Unsupervised Skill Discovery in Non-Markov Settings with Empowerment

Andrew Levy
Brown University
andrew_levy2@brown.edu

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

George Konidaris
Brown Univeristy
gdk@brown.edu

Abstract

General purpose agents must be able to execute a large number of skills in non-Markov settings. Yet learning diverse sets of policies in these domains is challenging because agents also need to learn representations that preserve information about the underlying state found in histories of actions and observations. We introduce an Empowerment-based unsupervised skill discovery algorithm for building skillsets in non-Markov settings. The algorithm maximizes a mutual information objective with respect to both a recurrent neural network (RNN) and a skill-conditioned policy, enabling agents to simultaneously learn a representation and a large number of policies conditioned on the learned representation. We prove that our objective encourages RNNs that preserve information about the underlying state. We also demonstrate empirically that our approach can learn large skillsets ranging from hundreds to thousands of skills in three small, non-Markov settings.

1 Introduction

Non-Markov settings, in which an agent's history provides more information about the underlying state than the latest observation, are ubiquitous in real-world settings. As a result, general purpose agents must be able to learn large sets of skills in non-Markov settings. But building skillsets in these settings is difficult because it is unclear what representation should be used as input to these policies. Conditioning policies on a specific number of the most recent actions and observations produces a trade-off between information and computational cost. At one end of the spectrum, simply conditioning policies on the latest observation means the agent is acting with limited information about the underlying state, which can result in stochastic and redundant skillsets that target wide and overlapping regions of observations. At the other end of the spectrum, agents could condition skills on the full history of actions and observations. While this will lessen the problem of underlying state uncertainty that can produce redundant skillsets, it comes at the expense of processing long histories prior to each action, which can be impractical in long-horizon settings. A more compelling approach is to instead learn representations that preserve information about the underlying state, using an architecture such as a recurrent neural network (Rumelhart & McClelland, 1987; Hopfield, 1982) that only needs to process a limited number of inputs to generate each representation. Yet existing unsupervised skill discovery algorithms have not yet demonstrated how agents can learn both information-preserving representations and skills conditioned on these representations in non-Markov settings (Gregor et al., 2016; Eysenbach et al., 2019; Sharma et al., 2020; Park et al., 2024; Zheng et al., 2025; Levy et al., 2025).

Preprint.

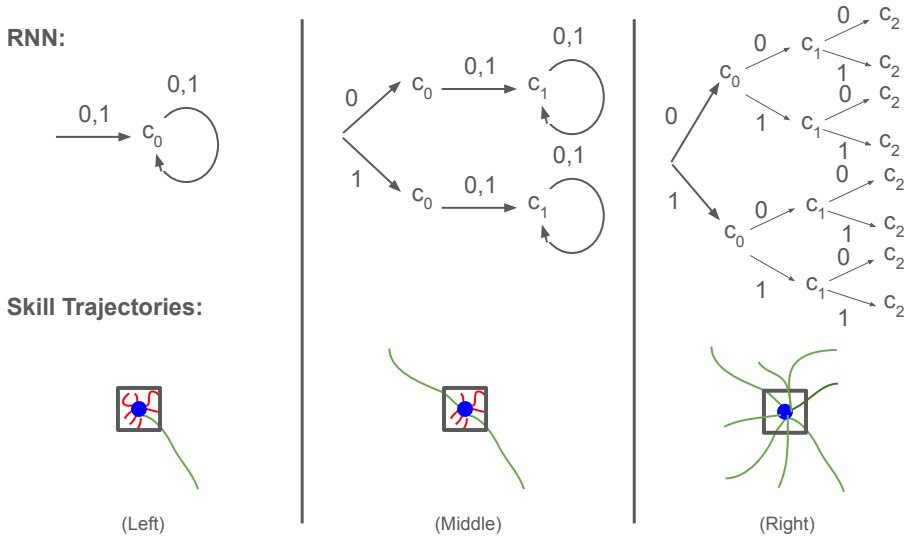

Figure 1: Visualizations of agent skillsets under different RNN and skill-conditioned policy settings. (Left) Skillset with low mutual information as RNN maps all 3-bit passwords to the same representation c_0 producing redundant skill trajectories. (Middle) Skillset with medium mutual information as RNN detects first bit of password but entangles the remaining bits. (Right) Skillset with high mutual information as RNN assigns different representations to different passwords, enabling skills to target different observations outside the cage. An RNN trained to maximize skillset size via mutual information will automatically prefer the extra information in the middle setting relative to the left setting and the right setting relative to the middle setting.

We introduce a new unsupervised skill discovery algorithm based on empowerment (Klyubin et al., 2005; Salge et al., 2013; Jung et al., 2012; Mohamed & Rezende, 2015; Karl et al., 2017; Gregor et al., 2016; Levy et al., 2023) that can help agents learn large skillsets in non-Markov settings. Empowerment has provided an intuitive approach for skill discovery in settings with Markov observations. For any skill-conditioned policy, the mutual information between skills and skill-terminating observations measures the number of distinct skills produced by the skill-conditioned policy under consideration. To build large skillsets, empowerment-based algorithms simply maximize this mutual information with respect to the skill-conditioned policy (i.e., the algorithms try to find skill-conditioned policies with more distinct skills). In the non-Markov setting, when an agent's policy is conditioned on a representation generated from an RNN, the mutual information depends both on the parameters of the RNN and the skill-conditioned policy. Thus, to scale empowerment to the non-Markov setting, we propose to simply maximize the mutual information between skills and observations with respect to both the recurrent neural network and the skill-conditioned policy.

Maximizing mutual information can encourage information-preserving representations because if an RNN has reduced the size of an agent's skillset by assigning similar representations to histories with different underlying state distributions, this objective should encourage the RNN to disentangle these histories. Consider the setting in Figure 1 where an agent is locked in a cage but can exit the cage if it enters the correct three-bit password. During each of the first three timesteps of an episode, the agent is given a single bit of a three-bit password, and then in the following three timesteps, the agent must output one bit of the password per time step. The agent receives no signal indicating whether the entered password is correct or not until all three bits have been entered. Figure 1 shows three skillsets with different RNNs and skill-conditioned policies at the third timestep when the agent begins to enter the password. Figure 1 (Left) shows what the agent's skillset could look like at the beginning of training when the RNN entangles histories by assigning similar representations to all combinations of the three-bit password. In this case, the mutual information, which measures the number of distinct skills in the skillset, is minimal as most skills produce similar outcomes with the agent entering the incorrect password and remaining locked in the cage. Figure 1 (Middle) shows an agent's skillset in which the RNN can now distinguish the first bit of the password, but then assigns similar representations to the following two bits. In this scenario, the mutual information of the

agent's skillset is larger as the agent more frequently executes the correct password, but the skillset is still largely redundant with many skills causing the agent to remain in the cage. Figure 1 (Right) shows the skillset where the RNN preserves information by assigning different representations to different passwords. In this case, the mutual information of the agent's skillset is large as most skills target distinct observations outside the cage. A key benefit of training an RNN to maximize skillset size using mutual information is that it will automatically encourage the information-preserving changes that for instance take the RNN from the left setting to the middle setting and then to the right setting because these changes increase skillset size. No regularization terms specializing in information gain need to be added for the RNN to disentangle histories corresponding to different underlying state distributions.

This work makes the following three contributions. First, we present new objective functions for maximizing the mutual information between skills and observations with respect to an RNN and a skill-conditioned policy. Second, we provide theoretical analysis showing that maximizing empowerment with respect to an RNN does encourage the RNN to preserve information with respect to the underlying state found in the agent's history of actions and observations. Specifically, we prove that our average empowerment objective is maximized when agents have learned an RNN that outputs sufficient statistic representations of histories with respect to the underlying state, which means that the RNN outputs representations that provide as much information about the underlying state as the agent's full history. In addition, we prove that if the agent is considering two RNNs and one RNN provides more information about the underlying state than the other, then the average empowerment of the RNN providing more information will be at least as large as the average empowerment from the other RNN. Third, we demonstrate empirically that our approach can learn large skillsets in three settings with non-Markov observations. All three settings are small in terms of observation dimensionality but require that the agents learn representations that remember one or more past actions and observations in order to build skillsets containing hundreds to thousands of skills. To our knowledge, our approach is the first unsupervised skill discovery algorithm to successfully perform both representation learning and unsupervised skill discovery in non-Markov settings.

