNN and skill-conditioned policy, there is a chicken-and-egg problem that arises. The RNN may be considering a change that adds information about the underlying state (i.e., reduces uncertainty) by assigning different representations to histories with different belief states. However, if the $f_\lambda(c)$ actor, which outputs skill-conditioned policies π_z, has not been trained on these possible new representations, the skill-conditioned policy may be poor, which may then cause the RNN to disregard this information-preserving change. By extending the length of the information signal, the agent's RNN will then consider histories in which the last observation has the signal about the underlying state. If the RNN happens to preserve information by assigning different representations to these final observations with different signals, the skill-conditioned policies can be trained to be effective in these representations. Then, in settings where there is no extended signal and the RNN is considering a change from some entangled representation to representations that preserve information and those representations already have good skill-conditioned policies, the RNN will be encouraged to make that change.

4.3.2 Baselines

Fully Observable Baselines: In the fully observable 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., 2025a). The main differences between these approaches and our approach are 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 achieve a tighter mutual information lower bound, these approaches 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 observations 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., 2024b), which is another recent leading 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 (i.e., this ablation maximizes a loose lower bound on mutual information), 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 2, 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 achieves this, but VIC does not treat the skill-conditioned policy parameters as the 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 downstream GCRL 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).

Partially Observable Baselines: We compare to two other algorithms in the three partially observable settings. First we compare to the fully observable version of our approach that trains an observation encoder to maximize empowerment rather than an RNN. This comparison will assess whether the RNN is learning representations that are closer to sufficient statistics of histories rather than simply encodings of the last observation. We also compare to our approach but with a fixed RNN. This comparison will test whether a randomly initialized RNN by default assigns representations that disentangle histories, which may mean no training of an RNN is necessary.

Except for the downstream task settings, all tasks evaluate agents based on the size of their skillsets, measured using average variational mutual information $\mathbb{E}_{c_0 \sim p(c_0|\eta)}[I^V(Z; O_n|c_0, \eta, \pi_z]$.

4.3.3 Results

Fully Observable Skill Discovery Results

Table 4.1 shows the fully observable variational mutual information results for all algorithms in all settings. Note that (i) the mutual information is shown in the logarithmic

Table 4.1: Average (±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

units of nats (e.g., in the Noisy 2D room domain, the agent learns 7.6 nats of skills or $\approx 2,000$ 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. This outperformance shows 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

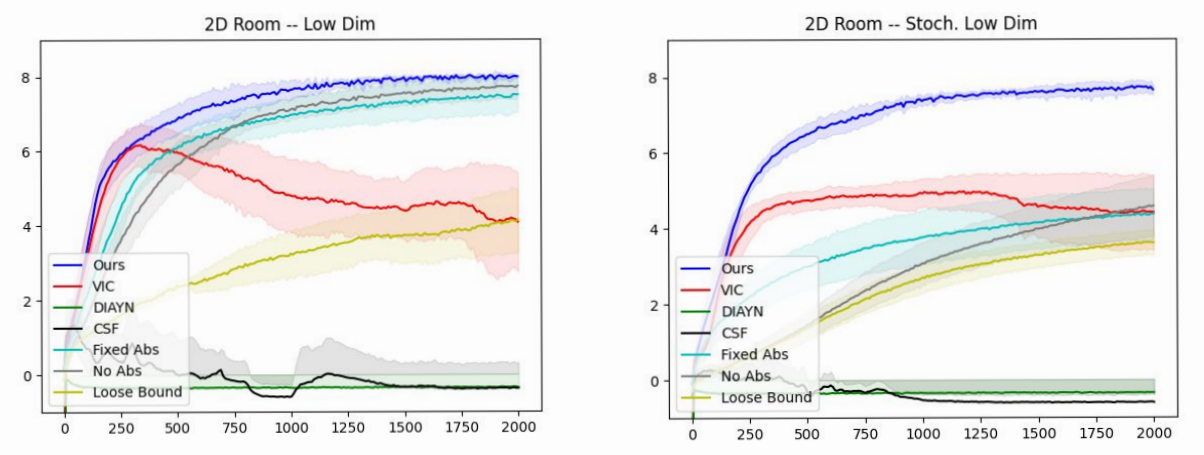


Figure 4.5: Learning curves for the low-dimensional tasks. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithm 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

skill-conditioned policies π_z. Further, Figures 4.5, 4.6, and 4.7 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 as can be seen by visuals of the different entropy terms in the mutual information objective such as $H(O_n)$, $H(Z)$, $H(Z|O_n)$. For instance, the left images in Figures 4.8 and 4.9 show 45 trajectories from randomly sampled skills from the Noisy 2D, Plus, Noisy Gray, and Push settings. The center images show 1,000 skill-terminating (x, y) positions for the agent (and object in Push) from randomly sampled skills. These images show the agent has learned skills that cover most of the observation space. The right images show sampled skills (colored squares) and for each skill, shows samples from the variational posterior. The tight variational posterior distributions confirm the agent is learning precise and distinct skills.

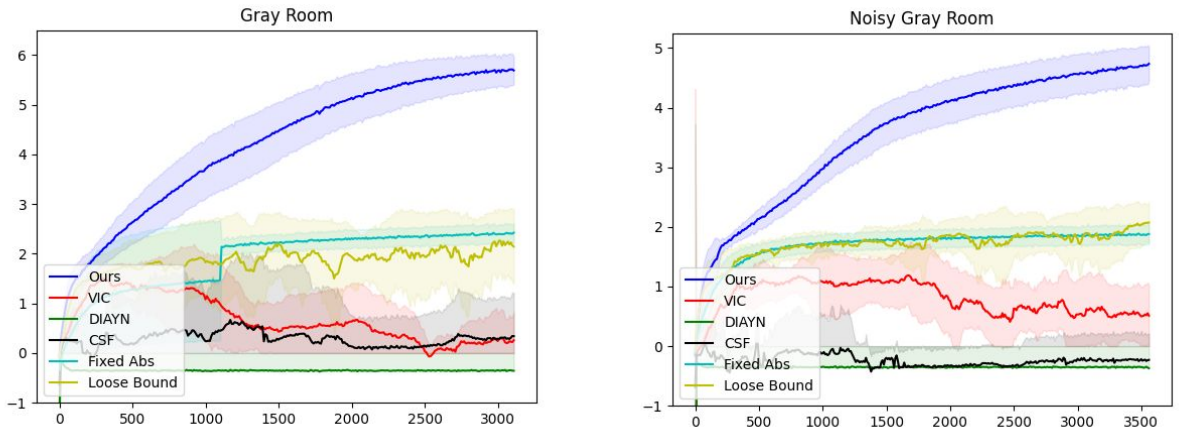


Figure 4.6: Learning curves for the regular and noisy grayscale rooms tasks. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithm 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

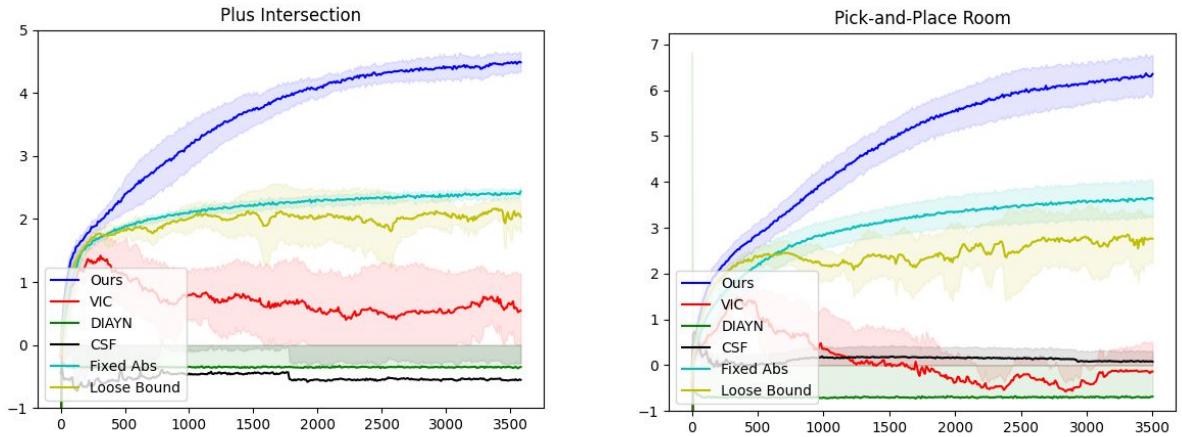


Figure 4.7: Learning curves for the plus intersection and push tasks in the first set of experiments. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithm 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

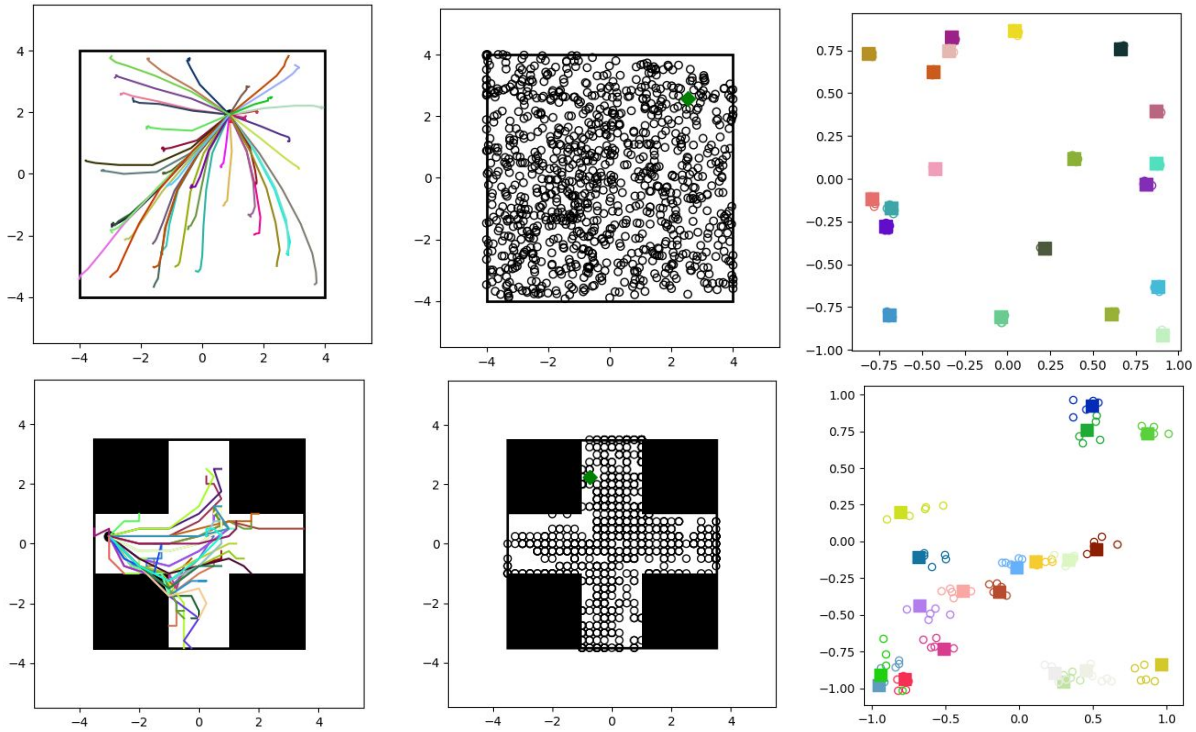


Figure 4.8: Some qualitative results from the Noisy 2D domain (top row) and Plus Intersection domain (bottom row). The left column shows the trajectories from a single starting observation produced by 45 randomly sampled skills. The center column shows the skill-terminating (x, y) positions from 1000 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.

