THE GEOMETRY OF EMPOWERMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Empowerment captures the capacity for an agent to actively control its environment. While conceptually appealing as an information-theoretic quantity, the connection between empowerment and structurally central states that provide broad access to future outcomes has remained an open question. In this work, we link empowerment maximization and skill-learning methods to provide new geometries for interpreting and analyzing empowerment. Our analyses answer longstanding open questions on the connections between empowerment and structural centrality. Our analyses also reveal distinctions between information and reward geometries, highlighting important theoretical implications to build scalable empowerment-maximization methods.[1]

1 INTRODUCTION

An intelligent agent should be in control of its future. For example, a well-trained hiker at a mountain can (1) choose and traverse down one of many trails from a mountain peak (Fig. 1(b)) *or* (2) navigate to and stay at the peak (Fig. 1(a)). While both useful, these behaviors are distinct: the first behavior characterizes "actively exerting control" while the latter characterizes "being in control."

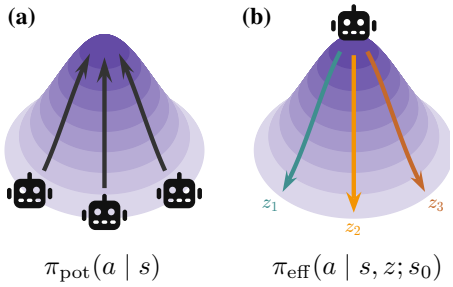


Figure 1: **Two trained empowerment-maximizing policies characterize central states in an MDP:** Consider an agent (robot) on a hill that is easier to descend than to ascend. In Fig. 1(a), the potential policy ascends the empowerment landscape to the top of the hill, using $\mathcal{E}_{\text{pot}}(s)$ as an intrinsic reward. In Fig. 1(b), The effective policy π_{eff} maximizes $I^{\pi}(Z; S_+ \mid S_0)$ where each skill z is a downhill route.

Cleanly parameterizing, optimizing, and separating these behaviors requires a tractable mathematical formulation of controllability. **Empowerment** is an information-theoretic object that quantifies controllability in terms of bits. As originally defined in (Klyubin et al., 2005a), **potential empowerment** measures the maximal bits an agent's actions can inform its future state from some current state. In other words, the potential empowerment measures how much control an agent *could* command from a given state. An alternative measure of empowerment, termed here as **effective empowerment**, is a mutual information that quantifies how readily an agent's policy *actively* informs its future state (Capdepuy, 2011). While sometimes conflated, these two objects are different. In the example of the mountain in Fig. 1, a potential empowerment policy ascends the mountain to a state from which the agent has options. Meanwhile, the effective empowerment policies exercise options and descend the mountain in different directions.

Prior works conjecture that empowerment is a generic way for agents to do goal-oriented tasks. Salge et al. (2014, §4.3.3) call this the AI Empowerment Hypothesis:

> *Empowerment provides a task-independent motivation that generate AI behaviour which is beneficial for a range of goal-oriented behaviour.*

However, empowerment does not have immediate obvious connections to downstream task performance. Indeed, while early works in empowerment suggest a link between empowerment and preferred or central states (Anthony et al., 2008; Jung et al., 2011; Salge et al., 2014), to the best of our knowledge, formally characterizing notions of MDP centrality and adaptation in relation to empowerment has remained a longstanding open problem. In this work, we formally show high potential empowerment states lie at centers of reachable regions in the state occupancy simplex

[1] Website: `https://tinyurl.com/27lg7ncl`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Theorem 4.1). The empowerment objective and long horizon generalizations also admit exact hitting time interpretation (Lemma 4.2) and capture centrality in a temporal distance geometry (Myers et al., 2024a). In addition to providing intuition, these notions of centrality directly lower bound an agent's ability to adapt to downstream tasks in tabular settings (Theorem 4.3), providing a partial answer to Salge's longstanding hypothesis. Potentially empowered agents can reach central states from which they can adapt, where effective empowerment quantifies the adaptation.

The above results theoretically answer questions in empowerment literature: (1) empowered states are central states and (2) in discrete settings, empowered states are states from which agents can adapt to downstream tasks. Attempting to extend the adaptation result to continuous-state MDPs reveals that information and reward geometries can diverge: an arbitrarily empowered state can provide arbitrarily small adaptation in a continuous MDP with continuous rewards (Proposition G.2). The same misalignment appears in discrete settings once the reward prior is anisotropic (Proposition H.1). Thus, empowered agents are *not generally* adaptive agents in continuous-state MDPs. Our work shows scalable formulations of empowerment maximization must address this discrepancy between information and reward adaptation geometries.

2 Related Work

As originally introduced, the empowerment at a state is the maximum possible information that can be communicated from open-loop action sequences to future states (Klyubin et al., 2008), though closed-loop variants have also been discussed (Salge et al., 2014; Mohamed and Jimenez Rezende, 2015).

An empowerment maximizing agent should maximize the empowerment of its occupied states. Capdepuy (2011) distinguishes this channel capacity form of empowerment, termed *potential empowerment*, from the mutual information between the action sequences and a future state *realized* by a specific policy, termed *actualized empowerment*. In this work, we refer to this form of empowerment as *effective empowerment*. These quantities are formalized in §. 3.

Where should an agent go if told that it will be asked to solve some soon-to-be-specified task? One approach is to consider the geometry of the state space and navigate to **central, connected states** or **bottleneck states** within the MDP: states from which the agent can easily visit many other states (Sabidussi, 1966). Such notions of MDP centrality have been studied in both reinforcement learning (Jung et al., 2011; Anthony et al., 2008; Şimşek and Barto, 2007; Turner et al., 2023) and neuroscience (Stachenfeld et al., 2014) community, often in the context of exploration. While these notions and definitions of centrality and bottlenecks are appealing, most objects studied in prior works are only well-defined in settings with discrete states and actions. Separately, prior empowerment literature has conjectured that potential empowerment measures the centrality of a state in an MDP (Anthony et al., 2008; Jung et al., 2011). Salge et al. (2014) note that potential empowerment is the log of the total reachable states in discrete, deterministic settings. Jung et al. (2011) mention the possibility that empowerment might identify "gateway" states. Anthony et al. (2008) compare empowerment to "closeness centrality". However, these works do not provide a general formal characterization of the relationship between empowerment and centrality, or even how centrality should be defined in continuous settings. Our work shows how temporal distance (Myers et al., 2024a), previously used for exploration and goal-reaching (Jiang et al., 2025; Myers et al., 2025; Park et al., 2026), relates to centrality and empowerment.

Estimating and Maximizing Empowerment in Practice. Prior works maximize both potential and effective empowerment (Levy et al., 2025; Gregor et al., 2017; Ringstrom, 2022). In particular, mutual information skill learning (MISL) methods maximize quantities like *effective empowerment* in high-dimensional, continuous settings to learn diverse behaviors, or "skills". These methods train skill policies that maximize mutual information between *skill-latents* and outcomes (Gregor et al., 2017; Eysenbach et al., 2018; Zheng et al., 2025; Park et al., 2023). Another recent line of work optimizes *potential empowerment* as objectives for assistance (Myers et al., 2024b; Yang et al., 2026; Ellis et al., 2026), where empowerment quantifies agency.

3 Preliminaries

We will define notions of empowerment in terms of occupancy measures over an Markov decision process (MDP), following a long line of work that casts decision making as probabilistic inference (Barreto et al., 2017; Eysenbach et al., 2020; Dayan, 1993; Gregor et al., 2017; Fu et al., 2018).

3.1 Discounted State Occupancy

MDP problem setting. Our analysis focuses on applying information theory to a fully-observed MDP with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, and environment dynamics $p(s' \mid s, a)$. Let the skill and starting state-conditioned policy be the function $\pi(a \mid s, z; s_0)$ with skill $z \in \mathcal{Z}$.

The latent skill z lives in a continuous skill space $\mathcal{Z} \subseteq \mathbb{R}^d$ (i.e. the surface of a hypersphere), including in tabular MDPs. The policy $\pi(a \mid s, z; s_0)$ is a distribution over actions given the current state s, the latent z, and the start state s_0. We write $\Pi_{\mathcal{Z}}$ for this class of skill- and starting-state conditioned policies, and Π for the class of ordinary policies $\pi(a \mid s)$ that carry no skill. Fixing a skill in a policy of $\Pi_{\mathcal{Z}}$ leaves an ordinary policy in Π. The start-state s_0 is the context of a skill.

Reachable state-occupancy polytope. For some initial state s_0, skill latent $z \in \mathcal{Z}$, and skill and start-state-conditioned policy $\pi(a \mid s, z; s_0)$, the discounted state occupancy measure (DSOM) is

$$p^\pi(s_+ = s \mid s_0, z) \triangleq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = s \mid s_0, z), \tag{1}$$

where $p^\pi(s_t = s \mid s_0, z)$ is the probability of visiting state s under policy π and the MDP dynamics after t time steps when initialized at state s_0 under skill z. Equivalently, the DSOM is the probability measure of the occupied future state S_+ at a geometric termination time $\tau \sim \text{Geom}(1-\gamma)$.

Averaging over the skill distribution $p(z \mid s_0)$ gives the skill-marginal DSOM,

$$p^\pi(s_+ \mid s_0) = \mathbb{E}_{p(z \mid s_0)}\big[p^\pi(s_+ \mid s_0, z)\big], \tag{2}$$

the future-state distribution induced by the policy when the skill is unknown. Every DSOM is a distribution over states within the simplex $\Delta(\mathcal{S})$. For a fixed MDP, an agent can only realize a strict subset of DSOM in the simplex. For a fixed start state s_0, the realizable DSOMs form the *reachable state-occupancy polytope*

$$\mathcal{P}(s_0) \triangleq \left\{ p^\pi(s_+ \mid s_0) \;:\; \pi \in \Pi \right\} \;\subseteq\; \Delta(\mathcal{S}), \tag{3}$$

a convex subset of the simplex whose linear constraints are the Bellman flow equations (see the polytope inside the simplex in Fig. 3a) (Eysenbach et al., 2021). We write $\mathcal{P}(s)$ for the polytope reachable from a generic start state s. Our analysis in §. 4 studies the geometry of this polytope.

3.2 Empowerment

We consider two forms of empowerment: potential (the maximal empowerment that *could* be achieved at a state) and effective (the actualized influence of a particular skill-conditioned policy (i.e., collection of policies) at a given state on the future state distribution).