2 Background

2.1 Modeling the Agent-Environment Interaction

We assume that the agent's interaction with the environment can be modeled as a Hidden Markov Control Process (HMCP), which is a Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) without a reward function. An HMCP is defined by the tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, p(s_0), p(s_{t+1}|s_t, a_t), p(o_t|s_t))$, in which $\mathcal{S}$ is the space of states, $\mathcal{O}$ is the space of observations, and $\mathcal{A}$ is the space of actions. $p(s_0)$ is the initial state distribution, $p(s_{t+1}|s_t, a_t)$ is the Markov state transition dynamics in which the distribution of the next state conditioned on the current state and action is independent of the history, and $p(o_t|s_t)$ is the observation distribution that is conditionally independent of the history given state s_t. We also assume that observations are non-Markov, which in this work will mean that the distribution over the underlying state s_t given observation o_t, $p(s_t|o_t)$, is not equal to the distribution over the underlying state given the history of actions and observations $h_t = (a_0, o_1, \ldots, a_{t-1}, o_t)$, $p(s_t|h_t)$, for all histories h_t. Using information theoretic terms, non-Markov observations are equivalent to noting that the mutual information between histories and environment states conditioned on the last observation, $I(H_t; S_t|O_t) > 0$, which means that on average, the history h_t provides extra information (i.e., reduces uncertainty) on the underlying state s_t when given the current observation o_t.

In this work, actions will be generated by skills or policies $\pi : \mathcal{C} \rightarrow \Delta(\mathcal{A})$ that map contexts, which are learned representations, to distributions over actions. The goal in this work is to learn a large number of skills that target distinct observations. We will define a set of skills (i.e., a skillset) as a distribution over policies $p(\pi)$. Context representations that serve as inputs to policies will be generated by the representation distribution $p_\eta(c_{t+1}|c_t, a_t, o_{t+1})$. This distribution will be modeled as a diagonal Gaussian $\mathcal{N}(c_{t+1}; [\mu, \sigma^2] = f_\eta(c_t, a_t, o_{t+1}))$, in which the mean and $\log$ variance are output by an RNN f_η that takes as input the current context c_t, action a_t, and next observation o_{t+1}. A special type of context representation is a sufficient statistic of a history with respect to the underlying state. Sufficient statistics $x_t \in \mathcal{C}$ have the property that the distribution over the underlying state s_t given the sufficient statistic x_t is independent of the history h_t: $p(s_t|x_t) = p(s_t|h_t)$, which is

equivalent to noting that the histories provide no extra information about the underlying state given the sufficient statistic (i.e., $I(H_t; S_t|X_t) = 0$), in contrast to single observations. Sufficient statistics x_t are sampled from the distribution $p(x_{t+1}|x_t, a_t, o_{t+1})$.

2.2 Empowerment

In this work, we will learn representations and skillsets using empowerment. In non-Markov settings, we will define the empowerment of a (c_0, η) pair, in which c_0 is a context representation and η is the parameters of an RNN, as the maximum mutual information between a policy random variable Π and a skill-terminating observation random variable O_n, in which the maximum is with respect to the skillset $p(\pi)$ that defines Π:

$$\mathcal{E}(c_0, \eta) = \max_{p(\pi)} I(\Pi; O_n|c_0, \eta). \tag{1}$$

The mutual information between Π and O_n for a skillset $p(\pi)$, $I(\Pi; O_n|c_0, \eta)$, measures the number of distinct skills within skillset $p(\pi)$ when the skillset is executed from c_0 and future contexts c_t are generated by an RNN defined by η (Cover & Thomas, 2006). Because it is unclear how to search through the space of all skillsets $p(\pi)$, it is common to instead use a lower bound of equation 1, in which the mutual information is between a skill random variable Z and skill-terminating observations O_n and is conditioned on a skill-conditioned policy π_{θ_z} parameterized by θ_z:

$$\mathcal{E}(c_0, \eta) = \max_{\theta_z} I(Z; O_n|c_0, \eta, \theta_z) \tag{2}$$

$$= \max_{\theta_z} \mathbb{E}_{z\sim p(z), o_n \sim p(o_n|c_0,\eta,\theta_z,z)}[\log p(z|c_0, \eta, \theta_z, o_n) - \log p(z)]. \tag{3}$$

A copy of the lower bound proof from Levy et al. (2025) is provided in section A of the appendix. In this definition of empowerment, agents are still searching for the $p(\pi)$ skillset with the most distinct skills but now the search is limited to sets of policies produced by sampling skills $z \in \mathcal{Z}$ from a fixed distribution over skills $p(z)$ and then inputting the skill z into a skill-conditioned policy $\pi_{\theta_z} : \mathcal{C} \times \mathcal{Z} \to \Delta(\mathcal{A})$ that maps contexts and skills to distributions over actions. Note that the channel distribution $p(o_n|c_0, \eta, \theta_z, z)$ marginalizes over the joint distribution $p(h, x_0, s_0, a_0, s_1, o_1, \ldots, c_{n-1}, x_{n-1}, a_{n-1}, s_n, o_n|c_0, \eta, \theta_z, z)$ containing the intermediate states $s_t \sim p(s_t|s_{t-1}, a_{t-1})$, actions $a_t \sim p_{\theta_z}(a_t|c_t, z)$, observations $o_t \sim p(o_t|s_t)$, contexts $c_t \sim p_\eta(c_t|c_{t-1}, a_{t-1}, o_t)$, and sufficient statistic $x_t \sim p(x_t|x_{t-1}, a_{t-1}, o_t)$ random variables. $h \sim p(h|c_0, \eta)$ refers to the histories of actions and observations that are mapped to c_0 using the current RNN parameters η. Note that the time subscripts in the variables do not necessarily refer to absolute times but rather the relative number of timesteps since the starting context c_0 as histories of different lengths can be mapped to the same c_0.

While the objective in line 2 computes the largest skillset for a specific context c_0 that is produced from a specific RNN parameterized by η, general purpose agents need to have large skillsets across a variety of contexts c_0 and they are also not limited to using a single set of RNN parameters η. As a result, in this work we are interested in maximizing the following average context empowerment objective with respect to the parameters of an agent's RNN:

$$\max_{\eta} \mathbb{E}_{c_0\sim p(c_0|\eta)}[\mathcal{E}(c_0, \eta)] = \max_{\eta, f_\lambda} \mathbb{E}_{c_0\sim p(c_0|\eta)}[I(Z; O_n|c_0, \eta, \theta_z = f_\lambda(c_0))]. \tag{4}$$

In this objective, the distribution $p(c_0|\eta)$ marginalizes over the joint distribution $p(h, c_0|\eta)$, in which the history $h \sim p(h|\eta) = p(h)$ is sampled from a replay buffer or some manually specified distribution. Per the right side of 4, the average empowerment objective involves maximizing an average mutual information with respect to (i) the RNN parameters η (i.e., representation learning) and (ii) the function f_λ that outputs the parameters θ_z of a skill-conditioned policy for a specific context c_0 (i.e., skill discovery).

2.3 Maximizing Empowerment In Practice

Levy et al. (2025) demonstrated how a mutual information objective similar to the one in line 4 can be maximized in practice. The key difference in the objective Levy et al. (2025) maximized was that it involved Markov settings. Thus, instead of training an RNN parameterized by η, the authors only trained an observation encoder $p_\eta(c_t|o_t)$ parameterized by η. To jointly maximize mutual information with respect to η and f_λ, Levy et al. (2025) used two actor-critics.