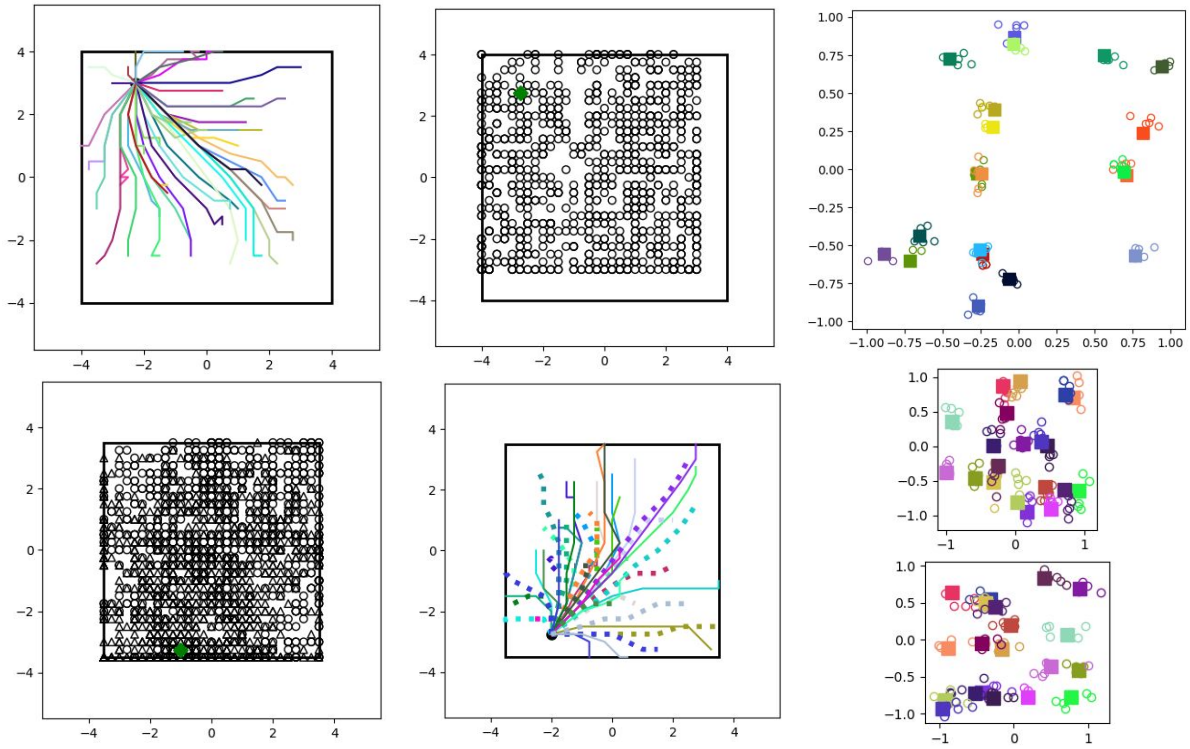


Figure 4.9: Some qualitative results from the Noisy Gray domain (top row) and Push domain (bottom row). The left column shows the trajectories from a single starting observation produced by 45 randomly sampled skills. The center column shows the skill-terminating (x, y) positions from 1000 randomly sampled 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.

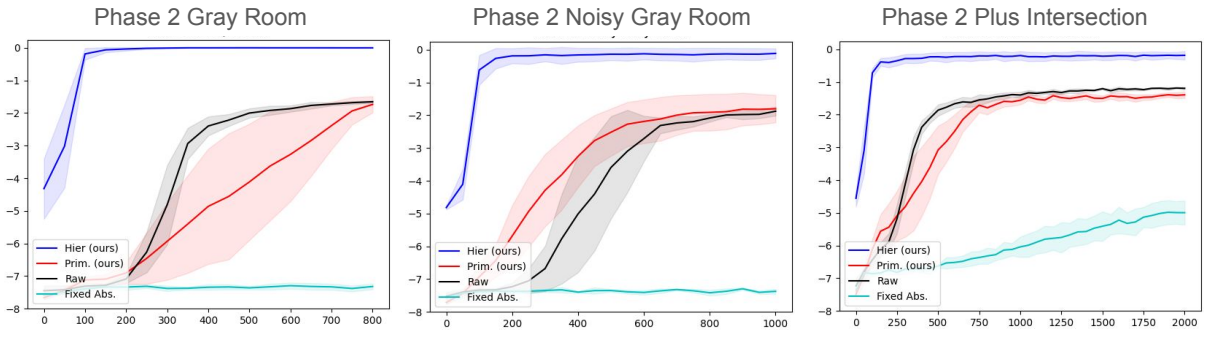


Figure 4.10: Learning curves for the phase 2 experiments. The x-axis shows the number of updates to the goal-conditioned policy and the y-axis shows the cumulative reward. The hierarchical policy should achieve lower cumulative reward as a result of the particular shortest path reward used and its temporally extended actions.

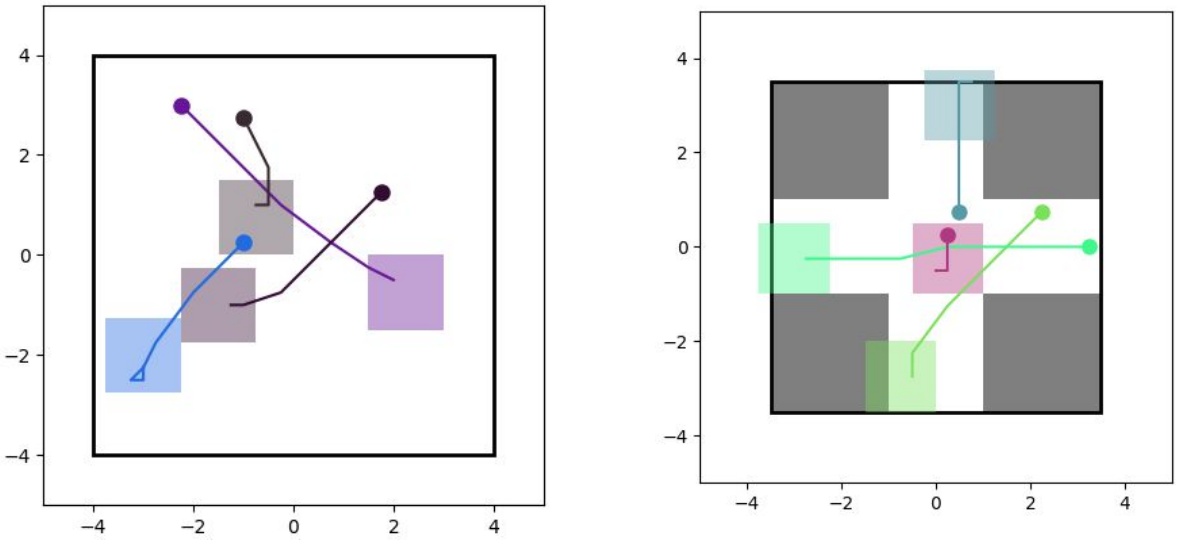


Figure 4.11: Phase 2 goal-conditioned trajectories for the grayscale room (Left) and Plus Intersection domains (Right) for the algorithm that learns a goal-conditioned policy outputting primitive actions and is conditioned on the learned representation space. Shaded regions are the episode goal and the line is the trajectory produced by the goal-conditioned policy.

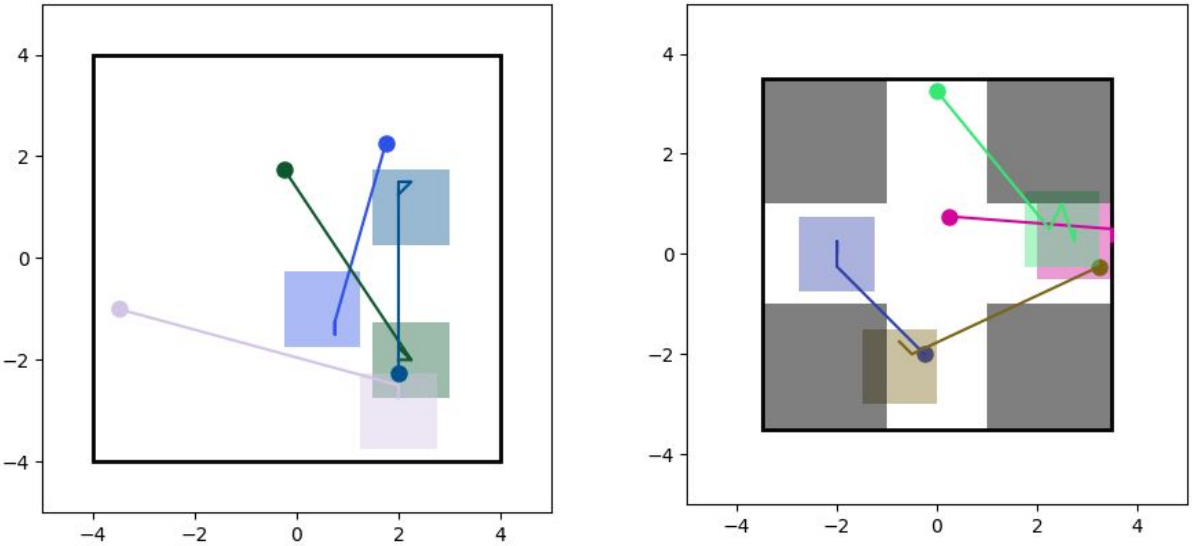


Figure 4.12: Phase 2 goal-conditioned trajectories for the grayscale room (Left) and Plus Intersection domains (Right) for the algorithm that learns a goal-conditioned policy outputting skills using the learned representation space and skills from pretraining. Shaded regions are the episode goal and the line is the trajectory produced by the goal-conditioned policy.

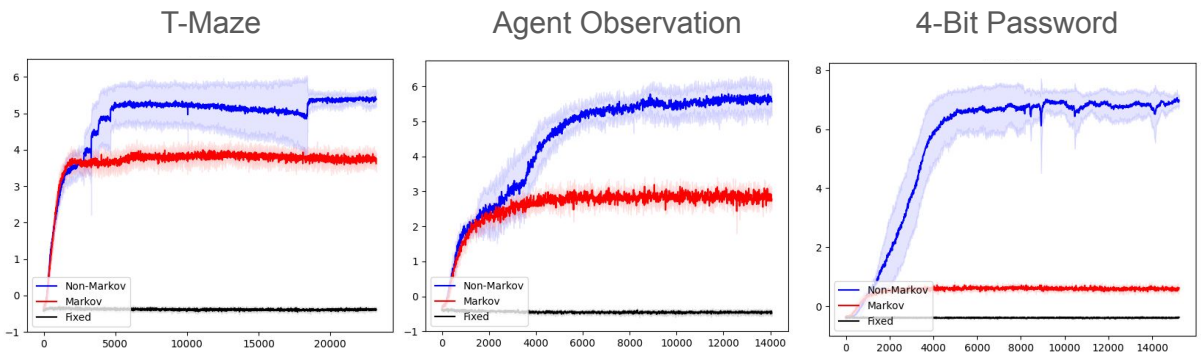


Figure 4.13: Average skillset size (in nats) vs. number of algorithm iterations. Skillset size is measured using variational mutual information. Mean and 1 std. of error computed with 5 random seeds.

Fully Observable Downstream RL Tasks

In addition to learning large skillsets, the second set of experiments in downstream RL tasks provides evidence that our approach can learn sufficient statistics of observations as the theory suggests. Figure 4.10 provides the learning curves for the second set of experiments where agents were required to learn goal-conditioned policies to target certain regions of the observation space. Figures 4.11 and 4.12 provide visuals of the goal-conditioned trajectories. Per Figure 4.10, 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 representations close to sufficient statistic representations. 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 evidence that empowerment is more effective at learning representations suitable for reinforcement learning than some randomly initialized function.

Partially Observable Skill Discovery Results

Figure 4.13 plots the results for all algorithms in the three domains. The y-axis measures skillset size using average variational mutual information. Note that skillset

size is measured in logarithmic units (nats). The x-axis reflects the number of iterations through algorithm 2. Our approach is the "Non-Markov" line (blue); the fully observable ablation of our approach, which follows algorithm 2 but trains an observation encoder, is shown by the "Markov" line (red); and the approach that uses algorithm 2 but does not update the representation learning actor-critic is the "Fixed" line (black).

Our approach successfully learns large skillsets in all domains. In T-Maze, Agent Observation, and 4-Bit Password our approach learned skillsets containing 5.4 nats (~ 220 skills), 5.7 nats (~ 300 skills), and 7.0 nats ($\sim 1,100$ skills) of skills, respectively, in around 5,000 iterations of Algorithm 2. These skillset sizes were 5.2x, 16.4x, and $555.6x$ larger than the performance of the fully observable ablation of our approach. The Fixed comparison was not able to learn a meaningful skillset in any domain. The significant outperformance relative to the comparisons shows the importance of learning representations of histories that preserve information in non-Markov settings.