Potential and Effective Empowerment. Our work highlights the difference between potential empowerment, as originally defined in Klyubin et al. (2005b) and the effective empowerment objects directly maximized in MISL methods (Mohamed and Jimenez Rezende, 2015; Eysenbach et al., 2018; Park et al., 2023; Zheng et al., 2025).

Klyubin et al. (2005b) originally defines (potential) empowerment as the degree to which action sequences of length H from a current state (i.e. input of the information channel) control the future state (i.e. output). Formally,

$$\mathcal{E}_{\text{pot}\,H}(s) = \max_{p(a_0^{H-1})} I(A_0^{H-1}; S_H \mid S_0 = s)$$

where $A_0^{H-1} \triangleq (A_0, \ldots, A_{H-1})$ denotes the action sequence. In our work, we make two changes. First, rather than considering entire sequences of actions as inputs (i.e. open-loop controls), we instead consider latents z that parameterize a *reactive* policy. We refer to these skill-conditioned, closed-loop controls as "skills". Potential empowerment then measures the maximal influence these skills exert on a future state. Second, rather than measure these skills' influence on a state H steps in the future, we measure how much skills control states at a geometric stopping time, so the discount factor sets the timescale of the empowerment.

Definition 3.1 (Potential empowerment)**.** Let policies $\pi \in \Pi_{\mathcal{Z}}$ condition on skill latents $z \in \mathcal{Z}$ and let S_+ denote the random state $s_+ \sim p^\pi(s_+ \mid s_0, z)$ sampled from the DSOM. Potential empowerment is the channel capacity over a *family of policies* dependent on an initial state:

$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi \in \Pi_{\mathcal{Z}}} I^\pi(Z; S_+ \mid S_0 = s_0).$$

Definition 3.2 (Potential empowerment policy)**.** A potential empowerment policy $\pi_{\text{pot}} \in \Pi$ that takes the form $\pi_{\text{pot}}(a \mid s)$ maximizes the potential empowerment as a reward:

$$\pi_{\text{pot}}{}^{*} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi} \left[\sum_{t} \gamma^{t} \mathcal{E}_{\text{pot}}(s_t) \right].$$

While conceptually similar, mutual information skill learning methods do not maximize potential empowerment, and maximize effective empowerment. *Effective empowerment* captures the mutual information (MI) itself:

Definition 3.3 (Effective empowerment)**.** Consider skill and start state-conditioned policies $\pi \in \Pi_{\mathcal{Z}}$ that take the form $\pi(a \mid s, z; s_0)$. *Effective empowerment* is the mutual information realized by a particular skill and start state-conditioned policy at a particular starting state s_0 under the skill source $p(z \mid s_0)$:

$$\mathcal{E}_{\text{eff}}(\pi, s_0) = I^{\pi}(Z; S_+ \mid S_0 = s_0) = \mathbb{E}_{p(z|s_0)} D_{\text{KL}}[p^{\pi}(s_+ \mid s_0, z) \| p^{\pi}(s_+ \mid s_0)]$$

Correspondingly, we refer to the skill and start state-conditioned policy $\pi(a \mid s, z; s_0)$ as the *effective empowerment policy*. This objective is similar to the objectives directly maximized in modern skill-learning algorithms (Gregor et al., 2017; Eysenbach et al., 2018; Zheng et al., 2025); note that MISL methods generally assume a fixed starting state s_0, so the learned skills are specific to s_0. We refer to the effective empowerment policies as skills.

Effective empowerment involves no maximization and measures the MI *achieved* by a particular skill-conditioned policy $\pi \in \Pi_{\mathcal{Z}}$ that begins from a starting state s_0. This MI depends on both the *channel* $z \mapsto p^{\pi}(s_+ \mid s_0, z)$, determined by the policy, and the *source* $p(z \mid s_0)$ set to uniform. Potential empowerment is the maximum-achievable effective empowerment for a given starting state s_0, so that

$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi \in \Pi_{\mathcal{Z}}} \max_{p(z|s_0)} I^{\pi}(Z; S_+ \mid S_0 = s_0). \tag{4}$$

Under a uniform skill prior, the policy can represent any skill source by simply relabeling skills (Lemma D.1). Fixing the source $p(z \mid s_0)$ to uniform therefore loses no generality. We include a full statement of this result in Appendix D. Thus, a learned skill set (effective empowerment policies) acts as a probe of the potential empowerment at a starting state. If skills initialized at s_0 induce easily distinguishable future occupancies $p^{\pi}(s_+ \mid s_0, z)$, then s_0 affords many controllable futures and has high potential empowerment $\mathcal{E}_{\text{pot}}(s_0)$.

Notation convention. Throughout the paper, we write π for the effective policy when context is clear and π_{eff} only when both policies appear in the same expression. The potential policy is always written as π_{pot}. The object $\mathcal{E}(s)$ always refers to potential empowerment measured at starting-state s.

4 The Geometry of Empowerment

Prior work has empirically observed a close connection between geometry and empowerment (Anthony et al., 2008; Jung et al., 2011), wherein states that are more centralized in an MDP have higher empowerment. However, to the best of our knowledge, such connections have not been proven theoretically. In this section, we show that empowerment is closely related to a probabilistic notion of centrality in information geometry (Polyanskiy and Wu, 2022) and temporal distance geometry (Myers et al., 2024a):

- §. 4.1: Effective empowerment corresponds to the information-theoretic distance between states a *specific* skill-conditioned policy may reach in the future. Potential empowerment is the maximum over these distances from a starting state, measuring the radius of a reachable space.
- §. 4.2: High empowerment states are central in temporal distance geometry. From a current state, the potential empowerment measures the average time saved to reach future states when committing to a skill.

We then connect empowerment to adaptation on downstream tasks (§. 4.3). We show a positive result, a lower bound on adaptation, but also prove important limitations that highlight the fundamental difference between information and reward geometries:

- §. 4.3, Theorem 4.3: The potential empowerment of a state lower bounds adaptation to unknown reward functions drawn from an isotropic distribution.

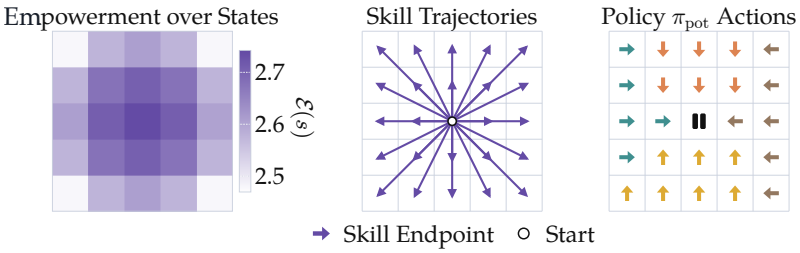


Figure 2: **Skills measure the centrality of a state.** We visualize the potential empowerment field $\mathcal{E}(s)$ (a) for each state in a tabular grid. The central state has high potential empowerment, because (b) the central state gives an agent access to a maximal set of distinguishable skills. A potential empowerment policy ascends the potential empowerment landscape to the center state in (c), from which it can execute different skills.

- §. 4.3, Proposition G.2: In continuous settings, a state can have unbounded potential empowerment while providing arbitrarily small adaptation to a continuous reward.
- §. 4.3, Proposition H.2: For anisotropic reward distributions, an empowered state can provide no adaptation.

4.1 Information Geometry of Potential and Effective Empowerment Policies.

Geometric intuition. Prior work has found empirically that high empowerment states are central states in an MDP, relating empowerment to the volume of reachable states in discrete, deterministic settings (Salge et al., 2014). Here, we extend definitions of centrality to the full MDP setting with stochasticity and continuous states. Effective empowerment is a mutual information. In the dual form, effective empowerment maximization finds state occupancy distributions that lie at the center of extremized state distributions. Then, potential empowerment-seeking policies search for the states that lie at the centers of the largest radii reachable polytopes.

Before presenting formal results, we visualize the information geometry of empowerment in Fig. 3. Let the triangle represent the state occupancy simplex reachable from different states s_1 and s_2. Maximizing the mutual information (effective empowerment) for a given starting state corresponds to learning skills that reach vertices of the set of reachable state occupancies. The agent maximizes effective empowerment separately at every possible starting state (orange for s_1, blue for s_2) and learns skills (orange and blue arrows) that span reachable regions. Then, to maximize potential empowerment, an agent searches among these starting states for the largest reachable region. In this case, the blue state s_2 corresponds to the state with the highest potential empowerment and occupies a more centralized point within the information geometry.

High Empowerment States are Central in Information Geometry. Formally, consider the state probability simplex $\Delta(\mathcal{S})$, where DSOM are elements of the simplex $p^\pi(s_+ \mid s_0, z) \in \Delta(\mathcal{S})$. A straightforward extension of prior work (Eysenbach et al., 2021) shows that effective empowerment maximization is a center localization problem in the initial state s_0-conditioned state probability simplex, where distance is given by the Kullback–Leibler divergence (Megiddo, 1983). Then, potential empowerment measures the min-max distance to the reachable state-occupancy polytope:

Theorem 4.1 (Potential empowerment is the Kullback-Leibler (KL) radius of the reachable polytope, extending Eysenbach et al. (2021, Lem. 6.1–6.2))**.** *Let $\mathcal{P}(s_0) \subset \Delta(\mathcal{S})$ be the reachable state-occupancy polytope of Eq. (3). Jointly maximizing the mutual information over the skill source and the skill-conditioned policy solves the KL 1-center localization problem for $\mathcal{P}(s_0)$:*

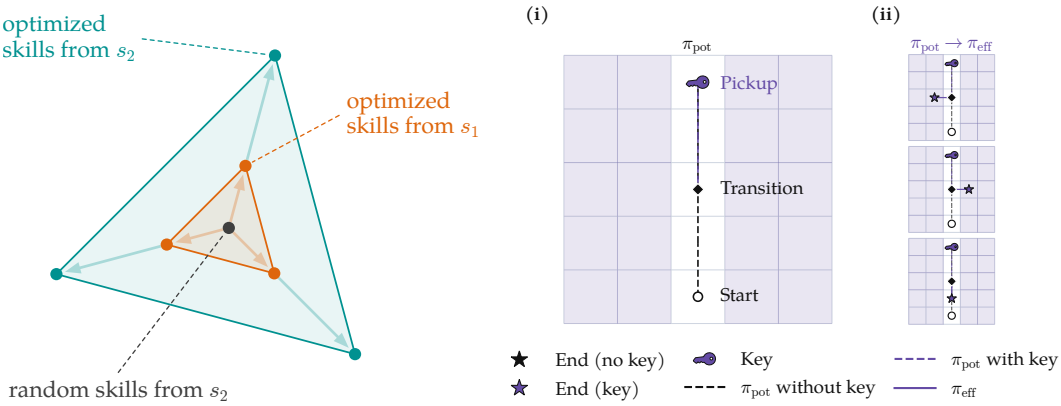
$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi \in \Pi_{\mathcal{Z}},\, p(z|s_0)} I^\pi(Z; S_+ \mid S_0 = s_0) = \min_{q(s_+) \in \Delta(\mathcal{S})} \max_{p \in \mathcal{P}(s_0)} D_{\text{KL}}\big(p(s_+) \| q(s_+)\big). \quad (5)$$

Moreover, every joint optimum $(\pi, p(z \mid s_0))$ places every skill on the boundary of the optimal KL ball measured from the optimal central occupancy $q^\star$:

$$D_{\text{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big) = \max_{p \in \mathcal{P}(s_0)} D_{\text{KL}}\big(p(s_+) \| q^\star(s_+)\big) \quad \textit{for every skill } z. \quad (6)$$

In other words, the optimized skills span the maximal information radius of the reachable polytope given starting state s_0.