The first actor-critic trained the actor, f_λ, to output more diverse skillsets θ_z for specific contexts c_0, assuming a fixed observation encoder defined by η. To guide the actor to larger skillsets, the critic $Q_\alpha(c_0, \theta_z)$ was used to measure the variational mutual information $I^V(Z; O_n|c_0, \theta_z)$ for various skillsets θ_z. $I^V(Z; O_n|c_0, \theta_z)$ is a lower bound on the true mutual information and is formed by replacing the posterior $p(z|c_0, \theta_z, o_n)$, which is intractable to compute in continuous settings, with the variational posterior $q_\psi(z|c_0, \theta_z, o_n)$ parameterized by ψ (Barber & Agakov, 2003; Poole et al., 2019). A key change from prior work is that Levy et al. (2025) include the skillset θ_z in the conditioned variables of the variational posterior and train $q_\psi(z|c_0, \theta_z, o_n)$ to match the true posterior $p(z|c_0, \theta_z, o_n)$ for any θ_z under consideration. The authors show that these changes produce a tighter bound on variational mutual information, enabling agents to more accurately measure the size of a skillset θ_z. Yet taking θ_z, which represents the parameters of a skill-conditioned policy and can be thousands of parameters long, as an input makes training the variational posterior and the critic impractical. Levy et al. (2025) overcome this problem by showing that all that is needed to estimate an accurate gradient with respect to the actor f_λ is how the variational mutual information changes with respect to each parameter of θ_z while the others remain constant. As a result, instead of training a single variational posterior and critic, they train $|\theta_z|$ variational posteriors $q_{\psi^i}(z|c_0, \theta_z^i, o_n)$ and critics $Q_{\alpha^i}(c_0, \theta_z^i)$ in parallel for $i = 0, \ldots, |\theta_z| - 1$, in which $|\theta_z|$ is the number of parameters in θ_z. θ_z^i is a scalar representing the skillset θ_z in which all parameters take on their greedy value from $f_\lambda(c_0)$ except for the i-th parameter, which is set to θ_z^i. The left side of Figure 2 visualizes how the parameter-specific critics attach to the f_λ actor for this actor-critic responsible for skill discovery.

The second actor-critic trains an actor f_γ to output RNN parameters η that produce larger average mutual information $\mathbb{E}_{c_0 \sim p(c_0|\eta)}[I(Z; O_n|c_0, \eta)]$. Similar to the first actor-critic, parameter-specific variational posteriors $q_{\psi^i}(z|c_0, \eta^i, o_n)$ and critics $Q_{\xi^i}(\eta^i)$ for $i = 0, \ldots, |\eta| - 1$ are used to measure the average variational mutual information of different RNN settings η. Similar to θ_z^i, η^i is a scalar representing a vector of parameters η in which all parameters take on their greedy values from $f_\gamma(a)$ except for the i-th parameter, which is set to η^i. a is simply a fixed vector as the actor f_γ does not take as input a variable vector. The right side of Figure 2 visualizes this second actor-critic architecture.

3 Building Skillsets in Non-Markov Settings with Empowerment

In this section, we present our approach for jointly performing representation learning and skill discovery with empowerment. Our approach extends the work of Levy et al. (2025) to the non-Markov setting by training an RNN to output representations instead of an observation encoder. We then provide theoretical analysis showing that maximizing average empowerment with respect to the parameters of an RNN encourages the agent to preserve information about the underlying state found in the agent's actions and observations.

3.1 Algorithm

Algorithm 1 provides the algorithm for updating both the skill discovery and representation learning actor-critics, which are visualized in Figure 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 $q_{\psi^i}(z|c_0, \theta_z^i, o_n)$ are first trained to match the true posteriors $p(z|c_0, \theta_z^i, o_n)$ for noisy θ_z for M iterations ($M = 300$ in our experiments). Then the parameter-specific critics $Q_{\alpha^i}(c_0, \theta_z^i)$ are trained to approximate the variational mutual information $I^V(Z; O_n|c_0, \theta_z^i)$ for noisy θ_z 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 $J(\gamma) = \sum_{i=0}^{|\eta|-1} Q_{\xi^i}(\eta^i = f_\gamma(a)[i])$, where $f_\gamma(a)[i]$ outputs the i-th component of the vector $f_\gamma(a)$.

3.2 Theoretical Analysis

To learn large and distinct skillsets in non-Markov settings, the representations need to preserve information about the underlying state. Otherwise, skillsets can be overly stochastic, yielding redundant skills. In this section, we prove that maximizing mutual information with respect to the

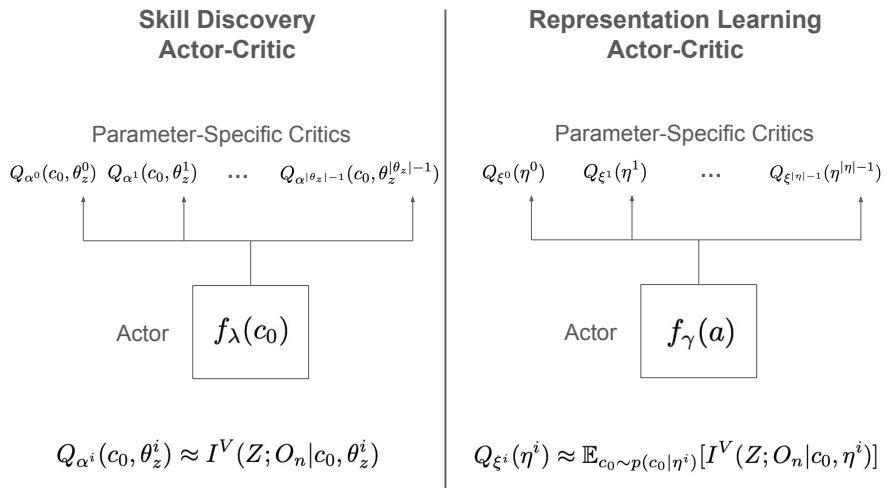

Figure 2: We use two actor-critic architectures to maximize our average mutual information with respect the parameters of a skill-conditioned policy and an RNN. The actor-critic on the left is designed to learn diverse skillsets across various contexts c_0. The actor $f_\lambda(c_0)$ maps a context to 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 1 Skill Discovery and Representation Learning with Empowerment

▷ Update Skill Discovery Actor-Critic
for all dimensions $i = 0, \ldots, |\theta_z| - 1$ in parallel **do**
  for M iterations **do** ▷ Update Variational Posterior
    Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \epsilon \nabla_{\psi_i}(D_{KL}(p(z|c_0, \theta_z^i, o_n) || q_{\psi^i}(z|c_0, \theta_z^i, o_n)))$ with noisy θ_z^i
  end for
  for M iterations **do** ▷ Update Critic
    Update Q_{α^i}: $\alpha^i \leftarrow \alpha^i - \epsilon \nabla_{\alpha^i}((Q_{\alpha^i}(c_0, \theta_z^i) - \text{Target})^2)$ with noisy θ_z^i,
    Target $= \mathbb{E}_{z \sim p(z), o_n \sim p(o_n|c_0, \theta_z^i, z)}[\log q_{\psi^i}(z|c_0, \theta_z^i, o_n) - \log p(z)]$
  end for
end for
Update f_λ: $\lambda \leftarrow \lambda + \epsilon \nabla_\lambda(\sum_{i=0}^{|\theta_z|-1} Q_{\alpha^i}(c_0, \theta_z^i = f_\lambda(c_0)[i]))$ ▷ 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 $= \mathbb{E}_{z \sim p(z), o_n \sim p(o_n|c_0, \eta^i, z)}[\log q_{\psi^i}(z|c_0, \eta^i, o_n) - \log p(z)]$
  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

RNN encourages the RNN to preserve information. Specifically, in Theorem 1, we prove that the average empowerment is maximized when an RNN outputs sufficient statistic representations of histories with respect to the underlying state. That is, average empowerment is maximized when an RNN outputs representations of histories that preserve as much information as the full history. In Theorem 2, we prove that if there are two RNNs and one RNN outputs representations that provide more information about the underlying state, then the average empowerment provided by the RNN that provides additional information is at least as large as the average empowerment produced by the other RNN. This result is notable because it implies that empowerment can be a relatively dense reward for training RNNs. Agents can be rewarded every time the agent gains certain bits of information about the underlying state similar to how the size of the agent's skillset grew in Figure 1 each time the agent remembered more bits of the password.