For additional evidence on the successful performance of our approach, we also provide visuals of the different entropy terms included in the symmetric definitions of $I(Z; O_n)$: $H(O_n),\ H(O_n|Z), H(Z),\ H(Z|O_n)$ for all tasks in Figures 4.14, 4.15, and 4.16. $H(O_n)$, which represents the distribution of skill-terminating observations produced by the trained skillset, is visualized with both agent trajectories and by marking skill-terminating observations. In all settings, the agents learn a skillset that mostly covers the observation space that can be targeted. For instance, in T-Maze, the agent learns skills that can target most of the larger hallway and never attempts to move to the non-existent hallway. Similarly, in Agent Observation and 4-Bit Password, nearly all skills first pass through the bottleneck (i.e., move to (x, y) goal in Agent Observation or enter the correct password in 4-Bit Password) and then target a large area of observations. The $H(O_n|Z)$ visuals, which show the observations targeted by specific skills, show that each skill targets a precise region of the observation space. Similarly, the visualizations of $H(Z|O_n)$, which show samples of the variational posterior $q_\psi(z|o_n)$ forming tight distributions around the

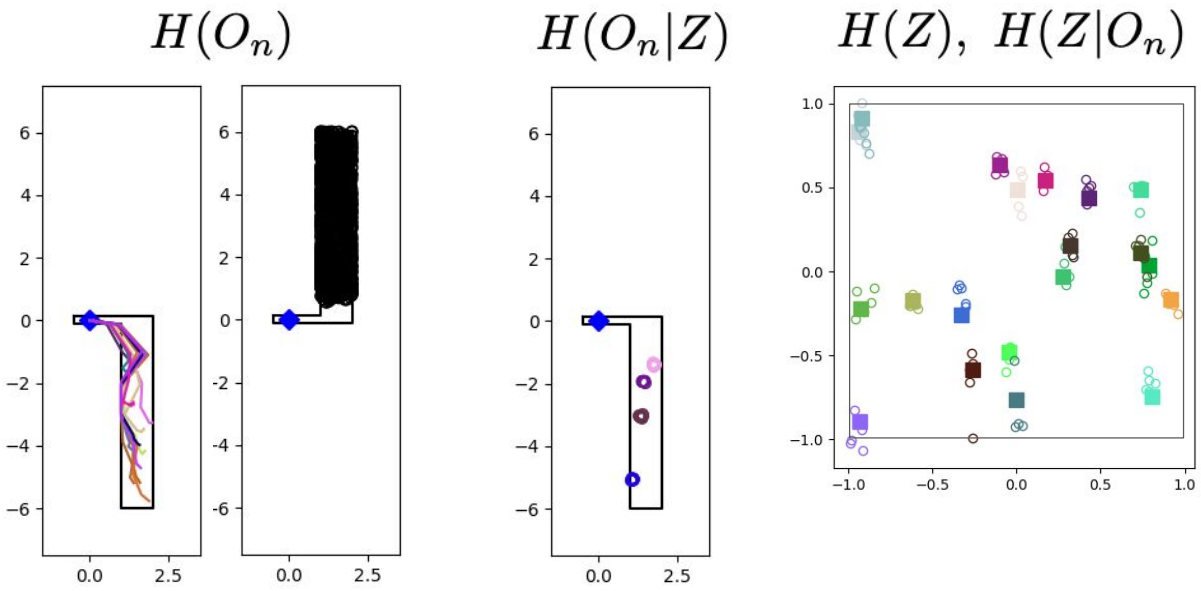


Figure 4.14: Entropy visualizations for the T-Maze domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The left most figure shows agent trajectories from randomly selected skills $z \sim p(z)$. The adjacent figure marks skill-terminating observations from 1000 randomly selected skills. As the images show, the skills target most of the larger hallway regardless of which direction the hallway opens. $H(O_n|Z)$ visualizes the skill-terminating observations from four randomly selected skills, showing five observations for each skill. The figure shows that each skill targets a precise regions of the (x, y) space. The right most figure visualizes both $H(Z)$ and $H(Z|O_n)$ by showing skills (filled squares) sampled from the fixed, uniform $p(z)$ (shown by the inner black square) as well as sampled from the variational posterior $q(z|o_n)$. Note that the samples from the variational posterior form tight distributions around the executed skill. All the entropy visualizations confirm the agent has learned a large set of distinct skills as the skillset covers a larger area of observations and each skill targets a precise region of the observation space. In addition, the very different policies that occur when the hallway opens north and south shows that the RNN is able to disentangle histories that do not have the same distribution over underlying states.

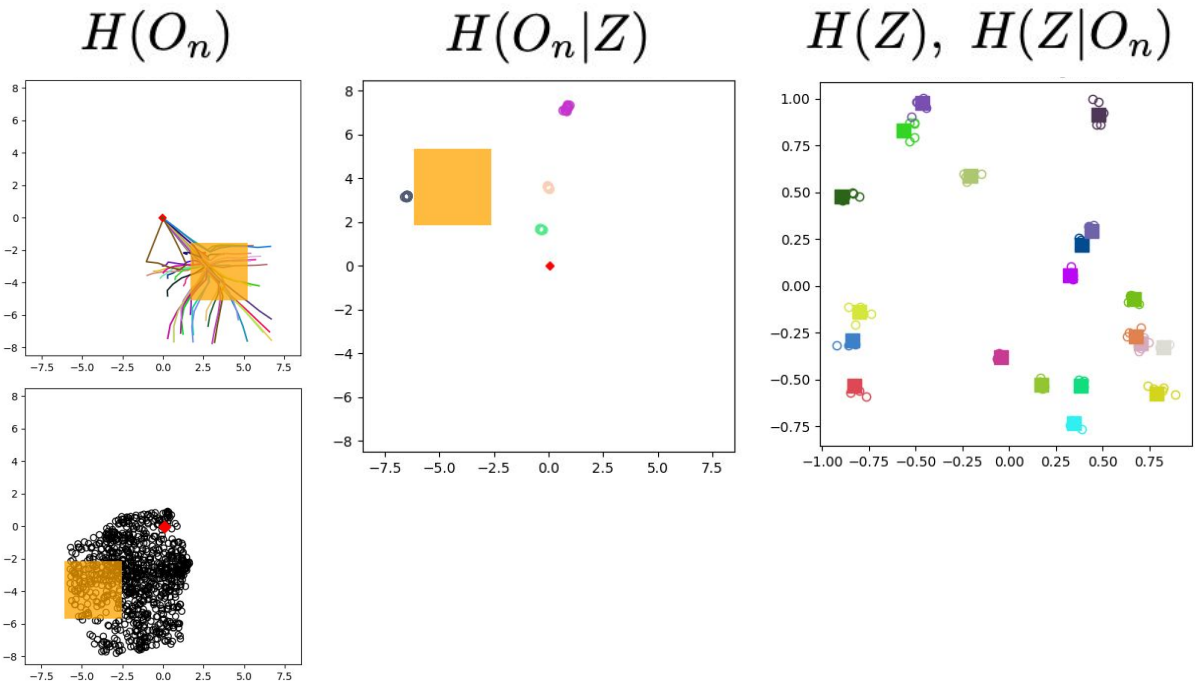


Figure 4.15: Entropy visualizations for the Decoy Agent domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The top figure on the left column shows agent trajectories from randomly selected skills $z \sim p(z)$. The bottom figure on the left column marks skill-terminating observations from 1000 randomly selected skills. Per the graphics, the agent's skillset first moves to the goal (orange square) and then targets a wide area of the observation space. $H(O_n|Z)$ visualizes the skill-terminating observations from four randomly selected skills, showing five observations for each skill. This figure demonstrates that each skill targets a precise region of the (x, y) space. The right most figure visualizes both $H(Z)$ and $H(Z|O_n)$ by showing skills (filled squares) sampled from the fixed, uniform $p(z)$ (in the shape of a 2D box in the ranges [-1,1]) as well as samples from the variational posterior $q(z|o_n)$. Note that the samples from the variational posterior form tight distributions around the executed skill. All the entropy visualizations confirm the agent has learned a large set of distinct skills as the skillset covers a larger area of observations and each skill targets a precise region of the observation space. In addition, the different policies that occur when the goal changes shows that the RNN is able to disentangle histories that do not have the same distribution over underlying states.

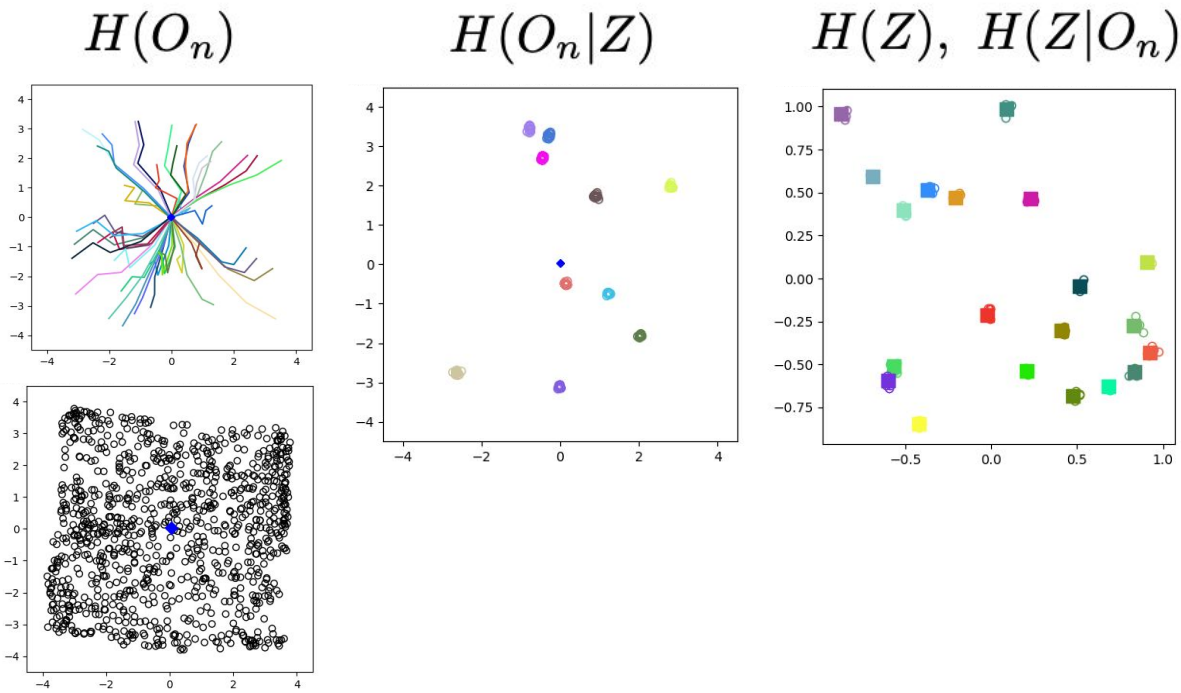


Figure 4.16: Entropy visualizations for the 4-Bit Password domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The top figure on the left column shows agent trajectories from randomly selected skills $z \sim p(z)$. The bottom figure on the left column marks skill-terminating observations from 1000 randomly selected skills. Note that because each skill lasts eight actions and because the agent is frozen in place during the first four actions when it outputs a password, the agent can not move more than four units in any direction. Per the figures, most skills in the skillset are executing the correct password and then target a wide range of observations. $H(O_n|Z)$ visualizes the skill-terminating observations from four randomly selected skills, showing five observations for each skill. Per the figure, each skill is targeting a precise region of the (x, y) space. The right most figure visualizes both $H(Z)$ and $H(Z|O_n)$ by showing skills (filled squares) sampled from the fixed, uniform $p(z)$ (in the shape of a 2D box in the ranges [-1,1]) as well as samples from the variational posterior $q(z|o_n)$ that form tight distributions around the executed skill. The entropy visualizations confirm the agent has learned a large set of distinct skills as the skillset covers a larger area of observations and each skill targets a precise region of the observation space. In addition, the fact that the agent can still learn skills that cover the available observation space despite the randomly selected password shows that the RNN is able to disentangle histories that do not have the same distribution over underlying states.

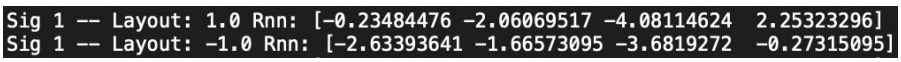


Figure 4.17: RNN representations in the T-Maze for when agent is in same starting (x, y) position but received a different binary signal for the direction of the T-Maze 15 timesteps earlier. Early in training these representations were virtually identical meaning that early in training the agent was not able to disentangle these histories. But later in training, the RNN was able to separate these histories as shown by some of the dimensions of the 4-dim vector that are more than 2 units apart.

executed skill, further demonstrate that the agent is learning diverse skillsets.

With respect to representation learning, the consistently large skillsets shown in the entropy visualizations despite the underlying state randomly changing every episode shows the RNN is able to disentangle histories representing different distributions of underlying states. If the agent was failing at preserving information, such as the agent assigning similar representations to the 16 different possible passwords in 4-Bit Password, the agents would not have been able to learn large skillsets. In Figures 4.17, 4.18, 4.19, and 4.20, we also show some of the actual learned representations of the trained agents for different underlying states. In all tasks, early in training the agent was not able to disentangle the histories representing different underlying states. For instance, in T-Maze, the agent would assign nearly the exact same representation when the agent was in some (x, y) position regardless of the signal the agent had received about the direction the hallway opened. But as training continued, agents in all tasks were able to correctly separate histories.