We prove a general version of Theorem 4.1 in Appendix B. This result shows that an exploration agent maximizing potential empowerment will seek central states in the information geometry. Prior work on empowerment was correct in conjecturing that empowerment maximization has a formal connection to centrality. However, the right geometry for studying this centrality is not (say) the Euclidean metric on states, but rather the KL divergence between state distributions induced by different skills.



(a) **State occupancy simplex** (b) **Potential policy finds central states**.

Figure 3: **Potentially empowered states are central states in the polytope.** (a) The skills starting from s_1 define a smaller polytope in the simplex than s_2 as a result of their transition dynamics, so the potential empowerment is higher at s_2 than s_1. (b) A potentially empowered agent picks up the key then navigates to the center of the room, correctly identifying the middle state *with the key* as the most central state.

A simple tabular experiment demonstrates this notion of centrality (Fig. 3b). In this MDP, the agent can pick up and drop off a key at the the bottom of the hallway in addition to {Left, Right, Up, Down, Stay} actions. Picking up this key allows access to the blue-shaded regions of the MDP; without the key, the agent cannot enter these spaces in the grid. The potential empowerment policy π_{pot} that ascends the potential empowerment landscape correctly reaches the true central state: the center of the MDP with the key. This experiment shows that states with maximal potential empowerment capture an information-geometric notion of centrality, rather than navigating towards the *physically* central state. See Appendix J.3 for experimental details.

4.2 Empowerment, Bottlenecks, and Temporal Distance

Beyond information geometry, the empowerment objective has a direct hitting time interpretation in discrete state spaces. Suppose that the agent has multiple trial trajectories to reach a future state, where each trajectory rolls out a skill from the same starting state until a geometric termination time. The empowerment measures the reduction in hitting time when the agent commits to a skill:

Lemma 4.2 (Bottleneck states are high potential empowerment states)**.** *In a discrete MDP, let the hitting time random variable $H^{\pi_{\text{eff}}}(s_0 \to s_+)$ be the number of rollouts from s_0 until a rollout hits s_+, for skills $Z \sim p(z \mid s_0)$. Let $H^{\pi_{\text{eff}}}(s_0 \to s_+ \mid z)$ be the corresponding hitting time when rolling out z every time. Then,*

$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi_{\text{eff}} \in \Pi_{\mathcal{Z}}} \mathbb{E}_{p^{\pi_{\text{eff}}}(z, s_+ \mid s_0)} \left[\log \mathbb{E}[H^{\pi_{\text{eff}}}(s_0 \to s_+)] - \log \mathbb{E}[H^{\pi_{\text{eff}}}(s_0 \to s_+ \mid z)] \right].$$

We prove Lemma 4.2 in Appendix C. In tabular settings, potential empowerment measures the reduction in time to reach future states *over independent rollouts* when committing to an option/skill.

However, hitting times are not well-defined in continuous geometries, where hitting a particular state is a measure-zero event. The temporal distance is an object that remains well-defined in continuous settings and reduces to hitting times in the discrete limit (Myers et al., 2024a). In the appendix, we show that a closely related empowerment objective over *contiguous* (versus repeated) skill rollouts, where the agent re-samples a skill at each geometric termination in a single trajectory, admits an exact temporal distance formulation (Myers et al., 2024a). In that construction, the empowerment object measures the reduction in temporal distance to future states after committing to a skill (Proposition C.1), recovering a hitting-time interpretation in the discrete, deterministic limit.

Conceptually, these temporal formulations capture the bottleneck effect. From a bottleneck, an agent must commit to different skills in order to quickly reach different downstream regions which is conceptually the reduction in temporal distance and hitting times considered in our results. Tabular experiments in Fig. 4 illustrate this effect (Stachenfeld et al., 2014; Şimşek et al., 2005). Adding slim openings ("bottlenecks") to the grid shifts the high empowerment states from the original central state to the bottleneck openings, prioritizing states where committing to a skill (e.g. move left or move right) quickly reaches different regions. This interpretation also connects empowerment to closeness centrality (Anthony et al., 2008).

4.3 Is Empowerment Optimal for Downstream Adaptation?

Potential empowerment lower bounds skill adaptation under an uninformed reward prior. While empowerment is often introduced as an appealing "universal" (Klyubin et al., 2005b) objective, is there a precise sense in which empowerment measures the ability to adapt? In Appendix F, we formally prove that empowerment lower bounds skill adaptation to rewards drawn from an uninformed prior.

Because there are no target rewards during training, we must define an adaptation objective with respect to a reward prior distribution $\mathcal{R}$. A simple reward prior is a Gaussian distribution centered at a fixed baseline distribution, set to the uniform occupancy over the state space. Prior works use similar isotropic reward distributions when proving general downstream adaptation results (Ollivier, 2025).

Here, we define adaptation as a *skill adaptation* to a sampled reward. At test time, an agent must select a skill $z \in \mathcal{Z}$ to adapt to a sampled reward function $r \sim \mathcal{R}$. For example, in the case of the 5×5 grid, an agent should select the skill "go left" if the reward provided at test time is on the left.

Theorem 4.3 (Informal statement that tabular empowerment lower bounds reward adaptation)**.** *Fix a start state s_0 and sample a reward $r \sim \mathcal{R}$ from the uninformed Gaussian prior. The agent adapts by selecting the skill with the highest mean return under r, and the adaptation $J(s_0)$ averages the best return over the reward prior. In whitened coordinates, write g for the reward and u_z for the offset of skill z's occupancy from the skill-marginal DSOM of Eq. (2), so the return gap of skill z over the marginal is $\langle g, u_z \rangle$. Then, for any skill space and skill prior,*

$$J(s_0) = \mathbb{E}_g \sup_{z \in \mathcal{Z}} \langle g,\ u_z \rangle \ \gtrsim\ I(Z; S_+ \mid S_0 = s_0).$$

In the best case, the agent selects the skill z with the highest expected return; note that the agent can do no better within the space of learned skills. This type of policy selection for adaptation/transfer has precedent in prior successor feature literature (Barreto et al., 2017; 2018). For the following result, we consider linear adaptation objective J for the skill set learned by an empowerment-maximizing agent, which is exactly the best skill's mean return averaged over the reward prior. This adaptation objective is notably different from prior results connecting MI and adaptation in Eysenbach et al. (2021), which considers regularized adaptation directly in the state occupancy. Here, the adaptation is the selection of a *pre-trained* skill behavior, which does not assume the ability to directly adapt state occupancy distributions:

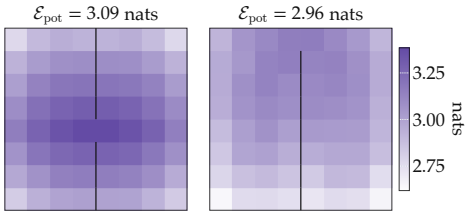


Figure 4: **Bottleneck states are high potential empowerment states.** Consider a grid with a door, or a "bottleneck," in the middle. We visualize how shifting the door position from the center (a) to the corner (b) changes the location of the high-empowerment states to the door openings. In §. 4.2 we show that empowerment measures the reduction in temporal distance to future states, which quantifies a bottleneck effect. See Appendix J.2 for experimental details.

We give the formal statement and proof in Theorem 4.3 (Appendix F), where $J(s_0)$ turns out to be the Gaussian width of the skill offsets $\{u_z\}_{z \in \mathcal{Z}}$ (Eq. (12)). Intuitively, for the simple 5×5 grid (Fig. 2), the central maximal potential empowerment state is *exactly* the state from which an agent can easily adapt to the left/right/upper/lower rewarding states in the grid. Thus, maximizing potential empowerment provides a new form of provable adaptation where agents navigate to central states *from which* they can later adapt, which existing skill-learning or effective empowerment methods do not capture on their own (Gregor et al., 2017).

Two didactic tabular experiments demonstrate the behavior of the bound across MDP parameters. The first set of MDPs varies the degree of commitment within an MDP, which measures how much probability mass an action places on a target state above uniform. Commitment measures the controllability of the MDP. As controllability increases, the bound tightens (Fig. 9). The second set of MDPs varies the number of possible distinct skills/actions while holding commitment fixed. As the number of possible skills grows, the bound loosens (Fig. 10). The bound is therefore most informative in MDPs where an agent exerts strong control over comparatively few distinguishable regions. Full experimental details are in Appendix F.

The First Limitation of Empowered Adaptation: Gaussian widths versus KL divergence. While the above result demonstrates that potential empowerment lower bounds adaptation in discrete MDPs, there remain open questions: (1) is the bound tight? and, correspondingly, (2) does the bound

extend to continuous geometries? The answer to both of these questions is generally no: without additional assumptions, the geometry of *reward adaptation* and that of information geometry are different.