Theorem 1. *Let η be the parameters of any RNN $f_\eta : \mathcal{C} \times \mathcal{A} \times \mathcal{O} \to \mathcal{C}$ and let η_x be the parameters of an RNN that outputs sufficient statistic representations of histories with respect to the underlying state, then the average empowerment produced by η is upper bounded by the average empowerment produced by η_x: $\mathbb{E}_{c_0 \sim p(c_0|\eta)}[\mathcal{E}(c_0, \eta)] \leq \mathbb{E}_{x_0 \sim p(x_0|\eta_x)}[\mathcal{E}(x_0, \eta_x)]$.*

Theorem 2. *Let η^a and η^b be the parameters of two RNNs, and let $p(h, c_0^a, c_0^b)$ be the joint distribution of a history h uniformly sampled from a dataset of histories and c_0^a and c_0^b be the contexts produced by inputting the history into the two respective RNNs. If (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) $p(s_t|c_0^a, c_0^b, c_t^b, z) = p(s_t|c_0^b, c_t^b, z)$ for all (c_0^a, c_0^b) with $p(c_0^a, c_0^b) > 0$ and for all $t \in [0, n)$, then $\mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[\mathcal{E}(c_0^a, \eta^a)] \leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[\mathcal{E}(c_0^b, \eta^b)]$.*

The proofs for both theorems are provided in section B of the appendix.

3.3 Limitations

Our approach enables agents to build skillsets in non-Markov settings by jointly performing representation learning and skill discovery in a manner that is reward-free and does not require knowledge of the exact transition dynamics $p(s_{t+1}|s_t, a_t)$ nor the observation distribution $p(o_t|s_t)$. However, to train the parameter-specific variational posteriors and critics in parallel, the algorithm needs to collect a large number of (skill z, observation o_n) tuples and so our approach does assume it can sample the distribution $p(o_{t+1}|h_t, a_t)$ thousands of times in parallel, which is not a realistic assumption. Future work can try to learn this potentially high-dimensional generative model similar to other work in POMDPs (Han et al., 2019; Igl et al., 2018; Lee et al., 2020). Alternatively, recent work by Levy et al. (2024) has shown how agents can maximize mutual information using easier-to-learn latent-predictive models that predict compact representations of observations. Future work can try to integrate similar models into our approach.

4 Experiments

4.1 Environments and Baselines

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 3. We briefly describe these environments next and provide additional detail in section E of the appendix.

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 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 θ_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

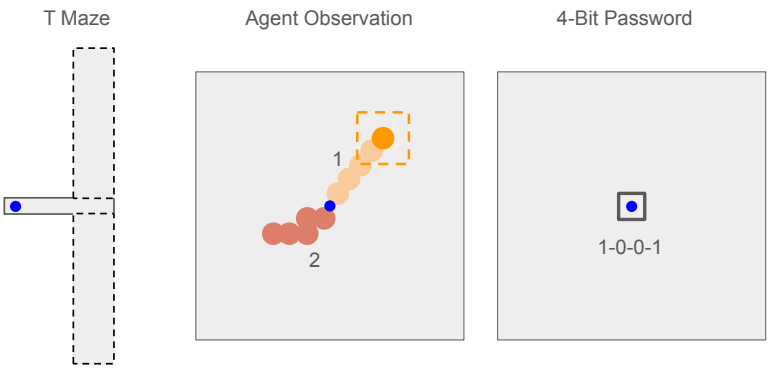

Figure 3: Visuals of the three non-Markov environments in which we evaluated our approach.

"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 3). 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 3). After these 10 timesteps, the primary agent then must attempt to move within a threshold of the goal (dashed orange square in Figure 3) 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 version of the password setting discussed previously. 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 a hint of the final bit. The purpose for 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

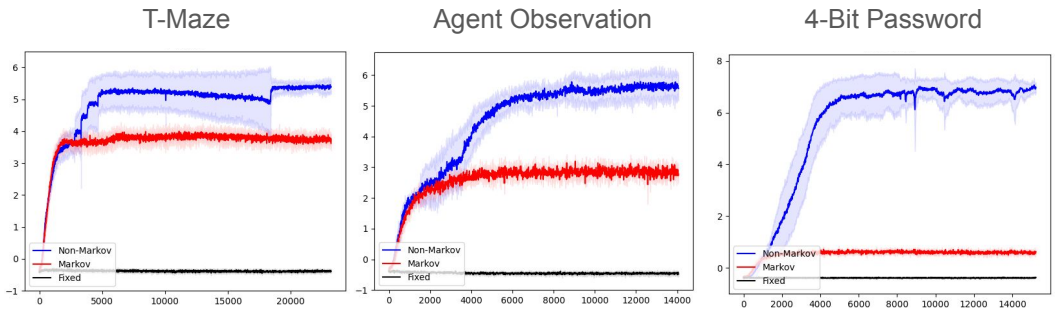

Figure 4: Average skillset size (in nats) vs. number of algorithm iterations. Skillset size is measured using variational mutual information. Mean and 1 std. of error computed with 5 random seeds.

a change that preserves information by assigning different representations to histories with different belief states. However, if the $f_\lambda(c)$ actor that outputs skill-conditioned policies θ_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.

We compare to two other algorithms in these three settings. First we compare to the empowerment approach introduced by Levy et al. (2025), which jointly performs representation learning and skill discovery with empowerment but is designed for the Markov setting. This comparison will test whether our approach actually encourages the RNN to preserve information. 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 would mean no training of an RNN is necessary. Because the focus of this paper is skill discovery in non-Markov settings, all agents are evaluated based on the size of their skillsets, measured using average variational mutual information $\mathbb{E}_{c_0 \sim p(c_0|\eta)}[I^V(Z; O_n|c_0, \eta, \theta_z = f_\lambda(c_0))]$.

4.2 Results

Figure 4 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 1. Our approach is the "Non-Markov" line (blue); the approach of Levy et al. (2025), which follows algorithm 1 but trains an observation encoder, is shown by the "Markov" line (red); and the approach that uses algorithm 1 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 1. These skillset sizes were 5.2x, 16.4x, and $555.6x$ larger than the performance of the next best approach from Levy et al. (2025). 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 5, 6, and 7. $H(O_n)$, 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 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 8, 9, 10, and 11, 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.

5 Related Work

Unsupervised Skill Discovery and Empowerment There have been several works that have attempted to use the mutual information between skills and observations to build skillsets in settings with compact and Markov observations (Gregor et al., 2016; Eysenbach et al., 2019; Warde-Farley et al., 2019; Achiam et al., 2018; Hansen et al., 2020; Sharma et al., 2020; Zhang et al., 2021; Campos et al., 2020; Choi et al., 2021; Levy et al., 2023). As a result of the inconsistent performance of these methods, a new class of empowerment-adjacent methods emerged that either made some changes to the mutual information objective such as adding regularization terms to improve exploration (Laskin et al., 2022; Zheng et al., 2025; Kim et al., 2023; Strouse et al., 2022; Baumli et al., 2021; Hu et al., 2024; Wang et al., 2025) or used related metrics to learn distinct skills (Park et al., 2022, 2023, 2024). Yet, Levy et al. (2025) demonstrated that many of these works still struggled to learn large skillsets in simple settings.

Representation Learning and Empowerment There have also been several works that have used empowerment or empowerment-related methods to learn representations (Klyubin et al., 2008; Capdepuy, 2011; Bharadhwaj et al., 2022; Rudolph et al., 2024; Lamb et al., 2023). Unlike our approach, they did not perform representation learning in non-Markov settings and they did not simultaneously learn closed loop skills.

Representation Learning in POMDPs Also related are the numerous works that learn representations in POMDPs (Lin & Mitchell, 1993; Schmidhuber, 1990, 1991; Bakker, 2001; Hausknecht & Stone, 2015; Ni et al., 2022; Wierstra et al., 2007; Heess et al., 2015; Hafner et al., 2019; Allen et al., 2024). A key difference from our work is that these approaches rely on hand-crafted rewards to provide signals for learning representations that in practice can be costly to implement and/or overly sparse. Our approach uses skillset size, which is affected by whether or not an agent's representation preserves information about natural features of the environment such as the turns in T-Maze or the password bottleneck in 4-Bit Password. These approaches also do not jointly learn large skillsets while performing representation learning.