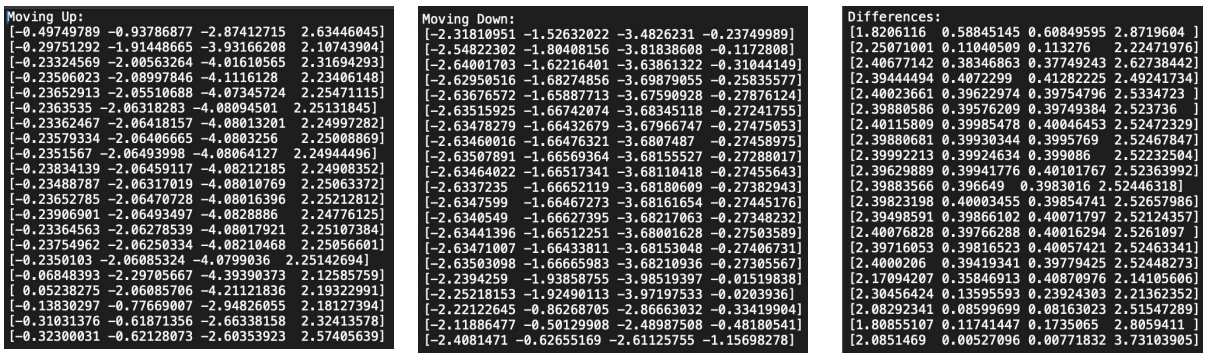


Figure 4.18: Figure shows sequences of RNN representations in T-Maze during episodes where the agent first receives the binary signal, then remains frozen for 15 timesteps, and then executes 5 actions to enter the larger hallway. The left table shows the RNN sequence for when the hallway opens up; the center table shows the RNN sequence for when the hallway opens down; and the right table shows the differences between the two. The consistent large differences in the RNN sequences after training shows that not only is the agent able to assign a different representation after the agent receives a different binary signal, but is able to maintain that difference both while remaining frozen in place and once the agent starts to move towards and enter the larger hallway.

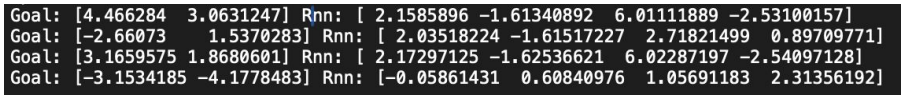


Figure 4.19: Figure shows the different RNN representations for different episode goals indicated by the movements of the first agent in the Decoy Agent setting. These representations are sampled after the first 10 timesteps of the episode when the agent has just finished observing the two agents in the environment. Early in training, these representation were similar regardless of the goal observed. But after training, the agent's RNN was able to learn different representation for different goals. For instance, the top and bottom lines show goals in the top right and bottom left, respectively. The agent was able to learn a representation that nearly differs by 5 units along two of the dimensions.

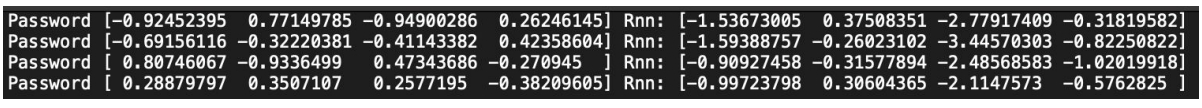


Figure 4.20: Figure shows the different RNN representations for different episode passwords in the 4-Bit password setting. These representations were sampled immediately after the fourth timestep when the agent had been given the last bit of the password. Note that, as shown in the table, the passwords provided to the agent were four-dim vectors of continuous numbers. Instead of bits, each dimension was either in the range $[0.25, 1.]$ or $[-1., -0.25]$. Early in training, these representation were nearly identical regardless of the password provided. But after training, the agent's RNN was able to learn different representation for different passwords.

CHAPTER 5

Reducing State Uncertainty from Partial Observability using Empowerment

The prior chapter showed that agents can address state uncertainty caused by poor representation learning with empowerment. Yet even if agents are able to generate sufficient statistic representations that provide as much information about the underlying state as the agent's full history, agents may still face significant state uncertainty that can lower the size of their skillsets. We refer to this remaining uncertainty, when the probability of the underlying state given an agent's history is still greater than zero, as state uncertainty caused by partial observability.

Agents in realistic settings continually face significant uncertainty from partial observability that affects the size of their skillset. For instance, given their limited view of the world, agents may enter some room and not know where certain objects are located as other agents may have moved them. Or the agent may see some device that may be slightly different than other devices the agent has used, and thus the agent is uncertain about certain properties of the device. In both of these situations when the agent does not

know where an object is located or an agent is unsure about certain important properties of a device, the agent's skillset may be smaller as it cannot definitively achieve observations that involve manipulating the object or effectively using the device.

To overcome state uncertainty caused by partial observability, agents can extend their histories. That is, they can execute actions in the environment and observe new observations that together may provide information about the features of the state they are uncertain about. Indeed, realistic settings are often filled with sources of information about the underlying state. For instance, if an agent is unsure where some object is located, it can on its own probe different locations in the environment. Or if an agent is unsure of some properties of a device, it can interact with the device (e.g., touch different buttons) to seek information about these properties. But there are also more efficient ways than trial-and-error in which agents can gain information from their environment. For instance, agents can watch other agents act in the world. If an agent is unsure of some properties of a device (e.g., what button turns on a device), the agent can watch another agent use the device. More sophisticated agents can seek information even more efficiently by asking other agents (e.g., chatbots) questions or searching the internet (e.g., YouTube).

The purpose of this chapter is to demonstrate that empowerment encourages information seeking behaviors that can reduce state uncertainty caused by partial observability. Specifically, when empowerment is maximized with respect to a behavior policy that interacts with the environments, agents are encouraged to seek information about the underlying state. This should be an intuitive result. If there is some uncertainty about features of the underlying state that is depressing the size of an agent's skillset, an objective that encourages actions to grow the size of an agent's skillset should encourage agents to take actions that reduce this uncertainty.

In this chapter, we provide theoretical and empirical results to defend this hypothesis. We provide a theorem that shows that empowerment encourages information seeking

policies because if an agent is considering two behavior policies and one policy provides more information about the underlying state than the other, then the average empowerment by the information seeking policy will be at least as high as the other policy (and higher under some easily achievable conditions). We also demonstrate in a few simple settings that when empowerment is used as a reward and an agent's policy is trained to maximize this reward, the agent will seek information.

This chapter also introduces a longer horizon version of empowerment that can be used as a reward for decision making. Computing empowerment over long horizons (i.e., maximizing the mutual information between skills and observations far into the future) is difficult for a few reasons. First, it is hard to learn a single skill-conditioned policy that can target distant observations. Second, these skill-conditioned policies may need more parameters which would require more compute for training. Third, learning a world model that can predict far into the future, a required ingredient for computing long-horizon empowerment, is difficult to train. We show a more practical approach may be to use a different version of empowerment in which the mutual information is defined to be between skills and sequences of equally-spaced sufficient statistic representations. This definition of mutual information is helpful because it can be maximized using dynamic programming so that agents still only need to learn short-horizon mutual information terms between skills and observations.

5.1 Theoretical Analysis

This section shows that empowerment provably encourages information seeking through environment interaction and thus can reduce state uncertainty caused by partial observability. Theorem 4 shows that if there are two behavior policies and one policy provides more information about the underlying state than the other, then the average empowerment produced by the more informative policy will be at least as large as the other policy, and thus will be preferred by an empowerment-maximizing objective.

Theorem 4. *Let π^a and π^b be the behavior policies executed from the same same sufficient statistic representation x, and let $p(x^a, s^a, x^b, s^b)$ be the joint distribution of the policy-terminating sufficient statistic representations and states produced by the two behavior policies. If (i) for all x^a, the marginal distribution $p(s^a|x^a) = p(s^b|x^a)$, (ii) $I(X^b; S^b|X^a) > 0$, and (iii) $X_t^a \rightarrow X_t^b \rightarrow S_t$ form a Markov chain when conditioned on x^a, x^b for $t = 0, 1, \ldots, n-1$, then $\mathbb{E}_{x^a \sim p(x^a)}[\mathcal{E}(x^a)] \leq \mathbb{E}_{x^b \sim p(x^b)}[\mathcal{E}(x^b)]$.*

Proof.

$$\mathbb{E}_{x_0^a \sim p(x_0^a)}[\mathcal{E}(x_0^a)] = \mathbb{E}_{x_0^a \sim p(x_0^a)}[I(Z; O_n|x_0^a, \pi_z^{a,*})] \tag{5.1}$$

$$\leq \mathbb{E}_{x_0^a \sim p(x_0^a), x_0^b \sim p(x_0^b|x_0^a)}[I(Z; O_n|x_0^a, x_0^b, \pi_z^{a,*})] \tag{5.2}$$

$$\leq \mathbb{E}_{x_0^b \sim p(x_0^b)}[I(Z; O_n|x_0^b, \pi_z^b)] \tag{5.3}$$

$$\leq \mathbb{E}_{x_0^b \sim p(x_0^b)}[I(Z; O_n|x_0^b, \pi_z^{b,*})] \tag{5.4}$$

$$= \mathbb{E}_{x_0^b \sim p(x_0^b)}[\mathcal{E}(x_0^b)] \tag{5.5}$$

□

Line 5.1 applies the definition of the empowerment of a sufficient statistic representation. The RNN parameters, which form a function that outputs sufficient statistic representations, are ignored as they are fixed in this proof. $\pi^{a,*}$ represents the mutual information maximizing skill-conditioned policy parameters when the starting sufficient statistic representation is x_o^a.

The lower bound in line 5.2 applies the convexity property of mutual information with respect to the channel distribution (Cover and Thomas, 2006; Capdepuy, 2011). In our case, if $I(X^b; S|X^a) > 0$), then the channel distribution $p(o_n|x_0^a, \pi_z^{a,*}, z)$ is a weighted mixture of channels $p(o_n|x_0^a, x_0^b, \pi_z^{a,*}, z)$ (i.e., $p(o_n|x^a, \pi_z^{a,*}, z) = \int_{x_0^b} p(x_0^b|x_0^a) p(o_n|x_0^a, x_0^b, \pi_z^{a,*}, z)$) and so the mutual information of the mixed channel $I(Z; O_n|c_0^a, \eta_a, \pi_z^{a,*})$ is upper bounded by the average mutual information of the individual channels $\mathbb{E}_{x_0^b \sim p(x_0^b|x_0^a)}[I(Z; O_n|x_0^a, x_0^b, \pi_z^{a,*})]$.

Line 5.3 removes the dependence on sufficient statistic representations x_t^a for $t \geq 0$ by (i) using the sufficient statistic representations x_t^b for $t \geq 0$ and (ii) replacing $\pi_z^{a,*}$ with a specific skill-conditioned policy π_z^b. As we will show, this will replace each mutual information term, $I(Z; O_n | x_0^a, x_0^b, \pi_z^{a,*})$, with a new mutual information term, $I(Z; O_n | x_0^b, \pi_z^b)$, that is at least as large. For each context x_0^b, π_z^b will be constructed as follows. For each x_0^b, find the tuple (x_0^a, x_0^b) with the largest $I(Z; O_n | x_0^a, x_0^b, \pi^{a,*})$ as there can be multiple representations x_0^a associated with the same x_0^b. Then, for each x_0^b let π_z^b be the skill-conditioned policy distribution $p(a_t | x_0^b, x_t^b) = \int_{x_t^a} p(x_t^a | x_0^a, x_0^b, x_t^b, s_t) p(a^t | x_0^a, x_0^b, x_t^b, x_t^a, s_t) = \int_{x_t^a} p(x_t^a | x_0^a, x_0^b, x_t^b) p(a^t | x_t^a)$, in which $p(a^t | c_t^a)$ is the probability specified by the skill-conditioned policy defined by $\pi_z^{a,*}$. That is, the skill-conditioned policy π_z^b will have the same distribution over actions as executed by $\pi_z^{a,*}$ when conditioned on the contexts x_t^b. Next, we show that for all (x_0^a, x_0^b), the original mutual information $I(Z; O_n | x_0^a, x_0^b, \pi_z^{a,*})$ is upper bounded by the new mutual information $I(Z; O_n | x^b, \pi_z^b)$.

We first show that for each mutual information term $I(Z; O_n | x_0^a, x_0^b, \pi_z^{a,*})$ from which π_z^b was constructed in the previous step, $I(Z; O_n | x_0^b, \pi_z^b) = I(Z; O_n | x_0^a, x_0^b, \pi_z^{a,*})$. That is, we replace the original mutual information term with an equivalent mutual information term. Given that the source distributions $p(z)$ are the same by definition as they are fixed, to show that the mutual information terms are the same, we need to show that the channel distributions $p(o_n | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(o_n | x_0^b, \pi_z^b, z)$ are the same for all (z, o_n) tuples. We show this below by proving by induction that the joint distributions $p(x_{t-1}^b, s_{t-1}, a_{t-1}, s_t, o_t, x_t^b | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_{t-1}^b, s_{t-1}, a_{t-1}, s_t, o_t, x_t^b | x_0^b, \pi_z^b, z)$ for $t = 1, \ldots, n$. Then because the joint distribution $p(x_{n-1}^b, s_{n-1}, a_{n-1}, s_n, o_n, x_n^b | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_{n-1}^b, s_{n-1}, a_{n-1}, s_n, o_n, x_n^b | x_0^b, \pi_z^b, z)$ at the last time step, the channel distributions are equal: $p(o_n | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(o_n | x_0^b, \pi_z^b, z)$.