First, we must define the geometry of reward adaptation. *Gaussian width* measures the spread of the skills' future occupancies and governs the adaptation geometry (Polyanskiy and Wu, 2022). Because the cumulative reward is *linear* in the occupancy, the width is large when different skills reach futures with clearly different values (Fig. 5, (b,e)). This quantity differs from MI, where distances are given by the Kullback–Leibler divergence. Our first observation is that the adaptation result is a bound, not a ranking result. Consider the simple example of two starting states in a discrete MDP (Fig. 5, (a) and (d)). The first starting state in (a) can reach two states. The second starting state (d) can reach three states. Under an isotropic reward prior, an agent optimizing adaptation should choose the second starting state (Fig. 5, (e)) according to the Gaussian-width geometry that characterizes reward adaptation. Meanwhile, an agent optimizing the information radius will choose the first starting state (Fig. 5, (c)). This discrepancy is a direct consequence of the difference between information and reward adaptation geometry. Thus, while the lower bound holds, the bound does not guarantee a ranking result with provable counterexamples.

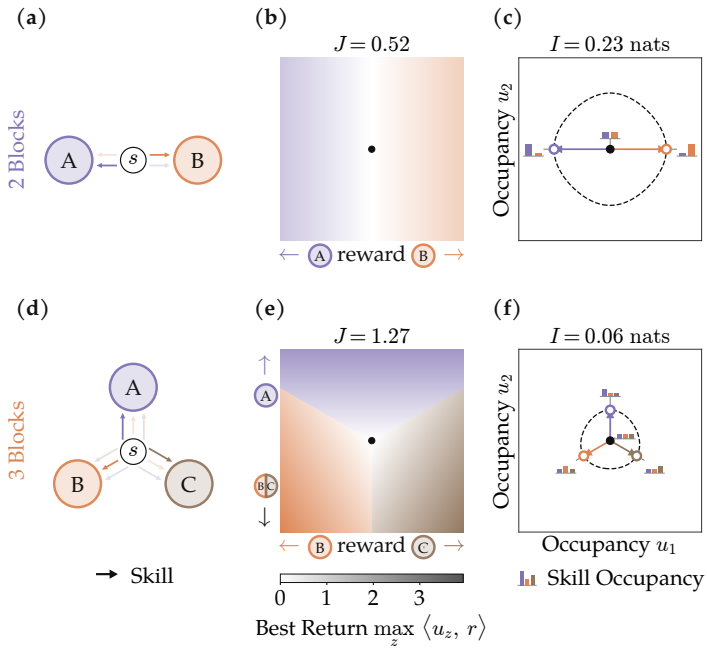


Figure 5: **Discrepancy between information geometry and reward adaptation geometry in discrete settings.** (a) By going left/right from s, an agent can reach node s_1 or s_2 with high likelihood, so there are two easily separable skills from s. (d) Targeting nodes s_1, s_2, or s_3 can make the agent inadvertently land in other nodes, so there are three somewhat separable skills from s. (b,e) The MDPs in (a) and (d) admit polytopes of feasible state occupancies in the simplex, visualized by the segment and triangle. (b) Axis u_1 trades off occupancy of s_1 vs. s_2. (e) Axis u_1 trades off occupying s_2 vs. s_3.; axis u_2 trades off occupying s_1 vs. s_2/s_3. The color indicates the best adapted skill for that reward coordinate. The polytope in (e) has a larger Gaussian Width than that in (b), supporting better adaptation. (c,f) However, increased skill separability in (a) gives larger information radius and MI, where the black contour indicates the maximal MI. Thus, even though empowerment is higher for the 2 state system (a) versus the 3 state system (d), the 2 state system has worse adapted returns. See Appendix J.4 for details.

In the continuous state limit, the difference between empowerment and adaptation geometry becomes more salient:

Theorem 4.4 (Informal statement that large MI does not necessarily imply adaptation to Lipschitz-continuous rewards in continuous state MDPs)**.** *For every skill count K and ϵ, there exists a continuous state MDP admitting K distinct skills that induce DSOMs within an ϵ-ball in arbitrary continuous state space.*

Taking $K \to \infty$, the MDP admits a set of skills that induce unbounded empowerment while adaptation to a Lipschitz-continuous reward can be arbitrarily small.

We give the formal statements and proofs in Propositions G.1 and G.2. Our result shows that skills can have unbounded distinguishability but fail to adapt to Lipschitz-continuous rewards, inducing identical returns across the skill set. Fig. 6 illustrates how information geometries can lead to collapsed returns in a continuous 2D state space, where the potential empowerment grows from left to right while returns homogenize. Thus, larger potential empowerment does not imply better adaptation to smooth rewards in continuous state settings. We fully discuss this collapse in Appendix G.

The Second Limitation of Empowered Adaptation: Controllable Degrees of Freedom versus Rewarding Degrees of Freedom. The second limitation of empowered adaptation is the misalignment

between controllable degrees of freedom and preferred degrees of freedom. While the adaptation result of Theorem 4.3 requires an isotropic reward prior, naive empowerment maximization cannot generically provide adaptation guarantees for anisotropic reward priors. Without external information, empowerment can only break symmetry across states via controllability. Then, the reward prior is free to vary in directions in the state occupancy simplex that empowerment does not naturally select:

Theorem 4.5 (Informal statement that large MI need not imply adaptation to anisotropic reward function priors in discrete settings)**.** *A state can have arbitrarily large MI while adaptation with respect to an anisotropic reward prior is arbitrarily close to zero.*

The formal statements and proofs are in Appendix H (Propositions H.1 and H.2). We can reinterpret Fig. 5 as an example where the maximizing MI selects for skills that do not control degrees of freedom relevant to the anisotropic return: if the variation in an anisotropic reward prior concentrates on three absorbing states in Fig. 5(d), an empowerment-maximizing agent will ignore those reward-relevant states and, instead, navigate towards the state with maximal channel capacity. Therefore, the skill set is suboptimal for adaptation to the anisotropic reward prior.

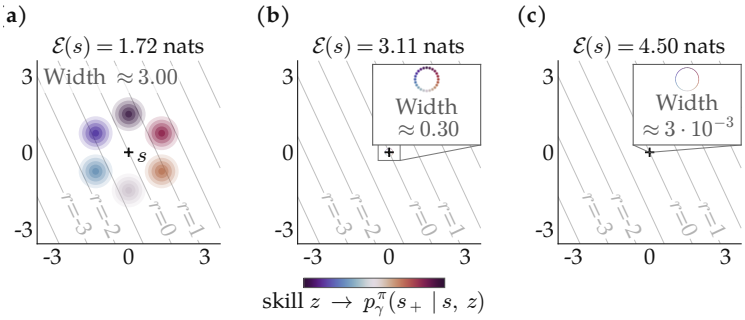


Figure 6: **The information ball can collapse in the continuous limit** (Proposition G.2)**.** Consider a smooth reward function (values in gray) and some MDP that admits Gaussian skill state occupancy measures (colored distributions). While the potential empowerment of the starting state $s \in \mathcal{S}$ increases from (a) $\rightarrow$ (c), the skills have increasingly identical returns and, thus, provide progressively smaller adaptation benefit. Therefore, large channel capacities do not imply adaptation for continuous MDPs and reward functions. See Proposition G.2 for the theoretical result.

5 Discussion, Limitations, & Future Work

In this work, we analyze the geometry of empowerment. Our theoretical results show that high potential empowerment states are central states in information geometry (Theorem 4.1). The empowerment objective and a multistep skill rollout variant also admit exact hitting time interpretations and capture centrality in a temporal distance geometry (Lemma 4.2 and Proposition C.1). Thus, policies that maximize potential empowerment naturally seek out central states within an MDP. Furthermore, effective empowerment policies, or skills, measure the degree of centrality (Fig. 2). Finally, in the tabular setting, this notion of centrality achieved by a set of skills directly lower bounds an agent's ability to adapt to unknown rewards (sampled from an uninformed prior). Our results formally link empowerment and the ability to complete downstream tasks (Theorem 4.3).

Limitations & Future Work. While our results support empowerment as a general objective, information geometry and reward adaptation are fundamentally different: reward adaptation geometry is optimized when working with Gaussian widths, while information-theoretic empowerment makes use of a weaker notion of diversity with KL divergence.

However, these approaches often optimize apples and oranges quantities: the units of extrinsic and intrinsic rewards can be fundamentally different, and our theoretical result highlights an extreme case. Until we consider exploration and adaptation methods that appropriately unify these geometries, there will remain a disconnect between intrinsic reward approaches and the broader RL problem setting. Indeed, our analysis can generically apply to any method that combines extrinsic and information-theoretic intrinsic rewards such as information gain (Houthooft et al., 2017). *We believe that scalably addressing these limitations is key in order to make information-based intrinsic rewards relevant to the broader RL community.*

Reproducibility Statement.

Full hyperparameter tables and methods are provided in the appendix. Our code is available at `https://tinyurl.com/27lg7ncl`

Generative AI Disclosure Statement.

We acknowledge the use of generative AI tools for figure generation, organizing hyperparameter tables, code implementation for the tabular experiments, and minor edits to improve readability. We did not use generative AI to develop the claims of the paper, the theoretical results, nor experiment design. The authors take full responsibility for the correctness of this manuscript, its results, and code.

References

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. *IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135 Vol.1, 2005a.

Philippe Capdepuy. Informational Principles of Perception-Action Loops and Collective Behaviours. 2011.

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

Tom Anthony, Daniel Polani, and Chrystopher L Nehaniv. On Preferred States of Agents: How Global Structure Is Reflected in Local Structure. 2008.

Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for Continuous Agent—Environment Systems. *Adaptive Behavior*, 19(1):16–39, 2011.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning Temporal Distances: Contrastive Successor Features Can Provide a Metric Structure for Decision-Making. *International Conference on Machine Learning*, 2024a.

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):e4018, 2008.

Shakir Mohamed and Danilo Jimenez Rezende. Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. *Neural Information Processing Systems*, 28, 2015.

Gert Sabidussi. The Centrality Index of a Graph. *Psychometrika*, 31(4):581–603, 1966.

Özgür Şimşek and Andrew G Barto. Betweenness Centrality as a Basis for Forming Skills. 2007.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend to Seek Power. aXiv:1912.01683, 2023.

Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. Design Principles of the Hippocampal Cognitive Map. *Neural Information Processing Systems*, 27, 2014.

Yuhua Jiang, Qihan Liu, Yiqin Yang, Xiaoteng Ma, Dianyu Zhong, Hao Hu, Jun Yang, Bin Liang, Bo Xu, Chongjie Zhang, and Qianchuan Zhao. Episodic Novelty Through Temporal Distance. arXiv:2501.15418, 2025.