6 Conclusion

Given that non-Markov settings are omnipresent in the real world, agents must be able to learn large skillsets in these settings. But learning large sets of skills in non-Markov settings is challenging as agents need to both learn a representation that preserves information and learn many policies conditioned on these representations. We show that an empowerment objective both in theory and

in practice enables agents to jointly learn information-preserving representations and large skillsets conditioned on these representations.

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.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62988–63028. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/73073ccb3bc559fd001e66b9079d6d5e-Paper-Conference.pdf.

Bram Bakker. Reinforcement learning with long short-term memory. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *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 (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

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

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

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

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

Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational empowerment as representation learning for goal-conditioned reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1953–1963. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/choi21b.html.

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

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

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

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

Dongqi Han, Kenji Doya, and Jun Tani. Variational recurrent models for solving partially observable control tasks. *CoRR*, abs/1912.10703, 2019. URL http://arxiv.org/abs/1912.10703.

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

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.

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.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, April 1982. ISSN 0027-8424. URL http://view.ncbi.nlm.nih.gov/pubmed/6953413].

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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76529–76552. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8c263f70550cc7d69dba3fc170a23e77-Paper-Conference.pdf.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2117–2126. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/igl18a.html.

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, 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.

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

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All else being equal be empowered. In Mathieu S. Capcarrère, Alex A. Freitas, Peter J. Bentley, Colin G. Johnson, and Jon Timmis (eds.), *Advances in Artificial Life*, pp. 744–753, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31816-3.

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.

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

Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: deep reinforcement learning with a latent variable model. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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*, pp. 271–280, Cambridge, MA, USA, 1993. MIT Press. ISBN 0262631490.

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

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 (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 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, 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 (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27225–27245. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/park23h.html.

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

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

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

David E. Rumelhart and James L. McClelland. *Learning Internal Representations by Error Propagation*, pp. 318–362. 1987.

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

J. Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pp. 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`.

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

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

Zizhao Wang, Jiaheng Hu, Caleb Chuck, Stephen Chen, Roberto Martín-Martín, Amy Zhang, Scott Niekum, and Peter Stone. Skild: unsupervised skill discovery guided by factor interactions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

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`.

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, pp. 697–706, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540746897.

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

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

A Empowerment Lower Bound Proof (from Levy et al. (2025))

The following proof shows that the empowerment objective in equation 2, which maximizes a mutual information between a skill random variable Z and a skill-terminating observation O_n with respect to a skill-conditioned policy θ_z, is a lower bound to equation 1, which maximizes the mutual information between a policy random variable Π and O_n with respect to any distribution over policies $p(\pi)$.

$$\max_{p(\pi)} I(\Pi; O_n|c_0) \geq \max_{\theta_z} I(\Pi; O_n|c_0) \tag{5}$$

$$\geq \max_{\theta_z} I(Z, \theta_z; O_n|c_0) \tag{6}$$

$$= \max_{\theta_z} I(Z; O_n|c_0, \theta_z) \tag{7}$$

The upper bound in line 5 results from reducing the search space from all possible skillsets (i.e., a search over all distributions over policies $p(\pi)$, to only skillsets $p(\pi)$ produced from sampling a skill

$z \sim p(z)$ and then plugging the skill z into a skill conditioned policy $\pi_{\theta_z} : \mathcal{C} \times \mathcal{Z} \to \Delta(A)$ that maps contexts and skills to distributions over actions. The lower bound in line 6 results from the Data Processing Inequality (Cover & Thomas, 2006) because the random variables $Z, \pi_z \to \Pi \to O_n$ form a Markov chain. In the final line 7, θ_z is moved to the list of conditioned variables as it is deterministic.

B Theoretical Analysis Proofs

Theorem 3. *Let η be the parameters of any RNN $f_\eta : \mathcal{C} \times \mathcal{A} \times \mathcal{O} \to \mathcal{C}$ and let η_x be the parameters of an RNN that outputs sufficient statistic representations of histories with respect to the underlying state, 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.

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

□

Line 8 applies the definition of the empowerment of a tuple of context and RNN parameters. θ^* represents the mutual information maximizing skill-conditioned policy parameters for the (context, RNN) tuple of (c_0, η).

The lower bound in line 9 applies the convexity property of mutual information with respect to the channel distribution (Cover & 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 $p(o_n|c_0, \eta, \theta_z^*, z) = \int_{x_0} p(x_0|c_0, \eta)p(o_n|c_0, x_0, \eta, \theta_z^*, z)$ is a weighted mixture of the channels $p(o_n|c_0, x_0, \eta, \theta_z^*, z)$ containing the sufficient statistic representation x_0. Consequently, $I(Z; O_n|c_0, \eta, \theta_z^*) \leq \mathbb{E}_{x_0 \sim p(x_0, \eta)}[I(Z; O_n|c_0, x_0, \eta, \theta_z^*)]$.

Line 10 removes the dependence on the RNN parameterized by η by (i) using the RNN defined by η_x to produce the skill representation x_t and (ii) replacing θ_z^* with a specific skill-conditioned policy θ_z^x. As we will show, this will replace each mutual information term, $I(Z; O_n|c_0, x_0, \eta, \theta^*)$, with a new mutual information term, $I(Z; O_n|x_0, \eta_x, \theta_z^x)$, that is at least as large. For each context x_0, θ_z^x will be constructed as follows. For each x_0, find the tuple (c_0, x_0) with the largest $I(Z; O_n|c_0, x_0, \eta, \theta^*)$ as there can be multiple contexts c_0 associated with the same sufficient statistic x_0. Then, for each x_0 let θ_z^x be the skill-conditioned policy distribution $p(a_t|x_0, x_t, t) = \int_{c_t} p(c_t|c_0, x_0, x_t)p(a^t|c_0, x_0, x_t, c_t) = \int_{c_t} p(c_t|c_0, x_0, x_t)p(a^t|c_0, c_t, t)$, in which $p(a^t|c_0, c_t, t)$ is the probability specified by the skill-conditioned policy defined by θ_z^*. That is, the skill-conditioned policy θ_z^x will have the same distribution over actions as executed by θ_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 $I(Z; O_n|c_0, x_0, \eta, \theta_z^*)$ is upper bounded by the new mutual information $I(Z; O_n|x_0, \eta_x, \theta_z^x)$.

We first show that for each mutual information term $I(Z; O_n|c_0, x_0, \eta, \theta_z^*)$ from which θ_z^x was constructed in the previous step, $I(Z; O_n|x_0, \eta_x, \theta_z^x) = I(Z; O_n|c_0, x_0, \eta, \theta_z^*)$. 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 $p(o_n|c_0, x_0, \eta, \theta_z^*, z) = p(o_n|x_0, \eta_x, \theta_z^x, z)$ are the same for all (z, o_n) tuples. We show this below by proving by induction that the joint distributions $p(x_{t-1}, s_{t-1}, a_{t-1}, o_t, x_t|c_0, x_0, \eta, \theta_z^*, z) = p(x_{t-1}, s_{t-1}, a_{t-1}, o_t, x_t|x_0, \eta_x, \theta_z^x, z)$ for $t = 1, \ldots, n$. Then because the joint distribution

$p(x_{n-1}, s_{n-1}, a_{n-1}, o_n, x_n|c_0, x_0, \eta, \theta_z^*, z) = p(x_{n-1}, s_{n-1}, a_{n-1}, o_n, x_n|x_0, \eta_x, \theta_z^x, z)$, the channel distributions are equal: $p(o_n|c_0, x_0, \eta, \theta_z^*, z) = p(o_n|x_0, \eta_x, \theta_z^x, z)$.