The proof by induction goes as follows. In the base case at $t = 1$, the distribution $p(x_0^b | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_0^b | x_0^b, \pi_z^b, z)$ because the same x_0^b appears in the conditioning variables. $p(s_0 | x_0^a, x_0^b, \pi_z^{a,*}, z) = p(s_0 | x_0^b, \pi_z^b, z)$ because $p(s_0 | x_0^a, x_0^b, z) = p(s_0 | x_0^b, z)$

per the Markov chain assumption in the theorem statement. $p(a_0|x_0^a, x_0^b, \pi_z^{a,*}, z, s_0) = p(a_0|x_0^a, x_0^b, \pi_z^{a,*}, z) = p(a_0|x_0^b, \pi_z^b, z)$ using the definition of π_z^b. $p(s_1, o_1|x_0^a, x_0^b, \pi_z^{a,*}, z, s_0, a_0) = p(s_1, o_1|x_0^b, \pi_z^b, z, s_0, a_0)$ as the next state and observation only depends on s_0 and a_0 and is independent of the other variables. Lastly, $p(x_1^b|x_0^a, x_0^b, \pi_z^{a,*}, z, s_0, a_0, o_1) = p(x_1^b|x_0^b, \pi_z^b, z, s_0, a_0, o_1)$ because the next sufficient statistic representation x_1^b only depends on x_0^b, a_0, o_1, which are the same in both cases. Thus, the base case of the induction proof is true as $p(x_0^b, s_0, a_0, s_1, o_1, x_1^b|x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_0^b, s_0, a_0, s_1, o_1, x_1^b|x_0^b, \pi_z^b, z)$.

Assuming the proof holds through $t = k-1$, then at step $t = k$, $p(x_{k-1}^b, s_{k-1}|x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_{k-1}^b, s_{k-1}|x_0^b, \pi_z^b, z)$ because the joint distribution $p(x_{k-2}^b, s_{k-2}, a_{k-2}, s_{k-1}, o_{k-1}, x_{k-1}^b|x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_{k-2}^b, s_{k-2}, a_{k-2}, s_{k-1}, o_{k-1}, x_{k-1}^b|x_0^b, \pi_z^b, z)$. $p(a_{k-1}|x_0^a, x_0^b, \pi_z^{a,*}, z, x_{k-1}^b, s_{k-1}) = p(a_{k-1}|x_0^a, x_0^b, \pi_z^{a,*}, z, x_{k-1}^b$ $p(a_{k-1}|x_0^b, \pi_z^b, z, x_{k-1}^b)$ using the definition of π_z^b. Again, $p(s_k, o_k, x_k^b|x_0^a, x_0^b, \pi_z^{a,*}, z, x_{k-1}^b, s_{k-1}, a_{k-1}) = p(s_k, o_k, x_k^b|x_0^b, \pi_z^b, z, x_{k-1}^b, s_{k-1}, a_{k-1})$ as s_k and o_k only depends on s_{k-1} and a_{k-1} and x_k^b only depends on x_{k-1}^b, a_{k-1}, o_k. Thus, the induction proof holds through step $t = k$ as $p(x_{k-1}^b, s_{k-1}, a_{k-1}, s_k, o_k, x_k^b|x_0^a, x_0^b, \pi_z^{a,*}, z) = p(x_{k-1}^b, s_{k-1}, a_{k-1}, s_k, o_k, x_k^b|x_0^b, \pi_z^b, z)$.

Thus, $I(Z; O_n|x_0^b, \pi_z^b) = I(Z; O_n|x_0^a, x_0^b, \pi_z^{a,*})$ for those (x_0^a, x_0^b) tuples from which π_z^b was constructed. For the other smaller $I(Z; O_n|x_0^a, x_0^b, \pi_z^{a,*})$ terms that were not used to construct π_z^b, these will also be replaced by the $I(Z; O_n|x_0^b, \pi_z^b)$ terms that are at least as large as $I(Z; O_n|x_0^a, x_0^b, \pi_z^{a,*})$. If the replacement mutual information terms are larger, then the inequality in line 5.3 will be replaced by a strictly less than.

In line 5.4, the lower bound results from replacing the skill-conditioned policy π_z^b with the optimal skill-conditioned policy $\pi_z^{b,*}$ for the specific x_0^b representation. If this replacement produces larger mutual information, then the inequality becomes a strictly less than. The final line 5.5 uses the definition of the empowerment of representation x_0^b.

5.2 A More Practical Longer Horizon Empowerment

Next we introduce a longer-horizon version of empowerment because maximizing the mutual information between skills and observations that are far into the future is difficult for a few reasons. First, it is hard to learn a single skill-conditioned policy that can target distant observations. Second, these skill-conditioned policies may need more parameters which would require more compute for training. Third, training a world model that can predict far into the future, which would be a required ingredient to performing long-horizon empowerment, would be challenging. We show that a more practical approach may be to use a different version of empowerment in which the mutual information is defined to be between skills and sequences of K equally-spaced sufficient statistic representations:

$$I(Z; X_n, X_{2n}, X_{3n}, \ldots, X_{Kn} | c, \pi_z). \tag{5.6}$$

The key benefit of this formulation of mutual information is that it can be maximized using dynamic programming. That is, it can be maximized by solving small subproblems that only require optimizing short horizon mutual information terms between skills and observations.

Consider the following mutual information term between skills Z^{2n} and two equally-spaced sufficient statistic representations X_n, X_{2n}:

$$I(Z^{2n}; X_n, X_{2n} | x_0, \pi_z^{2n}), \tag{5.7}$$

in which Z^{2n} denotes that the skills last for $2n$ primitive actions. Using the Chain Rule property of mutual information 2.4, this mutual information can be broken down into the

mutual information terms between skills and observations n steps into the future:

$$\begin{aligned} I(Z^{2n}; X_n, X_{2n}|x_0, \pi_z^{2n}) &= I(Z^{2n}; X_n|x_0, \pi_z^{2n}) + I(Z^{2n}; X_{2n}|x_0, X_n, \pi_z^{2n}) \\ &= I(Z^{2n}; X_n|x_0, \pi_z^{2n}) + I(Z^{2n}; X_{2n}|X_n, \pi_z^{2n}) \end{aligned} \tag{5.8}$$

In line 5.8, the x_0 conditioned variable in the second term is removed because the channel distribution is independent of x_0 given sufficient statistic representation x_n: $p(x_{2n}|x_0, x_n, \pi_z) = p(x_{2n}|x_n, \pi_z)$. Next, let's assume that X_n and X_{2n} are generated by executing nested skills, which will work as follows. To start, a higher level skill z^{2n} is sampled from the fixed distribution $z^{2n} \sim p(z^{2n})$. Then, one component of the skill-conditioned policy π_z^{2n} will propose some skill z^n given x_0 and z^{2n}. That skill will be passed to a skill-conditioned policy π_z^n that then executes the skill for n-primitive actions. After this skill completes, the second component of skill-conditioned policy π_z^{2n} will propose another skill z^n, which again is passed down to a skill-conditioned policy π_z^n which executes the skill for n steps. Note that below we will show that it is trivial to design that second component of π_z^{2n} to achieve high $I(Z^{2n}; X_{2n}|X_n, \pi_z^{2n})$.

With this nested structure for producing trajectories, using the data processing inequality (Cover and Thomas, 2006), upper bounds of both mutual information terms can be achieved by removing the outer skill-conditioned policy π_z^{2n} and then using the mutual information maximizing n-step skill-conditioned policies $\pi_z^{n,*}$:

$$\begin{aligned} I(Z^{2n}; X_n|x_0, \pi_z^{2n}) + I(Z^{2n}; X_{2n}|X_n, \pi_z^{2n}) &\leq I(Z^n; X_n|x_0, \pi_z^n) + I(Z^n; X_{2n}|X_n, \pi_z^n) \\ &\leq I(Z^n; X_n|x_0, \pi_z^{n,*}) + I(Z^n; X_{2n}|X_n, \pi_z^{n,*}) \end{aligned} \tag{5.9}$$

That is, the size of the agent's skillset over $2n$ actions, $I(Z^{2n}; X_n, X_{2n}|x_0, \pi_z^{2n})$, is upper bounded by the sum of two n-step empowerment terms. Note that a lower bound to these two mutual information terms in 5.9 (using the Data Processing Inequality), is the

maximum mutual information between skills and *observations*, which is the type of mutual information this thesis has focused on:

$$I(Z^n; O_n|x_0, \pi_z^n) + I(Z^n; O_{2n}|X_n, \pi_z^n) \leq I(Z^n; X_n|x_0, \pi_z^n) + I(Z^n; X_{2n}|X_n, \pi_z^n) \quad (5.10)$$

Thus, the process of computing the mutual information between skills and the sequence X_n, X_{2n} for a single skill-conditioned policy π_z^{2n} requires first solving the simpler subproblem of maximizing n-step mutual information.

Next, the mutual information $I(Z^{2n}; X_n, X_{2n}|x_0, \pi_z^{2n})$ needs to be maximized with respect to the two components of the skill-conditioned policy π_z^{2n}. One component outputs the first n-step skill z^n, while the second component outputs the second z^n skill. We first discuss how the second component of π_z^{2n} can be hardcoded to achieve a high value of $I(Z^{2n}; O_{2n}|x_n, \pi_z^{2n}) = I(Z^n; O_{2n}|x_n, \pi_z^{n,*})$, in which the latter $I(Z^n; O_{2n}|x_n, \pi_z^{n,*})$ is the n-step empowerment that was learned when n-step skillsets were trained. To achieve this high mutual information, the following procedure should be followed. First, the high level skill z^{2n} will be a concatenation of two z^n skills, in which each $z^n \sim p(z^n)$ is randomly sampled from the distribution over skills. Second, the first component of the skill-conditioned policy π_z^{2n} only takes as input the first z^n vector in z^{2n}. Then when the agent is in representation x_n, the second component simply proposes a skill equal to the second z^n vector in z^n. Given that (i) $p(o_n|x_n, z^{2n}) = p(o_n|x_n, z^n)$ and that (ii) this procedure induces a uniform $p(z^n)$ for the second skill, then $I(Z^{2n}; X_{2n}|x_n, \pi_z^{2n}) = I(Z^n; O_{2n}|x_n, \pi_z^{n,*})$.

The remaining optimization step is to maximize $I(Z^{2n}; X_n, X_{2n}|x_0, \pi_z^{2n})$ with respect to the first component of π_z^{2n}. As we show next, this maximization is very similar to maximizing n-step mutual information with respect to an n-step skill-conditioned policy. To derive the objective, we can first simplify the mutual information term to take into

account the way the second component of the skill-conditioned policy is designed:

$$I(Z^{2n}; X_n, X_{2n}|x_0, \pi_z^{2n}) = I(Z^{2n}; X_n|x_0, \pi_z^{2n}) + I(Z^{2n}; X_{2n}|X_n, \pi_z^{2n}) \tag{5.11}$$

$$\geq I(Z^{2n}; O_n|x_0, \pi_z^{2n}) + I(Z^{2n}; O_{2n}|X_n, \pi_z^{2n}) \tag{5.12}$$

$$= I(Z^{2n}; O_n|x_0, \pi_z^{2n}) + V^1(x_n) \tag{5.13}$$

$$\tag{5.14}$$

Line 5.11 applies the chain rule discussed earlier. The inequality in line 5.12 where observations replace sufficient statistic representations in the mutual information results from the data processing inequality. Line 5.13 accounts for the way the second component of π_z^{2n} is designed and replaces the second mutual information term with a scalar value $V^1(x_n) = I(Z^n; O_n|x_n, \pi_z^n)$, which is the mutual information learned with n-step skillsets were learned. A superscript of 1 is used to denote that the remaining skills are of length $1*n$ primitive actions. With this updated mutual information term, the mutual information objective to maximize with respect to π_z^{2n} is:

$$\max_{\pi_z^{2n}} I(Z^{2n}; O_n|x_0, \pi_z^{2n}) + V^1(x_n) \tag{5.15}$$

This is similar to the objective of learning a skill-conditioned to maximize n-step mutual information except now agents also have to take into account the size of the skillsets they can execute from future representations x_n as measured by the scalar $V^1(x_n)$. That is, when converting some high-level skills z^{2n} to a lower level skills z^n, π_z^{2n} needs to consider both the diversity of the O_n observation but also the diversity of the O_{2n} observations that can be achieved from representation x_n, which reflected in the $V^1(x_n)$ value. It could be the case π_z^{2n} outputs redundant skills z^n for different z^{2n} because it wants to have more skills that begin in representation x_n so that more observations O_{2n} can be targeted.

This mutual information maximization can be handled with the same actor-critic architecture discussed in Chapter 3. The actor maps sufficient statistic representations

to a vector of parameters representing the skill-conditioned policy π_z^{2n}. But now the parameter-specific critics will approximate $I(Z^{2n}; O_n|x_0, \pi_z^{i,2n}) + V^1(x_n)$, in which $\pi_z^{i,2n}$ is a scalar representing the skill-conditioned policy where all parameters $j \neq i$ are set to the greedy value of $f_\lambda(x_0)[j]$, except for the i-th parameter, which takes the value of $\pi_z^{i,2n}$. In addition, parameter-specific variational posteriors will be used to approximate the true posterior $p(z^{2n}|x_0, \pi_z^{i,2n}, o_n)$ in the mutual information term $I(Z^{2n}; O_n|x_0, \pi_z^{2n})$.