Vivek Myers, Bill Chunyuan Zheng, Benjamin Eysenbach, and Sergey Levine. Offline Goal-Conditioned Reinforcement Learning With Quasimetric Representations. arXiv:2509.20478, 2025.

Seohong Park, Deepinder Mann, and Sergey Levine. Dual Goal Representations. arXiv:2510.06714, 2026.

Andrew Levy, Alessandro G Allievi, and George Konidaris. Representation Learning and Skill Discovery With Empowerment. *Reinforcement Learning Conference*, 2025.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational Intrinsic Control. *International Conference on Learning Representations*, 2017.

Thomas J Ringstrom. Reward Is Not Necessary: How to Create a Modular & Compositional Self-Preserving Agent for Life-Long Learning. arXiv:2211.10851, 2022.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity Is All You Need: Learning Skills Without a Reward Function. *International Conference on Learning Representations*, 2018.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL Fly? Analysis and Ingredients for Mutual Information Skill Learning. *ICLR 2025*, 2025.

Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable Unsupervised RL With Metric-Aware Abstraction. *International Conference on Learning Representations*, 2023.

Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, and Anca Dragan. Learning to Assist Humans Without Inferring Rewards. *Neural Information Processing Systems*, 2024b.

Claire Yang, Claire Jie Zhang, Maya Cakmak, and Max Kleiman-Weiner. When Assisting One Disempowers Another. arXiv:2511.04177, 2026.

Evan Ellis, Vivek Myers, Jens Tuyls, Sergey Levine, Anca Dragan, and Benjamin Eysenbach. Training LLM Agents to Empower Humans. *International Conference on Machine Learning*, 2026.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor Features for Transfer in Reinforcement Learning. *Neural Information Processing Systems*, 30, 2017.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-Learning: Learning to Achieve Goals via Recursive Classification. *International Conference on Learning Representations*, 2020.

Peter Dayan. Improving Generalisation for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 1993.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational Inverse Control With Events: A General Framework for Data-Driven Reward Definition. *Neural Information Processing Systems*, 2018.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The Information Geometry of Unsupervised Reinforcement Learning. *International Conference on Learning Representations*, 2021.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All Else Being Equal Be Empowered. *European Conference on Artificial Life*, pp. 744–753, 2005b.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, first edition, 2022.

Nimrod Megiddo. The Weighted Euclidean 1-Center Problem. *Mathematics of Operations Research*, 8(4):498–504, 1983.

Özgür Şimşek, Alicia P Wolfe, and Andrew G Barto. Identifying Useful Subgoals in Reinforcement Learning by Local Graph Partitioning. *International Conference on Machine Learning*, pp. 816–823, 2005.

Yann Ollivier. Which Features Are Best for Successor Features? arXiv:2502.10790, 2025.

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in Deep Reinforcement Learning Using Successor Features and Generalised Policy Improvement. *International Conference on Machine Learning*, pp. 501–510, 2018.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational Information Maximizing Exploration. *Neural Information Processing Systems*, 2017.

Thomas M Cover and Joy Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2006.

Alison L Gibbs and Francis Edward Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 2002.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic Symmetries and Invariant Neural Networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99. Springer, Probability Theory and Stochastic Modelling, 2021.

Stas Tiomkin, Daniel Polani, and Naftali Tishby. Control Capacity of Partially Observable Dynamic Systems in Continuous Time. arXiv:1701.04984, 2017.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Approximation of Empowerment in the Continuous Domain. *Complex Systems*, 16(02n03):1250079, 2013.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Suguru Arimoto. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Richard Blahut. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

Jianyi Huang and S.P. Meyn. Characterization and Computation of Optimal Distributions for Channel Coding. *IEEE Transactions on Information Theory*, 51(7):2336–2351, 2005.

Paul J. Goulart and Yuwen Chen. Clarabel: An Interior-Point Solver for Conic Programs With Quadratic Objectives. arXiv:2405.12762, 2024.

A Preliminaries for Proofs

We restate four standard results in forms used in our statements. We cite referenced restatements of these results, which are more general versions of the results below.

The first result gives the standard concavity and convexity properties of mutual information (Cover and Thomas, 2006, Thm. 2.7.4) (Polyanskiy and Wu, 2022, Thm. 5.3).

Theorem A.1 (Concavity and convexity of mutual information)**.** *Let P_X be a source distribution and let $P_{Y|X}$ be a channel between standard Borel spaces. For a fixed channel, $I(X;Y)$ is concave in P_X. For a fixed source, $I(X;Y)$ is convex in $P_{Y|X}$.*

In our setting, the source is the skill prior $p(z \mid s_0)$ and the channel maps each skill z to a DSOM $p^\pi(s_+ \mid s_0, z)$. Thus, the MI $I^\pi(Z; S_+ \mid S_0 = s_0)$ is concave in the skill prior for fixed skill occupancies and convex in the occupancies for a fixed skill prior.

The second result expresses the channel capacity in information radius form as the smallest worst-case divergence from the channel outputs to a central distribution (Polyanskiy and Wu, 2022, Sec. 5.3).

Theorem A.2 (Dual form of channel capacity)**.** *Let $P_{Y|X}$ be a channel between standard Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, and write $P_{Y|X=x}$ for the output distribution at input $x \in \mathcal{X}$. Then*

$$\sup_{P_X \in \Delta(\mathcal{X})} I(X;Y) = \inf_{q(y)} \sup_{x \in \mathcal{X}} D_{\mathrm{KL}}\big(P_{Y|X=x}(y) \,\|\, q(y)\big).$$

If this value is finite, the infimum is attained at a unique center $q^\star$.

In our setting, the input is the skill, $X = Z$, and the output is the future state, $Y = S_+$, so each output distribution $P_{Y|X=x}$ is a skill occupancy $p^\pi(s_+ \mid s_0, z)$. We utilize this dual form to prove Theorem B.2.

The third result compares KL and χ^2 divergence (Polyanskiy and Wu, 2022, Eq. (7.31)); see also Gibbs and Su (2002) for the general statement.

Theorem A.3 (KL and χ^2 comparison)**.** *For any distributions $P \ll Q$,*

$$D_{\mathrm{KL}}(P\|Q) \le \log\big(1 + \chi^2(P\|Q)\big).$$

We apply this bound with

$$P = p^{\pi_{\mathrm{eff}}}(z \mid s_0, s_+) \quad \text{and} \quad Q = p(z \mid s_0).$$

Averaging the bound over s_+ gives the χ^2 control of mutual information that appears in Theorem 4.3.

The fourth result shows that a measurable transformation of noise can induce any distribution on a standard Borel space (Bloem-Reddy and Teh, 2020, Lem. 5) (Kallenberg, 2021, Thm. 8.17).

Theorem A.4 (Noise outsourcing)**.** *Let μ be a probability distribution on a standard Borel space $\mathcal{Z}$. Let U be the uniform distribution on a standard Borel space $\mathcal{V}$. If $\tilde{Z} \sim U$, then there exists a measurable map $T : \mathcal{V} \to \mathcal{Z}$ such that*

$$T(\tilde{Z}) \sim \mu.$$

We utilize this statement to prove that a uniform skill prior does not lose generality in Lemma D.1.

B Information Geometry of Empowerment

We prove a general theorem of the information radius result from §. 4.1 that applies to both continuous and discrete state space MDPs. We include the reduction of the result to the finite-state statement found in §. 4.1.

We start with intuition about the information channel, then explain potential simplications of this information channel. First, note that skill latents $z \in \mathcal{Z}$ can only influence future state distributions via a skill-conditioned policy. The information channel $P_{S_+|Z}$ is a function of the skill-conditioned policy $\pi \in \Pi_{\mathcal{Z}}$, which sends each skill to the induced occupancy:

$$\mathcal{Z} \xrightarrow{\pi} \mathcal{P}(s_0), \qquad z \longmapsto p^\pi(s_+ \mid s_0, z).$$

If this mapping is measurable, we can directly reduce the skills to their induced DSOMs. Specifically, there would exist a pushforward map from the skill random variable Z to a random state occupancy P drawn from a *distribution over* $\mathcal{P}(s_0)$ (e.g. a distribution over state occupancy distributions) which, by the data-processing inequality from the Markov chain $Z \to P_Z \to S_+$ and the sufficiency of P, would imply

$$I(Z; S_+ \mid S_0) = I(P_Z; S_+ \mid S_0).$$

Note, however, that not all policies admit a measurable mapping from skill space to occupancy space. We require the additional assumption that the stochastic policies $\pi(a \mid s, z)$ are jointly measurable over the state and skill space:

Assumption B.1 (Policies measurable in skill and state space)**.** Let $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{Z}$ denote the state space, action space, and skill spaces respectively, which we take to be standard Borel.

Take the skill-conditioned policy space $\Pi_{\mathcal{Z}}$ to be the space of policies measurable in the state space $\mathcal{S}$ and skill space $\mathcal{Z}$. That is,

$$(s, z) \mapsto \pi(\cdot \mid s, z) \in \Delta(\mathcal{A})$$

is a measurable function.

Accordingly, the state occupancy set $\mathcal{P}(s_0)$ is now occupancies reachable by measurable policies.

From this assumption, there exists a measurable mapping from $\mathcal{Z} \xrightarrow{\pi} \mathcal{P}(s_0)$. Thus, we can effectively alias the skill latent Z and the random state occupancy measure P_Z. For the information geometry result, this association allows direct application of Theorem A.2 to the information channel from skills to occupancies:

Theorem B.2 (General information geometry of empowerment)**.** *Fix a start state s_0. Then, with both sides taking values in $[0, \infty]$,*

$$\mathcal{E}_{\mathrm{pot}}(s_0) = \inf_{q(s_+)} \sup_{p \in \mathcal{P}(s_0)} D_{\mathrm{KL}}\big(p(s_+) \| q(s_+)\big).$$

If the empowerment is bounded, the infimum is attainable. Let the skill-conditioned policy $\pi \in \Pi_{\mathcal{Z}}$ and the skill source $p(z \mid s_0)$ over skills $z \in \mathcal{Z}$ attain $\mathcal{E}_{\mathrm{pot}}(s_0)$. Then, the skill marginal is the unique center:

$$q^\star = p^\pi(s_+ \mid s_0).$$

Moreover, for $p(z \mid s_0)$-almost every z,

$$D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big) = \mathbb{E}\mathcal{E}_{\mathrm{pot}}(s_0),$$

where $p^\pi(s_+ \mid s_0, z)$ is an extreme point of $\mathcal{P}(s_0)$. Thus, an optimal policy places every skill occupancy at the maximal KL radius, while an optimal source induces the average occupancy $q^\star$.