The proof by induction goes as follows. In the base case at $t = 1$, the distribution $p(x_0|c_0, x_0, \eta, \theta_z^*, z) = p(x_0|x_0, \eta_x, \theta_z^x, z)$ because the same x_0 appears in the conditioning variables. $p(s_0|c_0, x_0, \eta, \theta_z^*, z) = p(s_0|x_0, \eta_x, \theta_z^x, z)$ because per the definition of sufficient statistic representations $p(s_t|x_t) = p(s_t|h_t)$, in which the history h_t includes context representations. $p(a_0|c_0, x_0, \eta, \theta_z^*, z) = p(a_0|x_0, \eta_x, \theta_z^x, z)$ using the definition of θ_z^x. $p(o_1|c_0, x_0, \eta, \theta_z^*, z, s_0, a_0) = p(o_1|x_0, \eta_x, \theta_z^x, z, s_0, a_0)$ as the next observation o_1 only depends on s_0 and a_0 and is independent of the other variables. Lastly, $p(x_1|c_0, x_0, \eta, \theta_z^*, z, s_0, a_0, o_1) = p(x_1|x_0, \eta_x, \theta_z^x, z, s_0, a_0, o_1)$ 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 $p(x_0, s_0, a_0, o_1, x_1|c_0, x_0, \eta, \theta^*, z) = p(x_0, s_0, a_0, o_1, x_1|x_0, \eta_x, \theta_z^x, z)$.

Assuming the proof holds through $t = k - 1$, then at step $t = k$, $p(x_{k-1}|c_0, x_0, \eta, \theta_z^*, z) = p(x_{k-1}|x_0, \eta_x, \theta_z^x, z)$ because the joint distribution $p(x_{k-2}, s_{k-2}, a_{k-2}, o_{k-1}, x_{k-1}|c_0, x_0, \eta, \theta_z^*, z) = p(x_{k-2}, s_{k-2}, a_{k-2}, o_{k-1}, x_{k-1}|x_0, \eta_x, \theta_z^b, z)$. $p(s_{k-1}|c_0, x_0, \eta, \theta_z^*, z, x_{k-1}) = p(s_{k-1}|x_0, \eta_x, \theta_z^x, z, x_{k-1})$ because again $p(s_{k-1}|x_{k-1}) = p(s_{k-1}|h_{k-1})$ as x_{k-1} is a sufficient statistic. $p(a_{k-1}|c_0, x_0, \eta, \theta_z^*, z, x_{k-1}) = p(a_{k-1}|x_0, \eta_x, \theta_z^x, z, x_{k-1})$ using the definition of θ_z^x. Again, $p(o_1, x_k|c_0, x_0, \eta, \theta_z^*, z, x_{k-1}, s_{k-1}, a_{k-1}) = p(o_1, x_k|x_0, \eta_x, \theta_z^x, z, x_{k-1}, s_{k-1}, a_{k-1})$ as o_k only depends 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 $p(x_{k-1}, s_{k-1}, a_{k-1}, o_k, x_k|c_0, x_0, \eta, \theta_z^*, z) = p(x_{k-1}, s_{k-1}, a_{k-1}, o_k, x_k|x_0, \eta_x, \theta_z^x, z)$.

Thus, $I(Z; O_n|x_0, \eta_x, \theta_z^x) = I(Z; O_n|c_0, x_0, \eta, \theta_z^*)$ for those (c_0, x_0) tuples from which θ_z^x was constructed. For the other smaller $I(Z; O_n|c_0, x_0, \eta, \theta_z^*)$ terms that were not used to construct θ_z^x, these will also be replaced by the larger $I(Z; O_n|x_0, \eta_x, \theta_z^x)$. If these terms exist, the inequality in line 10, will be replaced by a strictly less than.

In line 11, the lower bound results from replacing the skill-conditioned policy θ_z^x with the optimal skill-conditioned policy $\theta_z^{x,*}$ 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 12 uses the definition of the empowerment of a context 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 an RNN defined by η_x that generates sufficient statistic representations.

Theorem 4. *Let η^a and η^b be the parameters of two RNNs, and let $p(h, c_0^a, c_0^b)$ be the joint distribution of a history h uniformly sampled from a dataset of histories and c_0^a and c_0^b be the contexts produced by inputting the history into the two respective RNNs. If (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) $p(s|c_0^a, c_0^b, c_t^b, z) = p(s|c_0^b, c_t^b, z)$ for all (c_0^a, c_0^b) with $p(c_0^a, c_0^b) > 0$ and for all $t \in [0, n)$, 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.

$$\mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[\mathcal{E}(c_0^a, \eta^a)] = \mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a)}[I(Z; O_n|c_0^a, \eta^a, \theta_z^{a,*})] \tag{13}$$

$$\leq \mathbb{E}_{c_0^a \sim p(c_0^a|\eta^a), c_0^b \sim p(c_0^b|c_0^a, \eta^a)}[I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})] \tag{14}$$

$$\leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[I(Z; O_n|c^b, \eta^b, \theta_z^b)] \tag{15}$$

$$\leq \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[I(Z; O_n|c^b, \eta^b, \theta_z^{b,*})] \tag{16}$$

$$= \mathbb{E}_{c_0^b \sim p(c_0^b|\eta^b)}[\mathcal{E}(c_0^b, \eta^b)] \tag{17}$$

□

Line 13 applies the definition of the empowerment of a tuple of context and RNN parameters. $\theta^{a,*}$ 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 14 applies the convexity property of mutual information with respect to the channel distribution (Cover & 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., $I(C_b; S|c_a)$ for each $c_a \sim p(c_a)$), then the channel distribution $p(o_n|c_0^a, \eta^a, \theta_z^{a,*}, z)$ is a weighted mixture of channels $p(o_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z)$ (i.e., $p(o_n|c^a, \eta^a, \theta_z^{a,*}, z) = \int_{c_0^b} p(c_0^b|c_0^a, \eta^a) p(o_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z)$) and so the mutual information of the mixed channel $I(Z; O_n|c^a)$ is upper bounded by the average mutual information of the individual channels $\mathbb{E}_{c_0^b \sim p(c_0^b|c_0^a)}[I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta^{a,*})]$.

Line 15 removes the dependence on the RNN parameterized by η^a by (i) using the RNN defined by η^b to produce the skill representation and (ii) replacing $\theta^{a,*}$ with a specific skill-conditioned policy θ^b. As we will show, this will replace each mutual information term, $I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta^{a,*})$, with a new mutual information term, $I(Z; O_n|c_0^b, \eta_b, \theta_z^b)$, that is at least as large. For each context c_0^b, θ_z^b will be constructed as follows. For each c_0^b, find the tuple (c_0^a, c_0^b) with the largest $I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta^{a,*})$ as there can be multiple contexts c_0^a associated with the same c_0^b. Then, for each c_0^b let θ_z^b be the skill-conditioned policy distribution $p(a_t|c_0^b, c_t^b, t) = \int_{c_t^a} p(c_t^a|c_0^a, c_0^b, c_t^b) p(a^t|c_0^a, c_0^b, c_t^b, c_t^a, t) = \int_{c_t^a} p(c_t^a|c_0^a, c_0^b, c_t^b) p(a^t|c_0^a, c_t^a, t)$, in which $p(a^t|c_0^a, c_t^a, t)$ is the probability specified by the skill-conditioned policy defined by $\theta_z^{a,*}$. That is, the skill-conditioned policy θ_z^b will have the same distribution over actions as executed by $\theta_z^{a,*}$ 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 $I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})$ is upper bounded by the new mutual information $I(Z; O_n|c^b, \eta^b, \theta_z^b)$.

We first show that for each mutual information term $I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})$ from which θ_z^b was constructed in the previous step, $I(Z; O_n|c_0^b, \eta^b, \theta_z^b) = I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})$. 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 $p(o_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(o_n|c_0^b, \eta^b, \theta_z^b, z)$ are the same for all (z, o_n) tuples. We show this below by proving by induction that the joint distributions $p(c_{t-1}^b, s_{t-1}, a_{t-1}, o_t, c_t^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(c_{t-1}^b, s_{t-1}, a_{t-1}, o_t, c_t^b|c_0^b, \eta^b, \theta_z^b, z)$ for $t = 1, \ldots, n$. Then because the joint distribution $p(c_{n-1}^b, s_{n-1}, a_{n-1}, o_n, c_n^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(c_{n-1}^b, s_{n-1}, a_{n-1}, o_n, c_n^b|c_0^b, \eta^b, \theta_z^b, z)$, the channel distributions are equal: $p(o_n|c_0^a, c_0^b, \eta_a, \theta_z^{a,*}, z) = p(o_n|c_0^b, \eta_b, \theta_z^b, z)$.