This approach can then be scaled to maximize the mutual information between skills and K-length sequences of sufficient statistic representations (i.e., $I(Z^{Kn}; X_n, X_{2n}, X_{3n}, \ldots, X_{Kn}|c, \pi_z^{Kn})$) using a dynamic programming approach. At each iteration $k = 2, \ldots, K$, set $V^{k-1}(x_n)$ equal to $I(Z^{(k-1)n}; O_{2n}|x_n, \pi_z^{(k-1)n}) + V^{k-2}(x_{2n})$ learned during the previous step. (Note that at $k = 2$, $V^{k-2} = 0$.) The mutual information objective is then

$$\max_{\pi_z^k} I(Z^k; O_n|x_0, \pi_z^k) + V^{(k-1)n}(x_n). \tag{5.16}$$

Algorithm 3 Maximizing $I(Z^{Kn}; X_n, X_{2n}, \ldots, X_{Kn})$ with dynamic programming

for $k = 2, \ldots, K$ **do**
  for all dimensions $i = 0, \ldots, |\pi_z^k| - 1$ in parallel **do**
    for M iterations **do**
      Update $q_{\psi^{i,k}}$: $\psi^{i,k} \leftarrow \psi^{i,k} - \alpha \nabla_{\psi_{i,k}} (D_{KL}(p(z^k|x_0, \pi_z^{i,k}, o_n)||q_{\psi^{i,k}}(z^k|x_0, \pi_z^{i,k}, o_n)))$
    end for
    for M iterations **do**
      Update $Q_{\eta^{i,k}}$: $\eta^{i,k} \leftarrow \eta^{i,k} - \alpha \nabla_{\eta^{i,k}} ((Q_{\eta^{i,k}}(x_0, \pi_z^{i,k}) - \text{Target})^2)$,
      $\text{Target} = \mathbb{E}_{z^k \sim p(z^k), o_n \sim p(o_n|x_0, \pi_z^{i,k}, z^k)} [\log q_{\psi^{i,k}}(z^k|x_0, \pi_z^{i,k}, o_n) - \log p(z^k)]$
    end for
  end for
  Update f_{λ^k}: $\lambda^k \leftarrow \lambda^k + \alpha \nabla_{\lambda^k} (\sum_{i=0}^{|\pi_z^k|-1} Q_{\eta^{i,k}}(x_0, \pi_z^{i,k} = f_{\lambda^k}(x_0)[i]))$
end for

Algorithm 3 provides the full algorithm for computing our version of long-horizon empowerment, including the equations for updating the actor, parameter-specific critics, and parameter-specific variational posteriors.

5.3 Training Skills, Representations, and Behavior Policies with Empowerment

We treat maximizing empowerment with respect to a behavior policy as a finite-horizon MDP. The representations are generated by the RNN that was trained to maximize empowerment. Although these representations may not be sufficient statistic representations of histories with respect to the underlying state, we will denote these representations as x_t and treat them as if they are sufficient statistics and thus Markov representations. The action space is the regular primitive action space. The reward $r(x_t)$ depends on the time step. For the first $n-1$ timesteps, the reward is 0. For the final time step, the reward $r(x_n)$ is $V^K(x_n) = I(Z^K; O_{2n}|x_n, \pi_z^K) + V^{K-1}(X_{2n})$ that was learned using Algorithm 3. Given that the reward depends on time, we implement n different actors $\pi_{\kappa^t} : \mathcal{X} \to \mathcal{A}$ and n different critics $Q_{\zeta^t} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ for $t = 0, \ldots, n-1$ so that there is an actor-critic pair for each of the n time steps when the behavior policy interacts with the environment. Alternatively, a time variable can be included in a single actor and critic.

Algorithm 4 provides our complete approach. In each iteration, the agent first interacts with the environment using an ϵ-greedy version of its behavior policy π_κ. Then the agent maximizes the mutual information with respect to (i) skill-conditioned policies, (ii) a representation learning function, and (iii) a behavior policy.

5.4 Experiments

5.4.1 Environments

We evaluated our complete approach in Algorithm 4 in three partially observable settings, which are visualized in Figure 5.1.

Password-Protected Cage: In this environment, an agent starts each episode locked in a password-protected cage. During the initial four timesteps, the agent can move

Algorithm 4 Learning Skills, Representations, and Behavior Policies with Empowerment

while not converged **do**

▷ Interact with Environment

Collect new history with ϵ-greedy policy: $\beta \leftarrow [a_0, o_1, \ldots, a_{n-1}, o_n]$

▷ Update Skill Discovery Actor-Critic

for all dimensions $i = 0, \ldots, |\pi_z| - 1$ in parallel **do**

for M iterations **do** ▷ Update Variational Posterior

Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \epsilon \nabla_{\psi_i}(\mathbb{E}_{c_0 \sim p(c_0|\beta)}[D_{KL}(p(z|c_0, \pi_z^i, o_n)||q_{\psi^i}(z|c_0, \pi_z^i, o_n))])$

end for

for M iterations **do** ▷ Update Critic

Update Q_{α^i}: $\alpha^i \leftarrow \alpha^i - \epsilon \nabla_{\alpha^i}(\mathbb{E}_{c_0 \sim p(c_0|\beta)}[(Q_{\alpha^i}(c_0, \pi_z^i) - \text{Target})^2])$

Target $= \mathbb{E}_{c_0 \sim p(c_0|\beta), z \sim p(z), o_n \sim p(o_n|c_0, \pi_z^i, z)}[\log q_{\psi^i}(z|c_0, \pi_z^i, o_n) - \log p(z)]$

end for

end for

Update f_λ: $\lambda \leftarrow \lambda + \epsilon \nabla_\lambda(\mathbb{E}_{c_0 \sim p(c_0|\beta)}[\sum_{i=0}^{|\theta_z|-1} Q_{\alpha^i}(c_0, \pi_z^i = f_\lambda(c_0)[i])])$ ▷ Update Actor

▷ Update Representation Learning Actor-Critic

for all dimensions $i = 0, \ldots, |\eta| - 1$ in parallel **do**

for M iterations **do** ▷ Update Variational Posterior

Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \epsilon \nabla_{\psi_i}(\mathbb{E}_{c_0 \sim p(c_0|\beta)}[D_{KL}(p(z|c_0, \eta^i, o_n)||q_{\psi^i}(z|c_0, \eta^i, o_n))])$

end for

for M iterations **do** ▷ Update Critic

Update Q_{ξ^i}: $\xi^i \leftarrow \xi^i - \epsilon \nabla_{\xi^i}((Q_{\xi^i}(\eta^i) - \text{Target})^2)$ with noisy η^i,

Target $= \mathbb{E}_{c_0 \sim p(c_0|\eta^i, \beta), z \sim p(z), o_n \sim p(o_n|c_0, \eta^i, z)}[\log q_{\psi^i}(z|c_0, \eta^i, o_n) - \log p(z)]$

end for

end for

Update f_γ: $\gamma \leftarrow \gamma + \epsilon \nabla_\gamma(\sum_{i=0}^{|\eta|-1} Q_{\xi^i}(\eta^i = f_\gamma(a)[i]))$ ▷ Update Actor

Compute long-horizon empowerment with Algorithm 3 (if applicable)

▷ Update behavior policy Actor-Critic

for each actor-critic pair $t = 0, \ldots, n-1$ **do**

for M iterations **do** ▷ Update Critic

Update $Q_{\zeta^i} : \zeta^i \leftarrow \zeta^i - \epsilon \nabla_{\zeta^i}(\mathbb{E}_{x_t \sim p(x_t|\beta)}[(Q_{\zeta^i}(x_t, a_t) - \text{Target})^2)$,

if $t + 1 == n$ **then** Target $= \mathbb{E}_{x_t \sim p(x_t|\beta), x_{t+t} \sim p(x_{t+1}|x_t, a_t)}[V^K(x_{t+1})]$

else Target $= \mathbb{E}_{x_t \sim p(x_t|\beta), x_{t+t} \sim p(x_{t+1}|x_t, a_t)}[Q_{\zeta^{t+1}}(x_{t+1}, a = \pi_{\kappa^t}(x_{t+1}))]$

end if

end for

Update $\pi_{\kappa^t} : \kappa^t \leftarrow \kappa^t + \nabla_{\kappa^t}(\mathbb{E}_{x_t \sim p(x_t|\beta)}[Q_{\zeta^t}(x_t, a_t = \pi_{\kappa^t}(x_t))])$ ▷ Update Actor

end for

end while

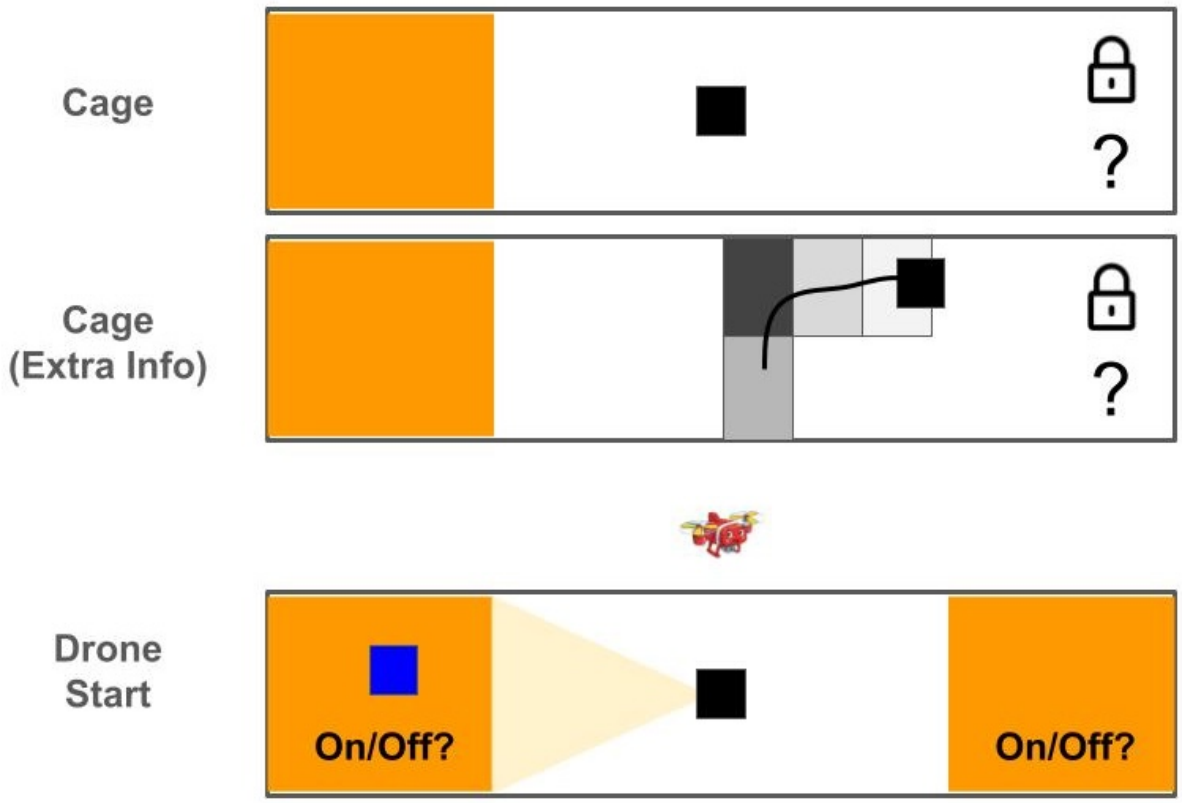


Figure 5.1: The three environments to which we applied our complete approach in Algorithm 4

.

around the cage but cannot escape. On the fifth timestep, the agent needs to output a password in the form of a scalar number. If this number is within a certain threshold of the true password, which is some number in the range $[-1, 1]$, the agent is free to leave the cage for the remainder of the episode. Otherwise, the agent remains stuck in place inside the cage for the remainder of the episode. During the initial four timesteps, the agent can obtain the password by both (i) moving to a particular region on the west side of the cage (shown by the orange square) and then (ii) looking up, which in this case is an extra action dimension that needs to be greater than 0.

Maximizing empowerment with respect to an RNN should cause the agent to learn different representations for histories that (i) end in different (x, y) positions and (ii) signal different passwords as well as no password. For instance, a history that contains a password of -0.5 should be encoded to a different representation than a history with a

password of 0.5. Maximizing empowerment with respect to the behavior policy should encourage the agent during the initial four timesteps to obtain the password by moving to the password region and looking up. If the agent were to compare two behavior policies: (i) move to some position in the password area, look up to get password, and then look down and (ii) move to the same position in the password region but then not look up, the first policy should on average achieve larger empowerment because this behavior enables the agent to leave the cage in the future, whereas the second behavior will often cause the agent to remain locked in the cage because the password is not known.