We now prove Theorem B.2 with the information-radius identity of Theorem A.2.

Proof. Fix $S_0 = s_0$. Each skill z induces a future state occupancy

$$p^\pi(s_+ \mid s_0, z) \in \mathcal{P}(s_0),$$

and the conditional distribution of S_+ depends on z only through this occupancy. From the measurability assumption Assumption B.1, we may alias skills that induce the same occupancy. Thus, jointly optimizing the policy and skill source is equivalent to choosing a source over $\mathcal{P}(s_0)$ for the channel

$$p \in \mathcal{P}(s_0), \quad S_+ \sim p.$$

In other words, the source is a distribution *over a set of state-occupancy distributions* that give the channel output S_+.

Applying Theorem A.2 with the random occupancy $P_Z \in \mathcal{P}(s_0)$ as the input and the future state S_+ as the output gives

$$\mathcal{E}_{\mathrm{pot}}(s_0) = \inf_{q(s_+)} \sup_{p \in \mathcal{P}(s_0)} D_{\mathrm{KL}}\big(p(s_+) \| q(s_+)\big).$$

Thus, assuming the potential empowerment is finite, a unique center $q^\star$ attains the infimum. Now, suppose that the skill policy and the prior attains the maximal capacity $C = \mathcal{E}_{\mathrm{pot}}(s_0)$. For any distribution q, we have that

$$\mathbb{E}_{p(z|s_0)}\left[D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q(s_+)\big)\right] = I^\pi(Z; S_+ \mid S_0 = s_0) + D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0) \| q(s_+)\big)$$

following the Golden Formula (Polyanskiy and Wu, 2022, Thm. 4.1). Taking $q = q^\star$, every reachable occupancy has divergence at most C. Hence,

$$\begin{aligned} C &\geq \mathbb{E}_{p(z|s_0)}\left[D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big)\right] && (7) \\ &= C + D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0) \| q^\star(s_+)\big) && \text{(Golden formula)} \\ &\geq C. && (8) \end{aligned}$$

Thus, the KL between the marginal state occupancy and $q^\star$ is 0, which implies

$$q^\star = p^\pi(s_+ \mid s_0).$$

and that every individual skill divergence is at most C:

$$D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big) = C \tag{9}$$

for $p(z \mid s_0)$-almost every z. Thus, an optimal source places skill occupancy mass on the boundary of the optimal KL ball, with the KL measured from the central marginal distribution $q^\star$.

Finally, suppose there exists an active skill occupancy that is not an extreme point of $\mathcal{P}(s_0)$. Then, for distinct $p_1, p_2 \in \mathcal{P}(s_0)$ and some $\lambda \in (0,1)$,

$$p^\pi(s_+ \mid s_0, z) = \lambda p_1 + (1-\lambda) p_2.$$

Convexity of the KL divergence in the first argument gives

$$\begin{aligned} D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big) &< \lambda D_{\mathrm{KL}}\big(p_1(s_+) \| q^\star(s_+)\big) + (1-\lambda) D_{\mathrm{KL}}\big(p_2(s_+) \| q^\star(s_+)\big) \\ &\leq C, \end{aligned}$$

which contradicts the equality in Eq. (9). Thus, $p(z \mid s_0)$-almost every skill induces an extreme point of $\mathcal{P}(s_0)$. □

The reduction to the finite tabular case directly follows:

Theorem 4.1 (Potential empowerment is the KL radius of the reachable polytope, extending Eysenbach et al. (2021, Lem. 6.1–6.2))**.** *Let $\mathcal{P}(s_0) \subset \Delta(\mathcal{S})$ be the reachable state-occupancy polytope of Eq. (3). Jointly maximizing the mutual information over the skill source and the skill-conditioned policy solves the KL 1-center localization problem for $\mathcal{P}(s_0)$:*

$$\mathcal{E}_{\mathrm{pot}}(s_0) = \max_{\pi \in \Pi_{\mathcal{Z}},\, p(z|s_0)} I^\pi(Z; S_+ \mid S_0 = s_0) = \min_{q(s_+) \in \Delta(\mathcal{S})} \max_{p \in \mathcal{P}(s_0)} D_{\mathrm{KL}}\big(p(s_+) \| q(s_+)\big). \quad (5)$$

Moreover, every joint optimum $(\pi, p(z \mid s_0))$ places every skill on the boundary of the optimal KL ball measured from the optimal central occupancy $q^\star$:

$$D_{\mathrm{KL}}\big(p^\pi(s_+ \mid s_0, z) \| q^\star(s_+)\big) = \max_{p \in \mathcal{P}(s_0)} D_{\mathrm{KL}}\big(p(s_+) \| q^\star(s_+)\big) \quad \textit{for every skill } z. \quad (6)$$

In other words, the optimized skills span the maximal information radius of the reachable polytope given starting state s_0.

Proof of Theorem 4.1. In the finite-state setting, $\mathcal{P}(s_0)$ is a polytope and $\mathcal{E}_{\mathrm{pot}}(s_0) \leq \log |\mathcal{S}|$, so the channel capacity is finite. Theorem B.2 shows that the center distribution is the skill-marginal where active skill occupancies lie on the boundary of the optimal KL ball at extreme points of $\mathcal{P}(s_0)$. The extreme points of a polytope are vertices, which proves the result. □

B.1 Log reachability interpretation of potential empowerment in discrete deterministic settings

For completeness, we include the log reachability interpretation of the action sequence potential empowerment (as originally defined in (Klyubin et al., 2005a)) in discrete deterministic settings. Salge et al. (2014) originally present this result.

We begin by defining the action sequence potential empowerment. Let $A_0^{H-1} = (A_0, \ldots, A_{H-1}) \in \mathcal{A}^H$ be an open-loop action sequence over the fixed horizon H. Let S_t denote the random variable of the state after t timesteps. Klyubin et al. (2005a) define (potential) empowerment as the capacity of the channel from this action sequence to the resulting state:

$$\mathcal{E}_{\mathrm{pot}\,H}(s) = \max_{p(a_0^{H-1})} I(A_0^{H-1}; S_H \mid S_0 = s).$$

This expression is equivalent to Definition 3.1, replacing Z with the action sequence A_0^{H-1} and S_+ with S_H.

Lemma B.3 (Discrete deterministic potential empowerment is log reachability, adapted from Salge et al. (2014))**.** *Fix a discrete, deterministic MDP and horizon H. Because dynamics are deterministic, the future state S_H is a deterministic function of S_0 and A_0^{H-1}. Then, the reachable set of states in H timesteps takes the form*

$$\mathcal{R}_H(s) = \{s_+ \in \mathcal{S} : \exists\, a_0^{H-1} \in \mathcal{A}^H \textit{ that reaches } S_H = s_+ \textit{ from } S_0 = s\},$$

and the empowerment at s is the log of the size of this set:

$$\mathcal{E}_{\mathrm{pot}\,H}(s) = \log |\mathcal{R}_H(s)|.$$

Proof. Because S_H is a deterministic function of A_0^{H-1} given $S_0 = s$, the conditional entropy $H(S_H \mid A_0^{H-1}, S_0 = s)$ is 0. Then,

$$I(A_0^{H-1}; S_H \mid S_0 = s) = H(S_H \mid S_0 = s).$$

Because the support of S_H is in the reachable set $\mathcal{R}_H(s)$, the entropy is at most $\log |\mathcal{R}_H(s)|$. By construction, we can choose one action sequence for each reachable $s_+ \in \mathcal{R}_H(s)$ and place uniform mass $p(a_0^{H-1})$ over the chosen action sequences. Thus, the random variable S_H is uniform on $\mathcal{R}_H(s)$ with entropy $\log |\mathcal{R}_H(s)|$. □

C Empowerment and Bottleneck States

We state and prove the hitting-time and temporal-distance statements from §. 4.2.

C.1 Bottleneck states are high-empowerment states

We first define the repeated rollout hitting time from the DSOM. A single rollout from s_0 samples a skill $z \sim p(z \mid s_0)$ and a geometric termination time $\tau \sim \mathrm{Geom}(1-\gamma)$. The agent executes skill z until τ, where the rollout terminates at $S_+ \triangleq S_\tau$.

Writing $S_+^{(k)}$ for the termination state of the kth rollout from s_0, the repeated-rollout hitting time measures the first rollout count that hits s_+:

$$H^\pi(s_0 \to s_+) \triangleq \inf\left\{k \geq 1 : S_+^{(k)} = s_+\right\},$$

where $H^\pi(s_0 \to s_+ \mid z)$ is the corresponding count when every rollout only uses skill z. Then, potential empowerment is exactly the largest expected reduction in log hitting time after committing to a skill:

Lemma 4.2 (Bottleneck states are high potential empowerment states)**.** *In a discrete MDP, let the hitting time random variable $H^{\pi_{\text{eff}}}(s_0 \to s_+)$ be the number of rollouts from s_0 until a rollout hits s_+, for skills $Z \sim p(z \mid s_0)$. Let $H^{\pi_{\text{eff}}}(s_0 \to s_+ \mid z)$ be the corresponding hitting time when rolling out z every time. Then,*

$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi_{\text{eff}} \in \Pi_{\mathcal{Z}}} \mathbb{E}_{p^{\pi_{\text{eff}}}(z, s_+ \mid s_0)} \left[\log \mathbb{E}[H^{\pi_{\text{eff}}}(s_0 \to s_+)] - \log \mathbb{E}[H^{\pi_{\text{eff}}}(s_0 \to s_+ \mid z)]\right].$$

Proof. The repeated-rollout hitting times are geometric random variables with success probabilities $p^\pi(s_+ \mid s_0)$ and $p^\pi(s_+ \mid s_0, z)$, respectively. Hence,

$$\mathbb{E}[H^\pi(s_0 \to s_+)] = \frac{1}{p^\pi(s_+ \mid s_0)}, \quad \mathbb{E}[H^\pi(s_0 \to s_+ \mid z)] = \frac{1}{p^\pi(s_+ \mid s_0, z)}.$$

The MI is the expected log ratio of the skill-conditioned to the marginal DSOM, so

$$\begin{aligned} I^\pi(Z; S_+ \mid S_0 = s_0) &= \mathbb{E}_{p^\pi(z, s_+ \mid s_0)} \log \frac{p^\pi(s_+ \mid s_0, z)}{p^\pi(s_+ \mid s_0)} \\ &= \mathbb{E}_{p^\pi(z, s_+ \mid s_0)} \left[\log \mathbb{E}[H^\pi(s_0 \to s_+)] - \log \mathbb{E}[H^\pi(s_0 \to s_+ \mid z)]\right]. \end{aligned} \tag{10}$$

This identity holds for any policy. Maximizing both sides over effective empowerment policies π_{eff} gives the result. □

C.2 Temporal distances over contiguous skill rollouts

The above result considers the empowerment objective used throughout the rest of the paper, which measures the controllability of a single skill on the termination state/discounted future observation. We now define a closely related objective over *contiguous* rollouts where the temporal distance interpretation from Myers et al. (2024a) is exact.