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

Assuming the proof holds through $t = k - 1$, then at step $t = k$, $p(c_{k-1}^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(c_{k-1}^b|c_0^b, \eta^b, \theta_z^b, z)$ because the joint distribution $p(c_{k-2}^b, s_{k-2}, a_{k-2}, o_{k-1}, c_{k-1}^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(c_{k-2}^b, s_{k-2}, a_{k-2}, o_{k-1}, c_{k-1}^b|c_0^b, \eta^b, \theta_z^b, z)$. $p(s_{k-1}|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z, c_{k-1}^b) = p(s_{k-1}|c_0^b, \eta^b, \theta_z^b, z, c_{k-1}^b)$ because $p(s_{k-1}|c_0^a, c_0^b, c_{k-1}^b, z) = p(s_{k-1}|c_0^b, c_{k-1}^b, z)$ per the assumption in the theorem statement. $p(a_{k-1}|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z, c_{k-1}^b) = p(a_{k-1}|c_0^b, \eta^b, \theta_z^b, z, c_{k-1}^b)$ using the definition of θ_z^b. Again, $p(o_1, c_k^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z, c_{k-1}^b, s_{k-1}, a_{k-1}) = p(o_1, c_k^b|c_0^b, \eta^b, \theta_z^b, z, c_{k-1}^b, s_{k-1}, a_{k-1})$ as o_k only depends 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 $p(c_{k-1}^b, s_{k-1}, a_{k-1}, o_k, c_k^b|c_0^a, c_0^b, \eta^a, \theta_z^{a,*}, z) = p(c_{k-1}^b, s_{k-1}, a_{k-1}, o_k, c_k^b|c_0^b, \eta^b, \theta_z^b, z)$.

Thus, $I(Z; O_n|c_0^b, \eta^b, \theta_z^b) = I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})$ for those (c_0^a, c_0^b) tuples from which θ_z^b was constructed. For the other smaller $I(Z; O_n|c_0^a, c_0^b, \eta^a, \theta_z^{a,*})$ terms that were not used to construct θ_z^b, these will also be replaced by the larger $I(Z; O_n|c_0^b, \eta^b, \theta_z^b)$. If these terms exist, the inequality in line 15, will be replaced by a strictly less than.

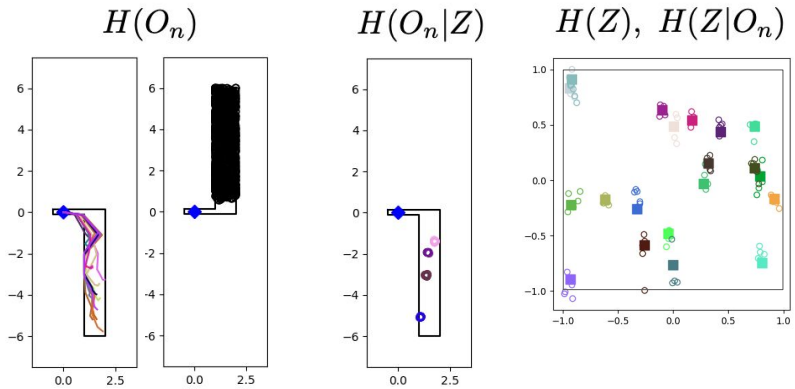

Figure 5: Entropy visualizations for the T-Maze domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The left most figure shows agent trajectories from randomly selected skills $z \sim p(z)$. The adjacent figure marks skill-terminating observations from 1000 randomly 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.

In line 16, the lower bound results from replacing the skill-conditioned policy θ_z^b with the optimal skill-conditioned policy $\theta_z^{b,*}$ 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 17 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.

C Entropy Visualizations

Figures 5, 6, and 7 provide visualizations of the entropy terms for the three domains.

D Representation Visualizations

Figures 8, 9, 10, and 11 shows samples of the disentangled learned representations for all settings.

E Additional Environment Detail

1. T-Maze
   - Action space: $a_t \in \mathbb{R}^2$ representing change in (x, y) directions
   - State space: $s_t \in \mathbb{R}^4$ representing (x position, y position, north or south hallway direction, part of hallway). The last component has three possibilities agent indicating which part of the hallway the agent is in: (i) thin hallway only, (ii) intersection of thin and large hallway, and (iii) large hallway. Note that this component alone does not indicate which direction the hallway opens and was added to help the agent turn.

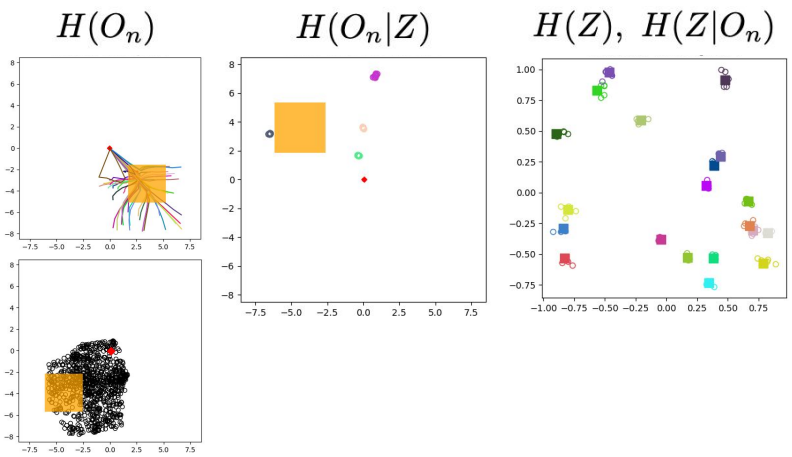

Figure 6: Entropy visualizations for the Agent Observation domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The top figure on the left column shows agent trajectories from randomly selected skills $z \sim p(z)$. The bottom figure on the left column marks skill-terminating 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.

- Observation space: $o_t \in \mathbb{R}^4$ representing (x position, y position, hallway direction flag $\in \{-1, 0, 1\}$, part of hallway).
- Non-Markov Detail: During evaluation episodes, agent is only provided hallway direction flag $\in \{-1, 1\}$ after first timestep of episode. For the remainder of episode the agent receives no signal (i.e., flag = 0), including the initial 15 time steps when the agent is frozen and then when the agent executes skills
- Skill length $n = 8$ actions
- Curriculum: During training episodes we vary the number of timesteps the binary signal lasts from 1 (same as evaluation episodes) to 16.

2. Agent Observation

- Action space: $a_t \in \mathbb{R}^2$ representing change in (x, y) directions
- State space: $s_t \in \mathbb{R}^6$ representing (agent x position, agent y position, agent goal achieved boolean, other agent x position, other agent y position, other agent goal achieved boolean)
- Observation space: $o_t \in \mathbb{R}^6$ representing (agent x position, agent y position, agent goal achieved boolean, other agent x position, other agent y position, other agent goal achieved boolean)
- Non-Markov Detail: During evaluation episodes, agent is not explicitly given the goal as part of each observation. Instead, agent must remember location of goal from when the first agent's boolean flag turned true.
- Skill length $n = 8$ actions
- Curriculum: During training episodes, the number of timesteps the primary agent watches the first agent achieve the goal is randomly sampled between 5 (no change

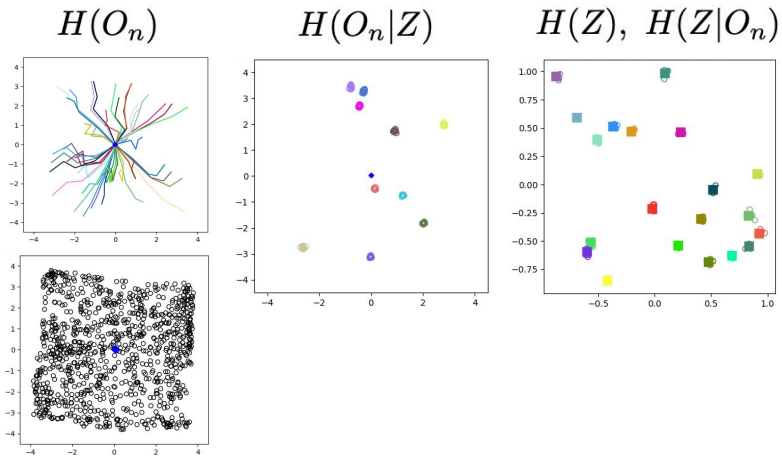

Figure 7: Entropy visualizations for the 4-Bit Password domain. $H(O_n)$ visualizes the distribution of skill-terminating observations in two ways. The top figure on the left column shows agent trajectories from randomly selected skills $z \sim p(z)$. The bottom figure on the left column marks skill-terminating 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.

```
Sig 1 -- Layout: 1.0 Rnn: [-0.23484476 -2.06069517 -4.08114624  2.25323296]
Sig 1 -- Layout: -1.0 Rnn: [-2.63393641 -1.66573095 -3.6819272  -0.27315095]
```