Password-Protected Cage (Extra Info): This is the same environment as the prior password-protected cage except we add extra state uncertainty that has no effect on the size of an agent's skillset. Specifically, we add an extra dimension to the agent's observation that shows the grayscale color of the tile the agent is currently on. In each episode, we divide the cage into $2 \times 10 = 20$ tiles (i.e., boxes) and randomly assign a number in the range $[-1.25, 1.25]$ to the tile. The observation the agent receives will include a dimension that shows the grayscale color for the tile the agent is currently on.

The purpose of this change was to add extra uncertainty in the underlying state (i.e., 20 new state features), but uncertainty that has no effect on the size of an agent's skillset. This is true because regardless of whether or not the agent knows the color of the tiles in the cage, the agent is still limited to learning skills that target (x, y) positions. In the case where the colors of tiles are unknown, each skill targets a set of observations that include a particular small (x, y) region and then all the possible colors of the tile the skills terminate on. If the colors are known, each skill targets a smaller set of observations that includes the same (x, y) and the exact color. However, the number of sets of observations that a skillset targets is not different.

Drone Start: This setting implements a variant of the drone setting discussed in chapter 1. The environment consists of two agents, a "drone", and two buttons. The buttons are shown by the orange regions in Figure 5.1. The primary agent can move

around a 2D room as well as potentially fly a drone in a larger 2D region provided that the agent first pushes the "On" button before the "Off" button. However, if the agent pushes the "Off" button first, the drone remains frozen in place for the remainder of the episode. The location of the "On" button is randomly selected each episode. In addition to moving around and potentially controlling the drone, the agent also has another action dimension enabling it to "look around". If the agent does look around, it will see the other agent on the correct "On" button.

Maximizing empowerment with respect to the agent's RNN should force the RNN to assign different representations to histories that (i) terminate with the agent and drone in different (x, y) positions, (ii) signal different "On" button locations, and (iii) show the drone is currently "On" versus "Off". Maximizing empowerment with respect to the behavior policy should at least encourage the agent to execute the look around to determine which button starts the drone.

5.4.2 Results

Table 5.1: Empowerment Maximization Results

Env	**Info Seek (%)**	**Mean Emp. Gain (nats)**
Cage	100	2.15
Cage (Extra Info)	100	1.5
Drone Start	100	3.13

The results in Table 5.1 show that our full approach works as expected and that the agent executes the correct information-seeking action. The first column shows the percentage of episodes after training that the agent performs the environment-specific information-seeking action, which was 100% in all episodes. The second column compares the average empowerment gain when comparing a policy that seeks information compared to another policy that achieves the same distribution over marginal states but does not seek information. For instance, in the drone setting, one pair of actions that was examined included (a) an action that stayed still but looked up for at least one time step to determine

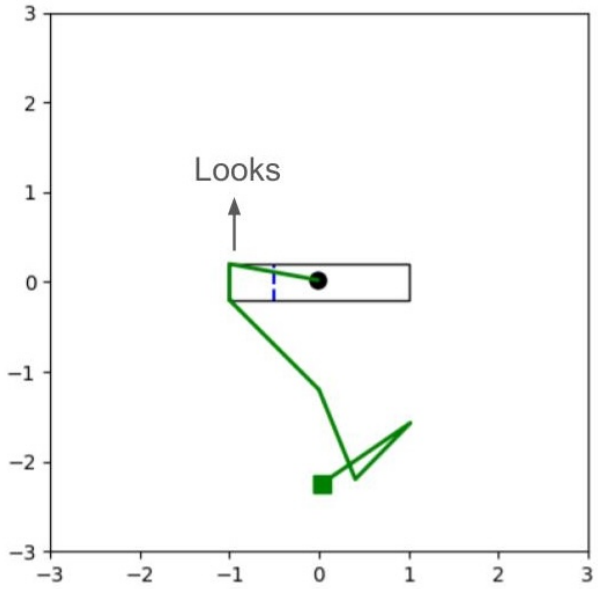


Figure 5.2: Agent executes the information-seeking action of moving to the password region on the left side of the cage and looking up to obtain the password.

the start button and (b) an action that stayed still but looked down for the entire policy. The empowerment gain for the information-seeking actions was significant. For instance, in the regular password-protected cage and drone start buttons, the mean empowerment gain from information seeking was 2.15 and 3.13 nats, respectively, which is equivalent to increasing the skillset by around 8.5x and 22.9x, respectively, from seeking information. Figures 5.2 and 5.3 show some sample trajectories in the password-protected cage and drone start environments. In the password-protected cage setting in Figure 5.2, the agent first moves to the password region across the blue dashed line while looking up, thereby attaining the password. Then, the agent enters the correct password and leaves the cage. In the drone start environment, the first action the agent takes before it has observed the password is to move towards the top left while looking around to see the other agent (i.e., the agent executes the correct information-seeking action). Then if the other agent appears on the left side, the agent moves left turning the drone on. On the other hand, if the other agent is on the right side, the agent moves right across the button threshold turning the drone on before again moving left.

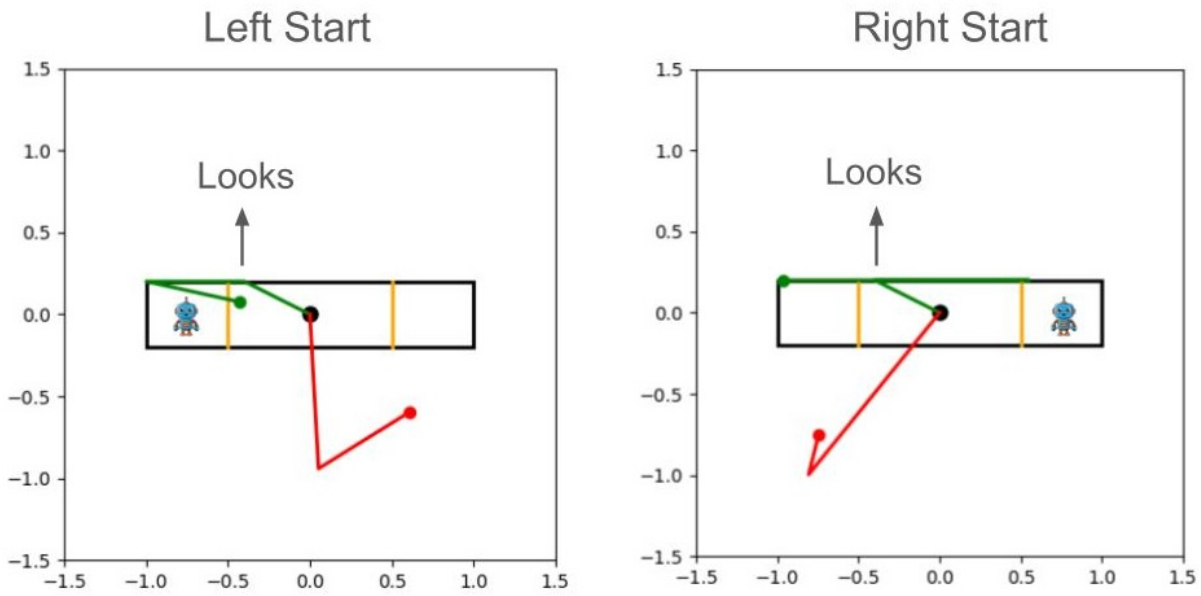


Figure 5.3: Agent executes the information-seeking action of looking around to see the other agent and then moving to the correct start button. Regardless of the location of the button, the agent first moves towards the top left while looking up. Then, if the agent observed that the button is on the left side, the agent moves to the left side as shown by green trajectory. However, if the agent observes that the button is on the right side, the agent first moves right to turn the drone on. The red trajectory shows the movement of the drone after the agent turns it on.

CHAPTER 6

Future Work and Conclusion

6.1 Open Problems in Empowerment Research

There are several remaining open problems for training empowerment-based agents. This section highlights a few of these problems.

6.1.1 World Models

One key challenge is that computing the empowerment of a representation requires a world model. The mutual information objective involves the expectation, $\mathbb{E}_{z\sim p(z),o_n\sim p(o_n|c_0,\pi_z,z)}[\cdot]$, which requires sampling many skills and executing them using the skill-conditioned policy π_z. The most practical way to sample from the expectation in an unbiased manner is through simulation, which is why a world model is needed. The experiments in this thesis assumed the agent has access to the world model (i.e., the distribution $p(o_{t+1}|h_t, a_t)$, in which h_t is the history of actions and observations through time t and a_t is the next action), but this is not a scalable approach. One possible solution is to try to learn this distribution (i.e., learn the world model) (Ha and Schmidhuber, 2018; Hafner et al., 2019a, 2020, 2021, 2024; Bruce et al., 2024). Yet predicting high-dimensional and stochastic observations is challenging.

A potentially more practical alternative is to learn latent world models, in which an agent predicts compressed representations of future observations (Grill et al., 2020; Guo et al., 2022; Ghugare et al., 2023; Assran et al., 2023; Bardes et al., 2024; Levy et al., 2023). The problem with many of these approaches is that they can suffer from mode collapse in which the agent learns to map all observations to similar latent representations. We have provided a potential solution to this problem: integrate the latent world model into the mutual information objective (Levy et al., 2024). Instead of measuring $I(Z; O_n|c_0, \pi_z)$, we proposed using the lower bound objective $I(Z; Z_{\text{enc}}|c_0, \pi_z)$ in which z_{enc} is sampled from an observation encoder $p_\tau(z_{\text{enc}}|o)$. Our work then showed how this mutual information objective can be trained in a way such that the observation encoder $p_\tau(z_{\text{enc}}|o)$ could output latent representations that can be modeled by a latent world model $p(z_{\text{enc}}|c_0, a_0)$. However, this method was not used for the experiments in this thesis.

6.1.2 Exploration

In order to train some type of accurate world model, agents need to have executed a variety of primitive actions from a variety of different histories. Thus, another key problem is how to encourage agents to explore different actions. Specifically, is there a way to adapt the empowerment maximization objective so that agents can explore while they are still maximizing empowerment? As this thesis and prior empowerment work (Jung et al., 2012; Mohamed and Rezende, 2015; Karl et al., 2017; Zhao et al., 2020) have demonstrated, maximizing empowerment produces various important behaviors like stabilization, predator avoidance, and information seeking. Ideally, agents can maximize some empowerment-related reward that still encourages these behaviors while also encouraging exploration.

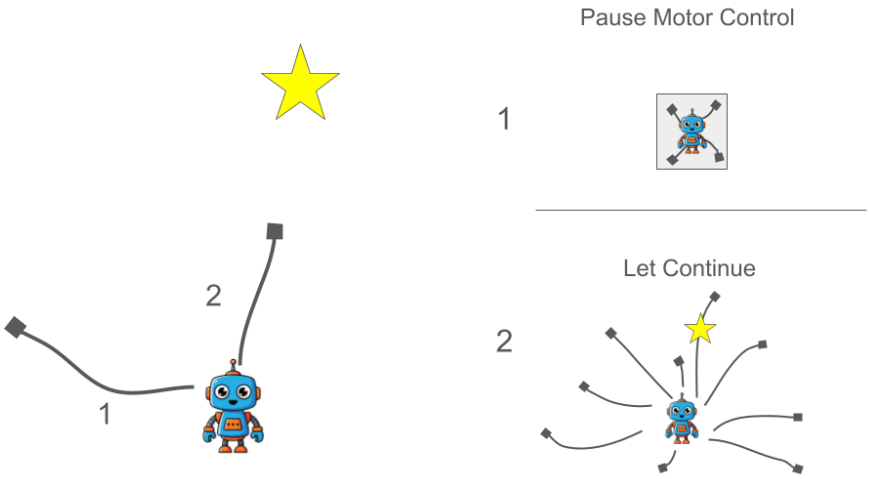


Figure 6.1: Humans can control empowerment-maximizing agents by manipulating the empowerment reward. For instance, consider the situation in the image where a human wants the empowerment-maximizing robot to move to the yellow star. The human can encourage the robot to do this by penalizing actions, like action 1, that go off course. Specifically, the human could take the action of pausing or limiting the robot's motor control, which can significantly limit its empowerment. On the other hand, if the robot progresses towards the goal, the human could let the robot continue, resulting in the agent achieving a representation with relatively high empowerment.

6.1.3 Controlling Empowerment-based Agents

Another open problem is how humans can control empowerment-based agents such as the ones described in this thesis. The conventional way agents trained via unsupervised skill discovery are controlled by humans is to divide training into two phases. In the first phase, agents build their skillsets using the USD algorithm. Then in the second phase, humans can control the agent by introducing a reward function, which forces the agent to learn a new policy that selects skills from the skillset learned during the prior phase that maximize reward. For empowerment-based agents, this approach is problematic because empowerment-based agents can learn massive skillsets that may be intractable to search through. For instance, empowerment-based agents that maximize the longer-horizon versions of empowerment such as the objective introduced in the prior chapter, learn skillsets that grow exponentially with time. Searching through these massive skillsets may not be feasible.