Rather selecting a single skill and rolling out until termination, the agent can execute contiguous skill rollouts in a single trajectory. An agent can begin by selecting a skill z. At some geometric termination time $\tau \sim \mathrm{Geom}(1-\gamma)$, the agent resamples a random skill and continues forth from the terminated state. This process repeats forever. Intuitively, this rollout scheme is a generalization of standard MDP dynamics, where skill Z now takes the role of the action A, and the β discount factor over units of rollouts controls the fundamental timescale, in addition to the standard γ.

We can mathematically write out the state occupancy measure of such a process. Let the random variable X_k be the state occupied at the k-th geometric termination time. Let the trajectory begin at state $X_0 = s_0$. At the k-th termination, the agent re-draws a skill $Z_k \sim p(z \mid X_k)$ and continues the trajectory. Then, the state occupancy of the process takes the form

$$M_\beta^\pi(g \mid s) \triangleq (1-\beta) \sum_{k=0}^{\infty} \beta^k p^\pi(X_k = g \mid X_0 = s).$$

We can correspondingly define the state and skill-conditioned DSOM of the contiguous process as

$$M_\beta^\pi(g \mid s, z) \triangleq (1-\beta) \sum_{k=0}^{\infty} \beta^k p^\pi(X_k = g \mid X_0 = s, Z_0 = z),$$

where the agent rolls out skill z until the first geometric time, then resamples skills from the prior $p(z \mid s_k)$ thereafter. Note that averaging the skill-conditioned DSOM gives the state-conditioned DSOM:

$$\mathbb{E}_{z\sim p(z|s)}\left[M_\beta^\pi(g \mid s, z)\right] = M_\beta^\pi(g \mid s). \tag{11}$$

The **effective contiguous empowerment** measures the degree to which the first skill controls the occupancy of the entire process. Let $Z \sim p(z \mid s_0)$ be the first skill and let $G \sim M_\beta^\pi(g \mid s_0, Z)$ be the state the process occupies at a geometrically sampled termination time. Taking the supremum over the policy and source gives the channel capacity form, or the **potential contiguous empowerment**:

$$\mathcal{E}_{\text{eff}}^{\text{contig}}(\pi, s_0) \triangleq I^\pi(Z; G \mid X_0 = s_0),$$
$$\mathcal{E}_{\text{pot}}^{\text{contig}}(s_0) \triangleq \sup_{\pi\in\Pi_{\mathcal{Z}},\, p(z|s)} \mathcal{E}_{\text{eff}}^{\text{contig}}(\pi, s_0).$$

An agent is effectively contiguously empowered if their intial skill/generalized action exerts control on their future state for the contiguous process. An agent is potentially contiguously empowered when occupying a state from which immediate skills provide long-term control over the future. We highlight that this modification of empowerment is conceptually similar to the definitions of empowerment considered in the rest of the paper, despite different mathematical forms.

The potential contiguous empowerment admits a temporal distance interpretation, following Myers et al. (2024a). The temporal distance captures a notion of the time to traverse between states that generalizes to continuous state settings. Let β be the discount factor. Let $M_\beta^\pi(g \mid s)$ denote the DSOM of state g for a process that begins at s. Then, the temporal distance from a state s to a future state g is defined as

$$d_\beta^\pi(s, g) \triangleq -\log \frac{M_\beta^\pi(g \mid s)}{M_\beta^\pi(g \mid g)}$$

where the denominator measures the self-reachability probability for a process that begins at g. Conditioning on the initial skill gives the skill-conditioned temporal distance

$$d_\beta^\pi(s, g \mid z) \triangleq -\log \frac{M_\beta^\pi(g \mid s, z)}{M_\beta^\pi(g \mid g)},$$

again normalized by the self-reaching probability. Let $M_\beta^\pi(z, g \mid s_0) = p(z \mid s_0)\, M_\beta^\pi(g \mid s_0, z)$ be the joint law of the initial skill Z and future state G. Adding and subtracting the self-reaching term unifies the contiguous empowerment and the temporal distance:

Proposition C.1 (Contiguous empowerment measures temporal distance reduction)**.** *For any skill policy $\pi \in \Pi_{\mathcal{Z}}$ and skill source $p(z \mid s)$,*

$$\mathcal{E}_{\text{eff}}^{\text{contig}}(\pi, s_0) = \mathbb{E}_{M_\beta^\pi(z,g|s_0)}\left[d_\beta^\pi(s_0, g) - d_\beta^\pi(s_0, g \mid z)\right].$$

Consequently,

$$\mathcal{E}_{\text{pot}}^{\text{contig}}(s_0) = \sup_{\pi\in\Pi_{\mathcal{Z}},\, p(z|s)} \mathbb{E}_{M_\beta^\pi(z,g|s_0)}\left[d_\beta^\pi(s_0, g) - d_\beta^\pi(s_0, g \mid z)\right].$$

Thus, the potential contiguous empowerment is the maximum expected reduction in temporal distance to future states when committing to an initial skill Z.

Proof. The result follows directly from definitions. The self-reaching term cancels:

$$d_\beta^\pi(s_0, g) - d_\beta^\pi(s_0, g \mid z) = \log \frac{M_\beta^\pi(g \mid s_0, z)}{M_\beta^\pi(g \mid s_0)}.$$

Taking the expectation over $M_\beta^\pi(z, g \mid s_0)$ gives the equality between the temporal distance reduction and effective contiguous empowerment. Taking the supremum of both sides over the skill-conditioned policy and source gives the potential contiguous empowerment expression. □

Unlike the hitting time expression, this form of the contiguous potential empowerment holds in continuous settings. In discrete spaces, the temporal distance reduces to a typical hitting time interpretation (Myers et al., 2024a). Let

$$H_{\text{contig}}^\pi(s \to g) \triangleq \inf\{k \geq 0 : X_k = g\}$$

be the time to hit g from s, counted in units of contiguous skill rollouts. Let $H_{\text{contig}}^\pi(s \to g \mid z)$ be the corresponding skill-conditioned hitting time. Following Myers et al. (2024a), we can rewrite the temporal distance in terms of these hitting times:

$$d_\beta^\pi(s, g) = -\log \mathbb{E}\left[\beta^{H_{\text{contig}}^\pi(s\to g)}\right], \quad d_\beta^\pi(s, g \mid z) = -\log \mathbb{E}\left[\beta^{H_{\text{contig}}^\pi(s\to g|z)}\right] \quad \text{for } g \neq s.$$

Thus, there is a clear hitting time interpretation of the temporal distance object in the discrete limit.

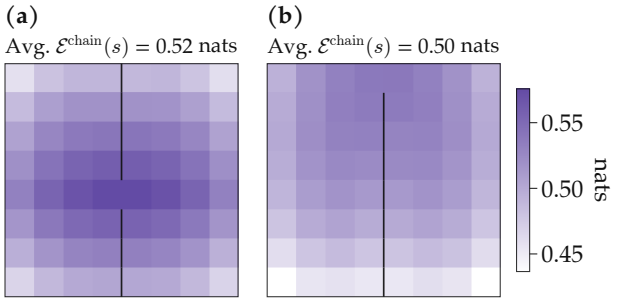


Figure 7: **Empowerment over contiguous rollouts concentrates at bottlenecks** (Proposition C.1)**.** The figures show the MI of the contiguous process at each start state on 5×5 grids with the door layouts of Fig. 4, with (a) a center opening and (b) a corner opening. The bottleneck openings have high potential contiguous empowerment.

Note on different empowerment expressions. The contiguous empowerment object differs from the empowerment objective used elsewhere in the paper. All other statements in the paper characterize the original potential empowerment; this result is an extension that provides a neat connection to temporal distance as defined in (Myers et al., 2024a), which (unlike hitting time) remains well-defined in continuous settings. Figure 7 shows the MI of the contiguous process from each start state for the Fig. 4 MDPs. The resulting potential contiguous empowerment field closely matches the potential empowerment field considered elsewhere in the paper in Fig. 4. In both experiments, the bottleneck states have high empowerment.

D Skill Selection: Fixed Continuous Uniform Priors are General

Lemma D.1 (A continuous uniform latent can represent any source distribution, discrete or continuous)**.** *Fix a start state s_0 and let the skill alphabet $\mathcal{Z}$ be standard Borel. Consider the mutual information $I^\pi(Z; S_+ \mid S_0 = s_0)$ under a skill-conditioned policy $\pi \in \Pi_{\mathcal{Z}}$, where $Z \sim p(z \mid s_0)$ is sampled once at s_0 and held fixed for the rollout. Let the reference source U be a continuous distribution on $\mathcal{Z}$, placing zero mass on every single point, e.g. the uniform distribution on $\mathcal{Z}$. Then*

$$\sup_{\pi \in \Pi_{\mathcal{Z}},\, p(z|s_0) \in \Delta(\mathcal{Z})} I^\pi(Z; S_+ \mid S_0 = s_0) = \sup_{\tilde{\pi} \in \Pi_{\mathcal{Z}},\, \tilde{Z} \sim U} I^{\tilde{\pi}}(\tilde{Z}; S_+ \mid S_0 = s_0),$$

where the maximization on the right ranges over all skill-conditioned policies.