Figure 8: 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.

```
Moving Up:
[-0.49749789 -0.93786877 -2.87412715  2.63446045]
[-0.29751292 -1.91448665 -3.93166208  2.10743904]
[-0.23324569 -2.00563264 -4.01610565  2.31694293]
[-0.23506023 -2.08997846 -4.1116128   2.23406148]
[-0.23652913 -2.05510688 -4.07345724  2.25471115]
[-0.2363535 -2.06318283 -4.08094501  2.25131845]
[-0.23362467 -2.06418157 -4.08013201  2.24997282]
[-0.23579334 -2.06406665 -4.0803256   2.25008869]
[-0.2351567 -2.06493998 -4.08064127  2.24944496]
[-0.23834139 -2.06459117 -4.08212185  2.24908352]
[-0.23488787 -2.06317019 -4.08010769  2.25063372]
[-0.23652785 -2.06470728 -4.08016396  2.25212812]
[-0.23906901 -2.06493497 -4.0828886   2.24776125]
[-0.23364563 -2.06278539 -4.08017921  2.25107384]
[-0.23754962 -2.06250334 -4.08210468  2.25056601]
[-0.2350103 -2.06085324 -4.0799036  2.25142694]
[-0.06848393 -2.29705667 -4.39390373  2.12585759]
[ 0.05238275 -2.06085706 -4.21121836  2.19322991]
[-0.13830297 -0.77669007 -2.94826055  2.18127394]
[-0.31031376 -0.61871356 -2.66338158  2.32413578]
[-0.32300031 -0.62128073 -2.60353923  2.57405639]
```

```
Moving Down:
[-2.31810951 -1.52632022 -3.4826231 -0.23749989]
[-2.54822302 -1.80408156 -3.81838608 -0.1172808]
[-2.64001703 -1.62216401 -3.63861322 -0.31044149]
[-2.62950516 -1.68274856 -3.69879055 -0.25835577]
[-2.63676572 -1.65887713 -3.67590928 -0.27876124]
[-2.63515925 -1.66742074 -3.68345118 -0.27241755]
[-2.63478279 -1.66432679 -3.67966747 -0.27475053]
[-2.63460016 -1.66476321 -3.6807487  -0.27458975]
[-2.63507891 -1.66569364 -3.68155527 -0.27288017]
[-2.63464022 -1.66517341 -3.68110418 -0.27455643]
[-2.6337235  -1.66652119 -3.68180609 -0.27382943]
[-2.6347599  -1.66467273 -3.68161654 -0.27445176]
[-2.6340549  -1.66627395 -3.68217063 -0.27348232]
[-2.63441396 -1.66512251 -3.68001628 -0.27503589]
[-2.63471007 -1.66433811 -3.68153048 -0.27406731]
[-2.63503098 -1.66665983 -3.68210936 -0.27305567]
[-2.2394259  -1.93858755 -3.98519397 -0.01519838]
[-2.25218153 -1.92490113 -3.97197533 -0.0203936]
[-2.22122645 -0.86268705 -2.86663032 -0.33419904]
[-2.11886477 -0.50129908 -2.48987508 -0.48180541]
[-2.4081471 -0.62655169 -2.61125755 -1.15698278]
```

```
Differences:
[1.8206116  0.58845145 0.60849595 2.8719604 ]
[2.25071001 0.11040509 0.113276   2.22471976]
[2.40677142 0.38346863 0.37749243 2.62738442]
[2.39444494 0.4072299  0.41282225 2.49241734]
[2.40023661 0.39622974 0.39754796 2.5334723 ]
[2.39880586 0.39576209 0.39749384 2.523736  ]
[2.40115809 0.39985478 0.40046453 2.52472329]
[2.39880681 0.39930344 0.3995769  2.52467847]
[2.39992213 0.39924634 0.399086   2.52232504]
[2.39629889 0.39941776 0.40101767 2.52363992]
[2.39883566 0.396649  0.3983016 2.52446318]
[2.39823198 0.40003455 0.39854741 2.52657986]
[2.39498591 0.39866102 0.40071797 2.52124357]
[2.40076828 0.39766288 0.40016294 2.5261097 ]
[2.39716053 0.39816523 0.40057421 2.52463341]
[2.4000206  0.39419341 0.39779425 2.52448273]
[2.17094207 0.35846913 0.40870976 2.14105606]
[2.30456424 0.13595593 0.23924303 2.21362352]
[2.08292341 0.08599699 0.08163023 2.51547289]
[1.80855107 0.11741447 0.1735065  2.8059411 ]
[2.0851469  0.00527096 0.00771832 3.73103905]
```

Figure 9: 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.

```
Goal: [4.466284  3.0631247] Rnn: [ 2.1585896 -1.61340892  6.01111889 -2.53100157]
Goal: [-2.66073    1.5370283] Rnn: [ 2.03518224 -1.61517227  2.71821499  0.89709771]
Goal: [3.1659575 1.8680601] Rnn: [ 2.17297125 -1.62536621  6.02287197 -2.54097128]
Goal: [-3.1534185 -4.1778483] Rnn: [-0.05861431  0.60840976  1.05691183  2.31356192]
```

Figure 10: Figure shows the different RNN representations for different episode goals indicated by the movements of the first agent in the Agent Observation 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.

```
Password [-0.92452395  0.77149785 -0.94900286  0.26246145] Rnn: [-1.53673005  0.37508351 -2.77917409 -0.31819582]
Password [-0.69156116 -0.32220381 -0.41143382  0.42358604] Rnn: [-1.59388757 -0.26023102 -3.44570303 -0.82250822]
Password [ 0.80746067 -0.9336499   0.47343686 -0.270945  ] Rnn: [-0.90927458 -0.31577894 -2.48568583 -1.02019918]
Password [ 0.28879797  0.3507107   0.2577195  -0.38209605] Rnn: [-0.99723798  0.30604365 -2.1147573  -0.5762825 ]
```

Figure 11: 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.

from evaluation episodes) to 10 (only goal achieving agent and not the decoy agent is observed).

3. 4-Bit Password
   - Action space: $a_t \in \mathbb{R}^3$ representing change in (x, y) directions and one dimension representing each "bit" of the password output by the agent. Agent outputs a continuous number for the password bit. If this bit and the same bit of the password are both positive or both negative, the bit is counted as correct.
   - State space: $s_t \in \mathbb{R}^7$ representing the four bit password, the number of bits the agent has output correctly, and the (x, y) position of the agent.
   - Observation space: $o_t \in \mathbb{R}^4$ representing (agent x position, agent y position, single bit of password, flag indicating whether password is correct (1), incorrect (-1), or incomplete (0)). Note that the flag indicating whether password is correct or not is only provided *after* the four timesteps when agent attempts to enter the correct password. Otherwise, the agent just received the incomplete flag.
   - Non-Markov Detail: Password is only provided 1 bit at a time, and actions are only entered 1 bit at a time. Thus, agent is not provided multiple bits of the password or shown the bits it has entered so far.
   - Skill length $n = 8$ actions
   - Curriculum: During training episodes, we assist the agent by executing the correct bit and providing the next bit (through an extra dimension of the observation) for a randomly selected number of timesteps between 0 (i.e., no help is provided similar to evaluation episodes) and 3 (i.e., we enter the first three bits correctly regardless of what the agent enters and then provide the next bit through an observation).