A more practical alternative for controlling empowerment-maximizing agents may be for humans to manipulate the empowerment reward of agents. Humans can manipulate the empowerment reward by altering the agent's transition dynamics. For instance, if the robot does something the human does not like, the human could pause or limit the robot's motor control such as in Figure 6.1 or place the agent in some location where there is not much to do (i.e., give the robot a timeout). If the robot uses money to pay for its energy, the human could potentially take some of this money away. All of these human interventions cause the agent to achieve relatively low empowerment representations, which should discourage these behaviors. On the other hand, humans can reward empowerment-maximizing agents through the opposite interventions (e.g., not pausing motor control, giving the agent more objects to interact with or more money to pay for energy). With this approach, humans can control the agent's behavior and the agent's action space remains the relatively small primitive action space.

6.2 Conclusion

In order to achieve general-purpose agents that can execute large sets of skills at a low price, the cost of skill discovery must be low. This will be difficult using the dominant frameworks of reinforcement learning and behavior cloning, where skill discovery can require expensive labor-intensive procedures like designing rewards and teleoperating robots for each new skill the agent needs to learn. This thesis demonstrates that empowerment can address two major issues in skill discovery — policy diversity and state uncertainty — in an *unsupervised* manner. Thus, this thesis provides evidence that Empowerment-based agents can build skillsets at a lower cost, and thus may provide a more scalable approach to training general-purpose agents.

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.

Siddhant Agarwal, Ishan Durugkar, Peter Stone, and Amy Zhang. f-policy gradients: A general framework for goal-conditioned RL using f-divergences. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EhhPtGsVAv.

Siddhant Agarwal, Caleb Chuck, Harshit Sikchi, Jiaheng Hu, Max Rudolph, Scott Niekum, Peter Stone, and Amy Zhang. A unified framework for unsupervised reinforcement learning algorithms. In *Workshop on Reinforcement Learning Beyond Rewards @ Reinforcement Learning Conference 2025*, 2025a. URL https://openreview.net/forum?id=tQtO75p5HA.

Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the behavior space of an RL agent. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=mUDnPzopZF.

Cameron Allen, Aaron Kirtland, Ruo Yu Tao, Sam Lobel, Daniel Scott, Nicholas Petrocelli, Omer Gottesman, Ronald Parr, Michael Littman, and George Konidaris. Mitigating partial observability in sequential decision processes via the lambda discrepancy. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 62988–63028. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/73073ccb3bc559fd001e66b9079d6d5e-Paper-Conference.pdf.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Pe-

ter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/453fadbd8a1a3af50a9df4df899537b5-Paper.pdf.

Edson Hiroshi Aoki, Arunabha Bagchi, Pranab Mandal, and Yvo Boers. A theoretical look at information-driven sensor management criteria. In *14th International Conference on Information Fusion*, pages 1–8, 2011.

Brenna D. Argall, Brett Browning, and Manuela Veloso. Learning robot motion control with demonstration and advice-operators. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 399–404, 2008. doi: 10.1109/IROS.2008.4651020.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning robot objectives from physical human interaction. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 217–226. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/bajcsy17a.html.

Bram Bakker. Reinforcement learning with long short-term memory. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/a38b16173474ba8b1a95bcbc30d3b8a5-Paper.pdf.

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

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4058–4068, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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

Glen Berseth, Daniel Geng, Coline Manon Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. {SM}irl: Surprise minimizing reinforcement learning in unstable environments. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=cPZOyoDloxl.

D.P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, 2 edition, 2001.

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *Interna-*

tional Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=DfUjyyRW90.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu, editors, *Robotics: Science and Systems*, 2023. ISBN 978-0-9923747-9-2. URL http://dblp.uni-trier.de/db/conf/rss/rss2023.html#BrohanBCCDFGHHH23.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

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.

Jonathan Colaço Carr, Qinyi Sun, and Cameron Allen. Focused skill discovery: Using per-factor empowerment to control state variables. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=eM5ZtMijef.

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, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1953–1963. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/choi21b.html.

Caleb Chuck, Fan Feng, Carl Qi, Chang Shi, Siddhant Agarwal, Amy Zhang, and Scott Niekum. Null counterfactual factor interactions for goal-conditioned reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2uPZ4aX1VV.

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.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993. doi: 10.1162/neco.1993.5.4.613.

Ishan Durugkar, Mauricio Tec, Scott Niekum, and Peter Stone. Adversarial intrinsic motivation for reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8622–8636. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/486c0401c56bf7ec2daa9eba58907da9-Paper.pdf.

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.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tc5qisoB-C.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, P. Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. *ArXiv*, abs/1709.04905, 2017. URL https://api.semanticscholar.org/CorpusID:22221787.

Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1oK8aoxe.

Zafeirios Fountas, Noor Sajid, Pedro Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11662–11675. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/865dfbde8a344b44095495f3591f7407-Paper.pdf.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2009.04.005. URL https://www.sciencedirect.com/science/article/pii/S136466130900117X.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John ODoherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience Biobehavioral Reviews*, 68:862–879, 2016. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2016.06.022. URL https://www.sciencedirect.com/science/article/pii/S0149763416301336.

Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal conditioned policies. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Hye9lnCct7.

Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Russ Salakhutdinov. Simplifying model-based RL: Learning representations, latent-space models, and policies with one objective. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MQcmfgRxf7a.

Alessandro Giusti, Jérôme Guzzi, Dan C. Cireşan, Fang-Lin He, Juan P. Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, Davide Scaramuzza, and Luca M. Gambardella. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1 (2):661–667, 2016. doi: 10.1109/LRA.2015.2509024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou

Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. *CoRR*, abs/2110.07058, 2021. URL https://arxiv.org/abs/2110.07058.

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

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pislar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, Michal Valko, Remi Munos, Mohammad Gheshlaghi Azar, and Bilal Piot. BYOL-explore: Exploration by bootstrapped prediction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=qHGCH75usg.

David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1803.html#abs-1803-10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Ma-*

chine Learning Research, pages 2555–2565. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *CoRR*, abs/1912.01603, 2019b. URL http://arxiv.org/abs/1912.01603.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

Matthew J. Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. *CoRR*, abs/1507.06527, 2015. URL http://arxiv.org/abs/1507.06527.

Ziming He, Chao Song, Jingchen Li, and Haobin Shi. Pdrl: Towards deeper states and further behaviors in unsupervised skill discovery by progressive diversity. *IEEE Transactions on Cognitive and Developmental Systems*, 17(3):495–509, 2025. doi: 10.1109/TCDS.2024.3471645.

Nicolas Heess, Jonathan J. Hunt, Timothy P. Lillicrap, and David Silver. Memory-based control with recurrent neural networks. *CoRR*, abs/1512.04455, 2015. URL http://arxiv.org/abs/1512.04455.

Jiaheng Hu, Zizhao Wang, Peter Stone, and Roberto Martín-Martín. Disentangled unsupervised skill discovery for efficient hierarchical reinforcement learning. In A. Globerson,

L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 76529–76552. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8c263f70550cc7d69dba3fc170a23e77-Paper-Conference.pdf.

Jiaheng Hu, Peter Stone, and Roberto Martín-Martín. Slac: Simulation-pretrained latent action space for whole-body real-world rl, 2025. URL https://arxiv.org/abs/2506.04147.

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.

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.

Leslie Pack Kaelbling. Learning to achieve goals. In *International Joint Conference on Artificial Intelligence*, 1993. URL https://api.semanticscholar.org/CorpusID:5538688.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1–2):99–134, May 1998. ISSN 0004-3702.

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.

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, pages 128–135 Vol.1, 2005. doi: 10.1109/CEC.2005.1554676.

W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: the tamer framework. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, page 9–16, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605586588. doi: 10.1145/1597735.1597738. URL https://doi.org/10.1145/1597735.1597738.

Chris Kreucher, Keith Kastella, and Alfred O. Hero III. Sensor management using an active sensing approach. *Signal Processing*, 85(3):607–624, 2005. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2004.11.004. URL https://www.sciencedirect.com/science/article/pii/S0165168404003068.

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, 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, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34478–34491. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/debf482a7dbdc401f9052dbe15702837-Paper-Conference.pdf.

Alexander Levine, Peter Stone, and Amy Zhang. Multistep inverse is not all you need,

2024. URL https://arxiv.org/abs/2403.11940.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018. doi: 10.1177/0278364917710318. URL https://doi.org/10.1177/0278364917710318.

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.

Andrew Levy, Alessandro Allievi, and George Konidaris. Latent-predictive empowerment: Measuring empowerment without a simulator, 2024. URL https://arxiv.org/abs/2410.11155.

Andrew Levy, Alessandro G Allievi, and George Konidaris. Representation learning and skill discovery with empowerment. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=w61h2RK8ni.

Long-Ji Lin and Tom M. Mitchell. Reinforcement learning with hidden states. In *Proceedings of the Second International Conference on From Animals to Animats 2: Simulation of Adaptive Behavior: Simulation of Adaptive Behavior*, page 271–280, Cambridge, MA, USA, 1993. MIT Press. ISBN 0262631490.

Michael Littman and Richard S Sutton. Predictive representations of state. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Michael L. Littman. A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125, 2009. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2009.01.005. URL https://www.sciencedirect.com/science/article/pii/S0022249609000042. Special Issue: Dynamic Decision Making.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105): eads5033, 2025. doi: 10.1126/scirobotics.ads5033. URL https://www.science.org/doi/abs/10.1126/scirobotics.ads5033.

Yecheng Jason Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned reinforcement learning via f-advantage regression. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *ICLR*, 2023.

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. *CoRR*, abs/1811.02790, 2018. URL http://arxiv.org/abs/1811.02790.

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, page 2125–2133, Cambridge, MA, USA, 2015. MIT Press.

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, editors, *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.

Soroush Nasiriany, Vitchyr H. Pong, Steven Lin, and Sergey Levine. *Planning with goal-conditioned policies*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, pages 9887–9897. Computer Vision Foundation / IEEE, 2020. ISBN 978-1-7281-7168-5. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2020.html#NgXJG20.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16691–16723. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ni22a.html.

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, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: offline goal-conditioned rl with latent states as actions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware

unsupervised skill discovery. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27225–27245. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/park23h.html.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

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

Dean A. Pomerleau. Alvinn: an autonomous land vehicle in a neural network. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'88, page 305–313, Cambridge, MA, USA, 1988. MIT Press.

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, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Nicholas Rhinehart, Jenny Wang, Glen Berseth, John Co-Reyes, Danijar Hafner, Chelsea Finn, and Sergey Levine. Information is power: Intrinsic control via information capture. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages

10745–10758. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/59112692262234e3fad47fa8eabf03a4-Paper.pdf.

Max Rudolph, Caleb Chuck, Kevin Black, Misha Lvovsky, Scott Niekum, and Amy Zhang. Learning action-based representations using invariance. *Reinforcement Learning Journal*, 1:342–365, 2024.

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

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1312–1320, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schaul15.html.

J. Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pages 1458–1463 vol.2, 1991. doi: 10.1109/IJCNN.1991.170605.

J.H. Schmidhuber. *Networks Adjusting Networks*. Forschungsberichte Künstliche Intelligenz. Report. Verlag nicht ermittelbar, 1990. URL https://books.google.com/books?id=VzUrzQEACAAJ.

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. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgLZR4KvH.

Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and

Scott Niekum. Score models for offline goal-conditioned reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oXjnwQLcTA.

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.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Stas Tiomkin, Christoph Salge, and Daniel Polani. Process empowerment for robust intrinsic motivation. *Journal of Physics: Complexity*, 6(3):035011, aug 2025. doi: 10.1088/2632-072X/adf2ec. URL https://dx.doi.org/10.1088/2632-072X/adf2ec.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MYEap_OcQI.

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.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: interactive agent shaping in high-dimensional state spaces. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative*

Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Daan Wierstra, Alexander Foerster, Jan Peters, and Jürgen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *Proceedings of the 17th International Conference on Artificial Neural Networks*, ICANN'07, page 697–706, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540746897.

Jason L. Williams. *Information theoretic sensor management*. PhD thesis, USA, 2007. AAI0818404.

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`.

Ruihan Zhao, Kevin Lu, P. Abbeel, and Stas Tiomkin. Efficient empowerment estimation for unsupervised stabilization. In *International Conference on Learning Representations*, 2020. URL `https://api.semanticscholar.org/CorpusID:234358796`.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.

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*, 2025a. URL `https://openreview.net/forum?id=xoIeVdFO7U`.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhe Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit

world modeling. In *Structured World Models for Robotic Manipulation*, 2025b. URL https://openreview.net/forum?id=rFiLBM4YCh.

Chuning Zhu, Xinqi Wang, Tyler Han, Simon Shaolei Du, and Abhishek Gupta. Distributional successor features enable zero-shot policy optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

K.J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(65)90154-X. URL https://www.sciencedirect.com/science/article/pii/0022247X6590154X.