Proof. Fix any pair (π, p) with $p \in \Delta(\mathcal{Z})$. By the transfer theorem (Theorem A.4; Kallenberg, 2021), there exists a measurable map $T_{s_0} : \mathcal{Z} \to \mathcal{Z}$ such that $T_{s_0}(\tilde{Z}) \sim p(z \mid s_0)$ for $\tilde{Z} \sim U$.[1] Set $Z = T_{s_0}(\tilde{Z})$. Define

$$\tilde{\pi}(a \mid s, \tilde{z}; s_0) = \pi(a \mid s, T_{s_0}(\tilde{z}); s_0).$$

Conditioned on $Z = T_{s_0}(\tilde{Z})$, the action kernels of π and $\tilde{\pi}$ agree at every timestep and the environment dynamics are identical. Therefore, the joint law of (Z, S_+) under (π, p) is the same as the joint law of $(T_{s_0}(\tilde{Z}), S_+)$ under $(\tilde{\pi}, U)$. Moreover, $\tilde{\pi}$ only receives information from $\tilde{z}$ through $T_{s_0}(\tilde{z})$, so S_+ is conditionally independent of $\tilde{Z}$ given $T_{s_0}(\tilde{Z})$, and

$$I^{\tilde{\pi}}(\tilde{Z}; S_+ \mid S_0 = s_0) = I^{\tilde{\pi}}\big(T_{s_0}(\tilde{Z}); S_+ \mid S_0 = s_0\big) = I^\pi(Z; S_+ \mid S_0 = s_0).$$

Taking the supremum over (π, p) shows that the the RHS upper bounds the LHS. Conversely, any policy $\tilde{\pi}$ with source $U \in \Delta(\mathcal{Z})$ is within the domain of maximization on the LHS, so we have an equality and are done. □

E Continuous versus Discrete Skill Spaces

Lemma E.1 (Uniform skill priors are not generally optimal with finitely many skills)**.** *With a finite skill alphabet, fixing the skill prior to be uniform can strictly reduce channel capacity.*

Proof. Consider an MDP consisting of a start state s_0 and two terminal states $\{s_1, s_2\}$. Let the skill set be the discrete skill set $\mathcal{Z} = \{0, 1\}$. Let the action space be a continuous action $a \in [0, 1]$. The transition is

$$p(s_1 \mid s_0, a) = \beta a, \qquad p(s_2 \mid s_0, a) = 1 - \beta a, \qquad 0 < \beta < 1.$$

[1] For $U = \text{Unif}[0, 1]$ this can be the inverse-CDF construction.

We can fully represent any deterministic, skill-conditioned policy by assigning each skill z to an action $a_z \in [0,1]$ that completely determines the transition probability $p(s_1 \mid s_0, z) = \beta a_z$.

First optimize the channel for a fixed source. By convexity of MI in the channel (Theorem A.1; Cover and Thomas, 2006; Polyanskiy and Wu, 2022), the maximum over $(a_0, a_1) \in [0,1]^2$ occurs at a vertex. The vertices $(0,0)$ and $(1,1)$ transmit no information. The informative vertices are $(0,1)$ and $(1,0)$, which are equivalent up to relabeling. Without loss of generality, set $(a_0, a_1) = (0,1)$, so that $p(s_1 \mid s_0, z=0) = 0$ and $p(s_1 \mid s_0, z=1) = \beta$.

Let $h(x) = -x \log x - (1-x)\log(1-x)$ denote the binary entropy. Under this channel, the output marginal is $p(s_1 \mid s_0) = p(Z=1 \mid s_0)\,\beta$, and the conditional output entropies are $h(0) = 0$ and $h(\beta)$. The MI is therefore

$$I(Z; S_+ \mid S_0 = s_0) = h\big(p(Z=1 \mid s_0)\,\beta\big) - p(Z=1 \mid s_0)\,h(\beta).$$

Differentiating in the source probability $p(Z=1 \mid s_0)$ gives

$$\frac{\partial\, I(Z; S_+ \mid S_0 = s_0)}{\partial\, p(Z=1 \mid s_0)} = \beta \log \frac{1 - p(Z=1 \mid s_0)\,\beta}{p(Z=1 \mid s_0)\,\beta} - h(\beta),$$

so the capacity-achieving source is

$$p^\star(Z=1 \mid s_0) = \frac{1}{\beta(1 + \exp(h(\beta)/\beta))}.$$

For $\beta = 1/2$, this gives $p^\star(Z=1 \mid s_0) = 2/5$, not $1/2$.

To see that the uniform prior is strictly worse, maximize over channels with $p(Z=1 \mid s_0) = 1/2$. The same convexity argument places the optimum at vertex $(a_0, a_1) = (0,1)$ or $(1,0)$. For $\beta = 1/2$, the best uniform-prior value is

$$h(1/4) - \frac{1}{2}h(1/2),$$

whereas the best arbitrary source value is

$$h(1/5) - \frac{2}{5}h(1/2).$$

The latter is strictly larger. Thus a uniform prior over two discrete skills cannot generally achieve the channel capacity. □

Comment on VIC (Gregor et al., 2017) and DIAYN (Eysenbach et al., 2018). A fixed continuous uniform prior is fully general for discrete and continuous MDPs, while a discrete prior is not. This result has implications for skill learning algorithms. VIC optimizes both a source distribution and channel (Gregor et al., 2017), with options (termed here as "skills") that may be finite or continuous. However, the authors note that the joint optimization is difficult and practically implement a uniform prior over a discrete skill set. DIAYN similarly fixes the source to a uniform categorical over a discrete skill set (Eysenbach et al., 2018). While neither work claims that fixing the prior preserves the objective, our result Lemma E.1 is a concrete counterexample that shows uniform discrete priors lose generality. Finally, Lemma D.1 shows that utilizing a *continuous* uniform skill prior recovers full generality *without the need for source learning*. Combined, our results theoretically suggest that skill learning methods should utilize continuous uniform sources; note that SOTA skill learning methods already utilize such priors (Zheng et al., 2025; Park et al., 2023).

F Empowerment Lower Bounds Tabular Adaptation

We now prove that empowerment lower bounds zero-shot adaptation in the tabular setting under an uninformed reward prior.

Theorem 4.3 (Informal statement that tabular empowerment lower bounds reward adaptation)**.** *Fix a start state s_0 and sample a reward $r \sim \mathcal{R}$ from the uninformed Gaussian prior. The agent adapts by selecting the skill with the highest mean return under r, and the adaptation $J(s_0)$ averages the best return over the reward prior. In whitened coordinates, write g for the reward and u_z for the offset of skill z's occupancy from the skill-marginal DSOM of Eq. (2), so the return gap of skill z over the marginal is $\langle g, u_z \rangle$. Then, for any skill space and skill prior,*

$$J(s_0) = \mathbb{E}_g \sup_{z \in \mathcal{Z}} \big\langle g,\ u_z \big\rangle \ \gtrsim\ I(Z; S_+ \mid S_0 = s_0).$$

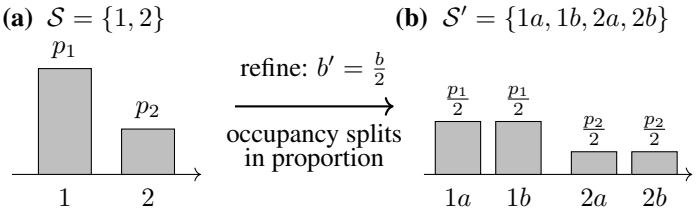


Figure 8: **State refinement leaves return distribution unchanged (Example F.1).** Bars display the occupancy $p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) = (p_1, p_2)$ of a skill over the coarse states (a) and the occupancy over the split states (b). Each panel lists the reference mass and per-state reward prior $r(s_+) \sim \mathcal{N}(0, 1/b(s_+))$. Splitting states halves the reference mass and doubles the reward variance. Thus, the return obeys the same distribution $\mathcal{N}(0,\ 2p_1^2 + 2p_2^2)$ in both MDPs.

Define the skill offset $u_z \triangleq \big(p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) - p^{\pi_{\text{eff}}}(s_+ \mid s_0)\big)/\sqrt{b}$. The whitened reward $g \triangleq \sqrt{b}\, r$ is standard normal under the prior, and the marginal's return $\langle g,\ p^{\pi_{\text{eff}}}(s_+ \mid s_0)/\sqrt{b}\rangle$ is mean-zero and z-independent, so

$$J(s_0) = \mathbb{E}_g \sup_{z \in \mathcal{Z}} \langle g,\ u_z \rangle, \tag{12}$$

is the Gaussian width of $\{u_z\}_{z \in \mathcal{Z}}$.

Before presenting the proof, we discuss the chosen reward prior and motivate the variance scaling. Under the $1/b$ scaling, splitting a state into substates leaves the distribution of the returns unchanged.

Example F.1 (Splitting a two-state MDP into a four-state MDP Fig. 8)**.** Let $\mathcal{S} = \{1, 2\}$ with uniform reference $b(1) = b(2) = \frac{1}{2}$. The prior samples $r(1), r(2) \sim \mathcal{N}(0, 2)$ independently (Fig. 8). For a skill z with DSOM $p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) = (p_1, p_2)$, the mean reward is a centered Gaussian:

$$\mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0, z)}[r(s_+)] = p_1\, r(1) + p_2\, r(2) \ \sim\ \mathcal{N}\big(0,\ 2p_1^2 + 2p_2^2\big)\,.$$

Splitting each state into two substates gives $\mathcal{S}' = \{1a, 1b, 2a, 2b\}$ where the refined prior samples each coordinate reward $r' \sim \mathcal{N}(0, 4)$. Furthermore, split every skill's occupancy in the same proportions, so the occupancy of z becomes $\big(\frac{p_1}{2}, \frac{p_1}{2}, \frac{p_2}{2}, \frac{p_2}{2}\big)$. The mean reward on the fine-grained space is

$$\tfrac{p_1}{2}\big(r'(1a) + r'(1b)\big) + \tfrac{p_2}{2}\big(r'(2a) + r'(2b)\big) \ \sim\ \mathcal{N}\Big(0,\ \tfrac{p_1^2}{4} \cdot 8 + \tfrac{p_2^2}{4} \cdot 8\Big) = \mathcal{N}\big(0,\ 2p_1^2 + 2p_2^2\big)\,,$$

since each sum $r'(\cdot a) + r'(\cdot b)$ has variance 8. Thus, the distribution of the mean reward is the same before and after splitting the state.

In this example, the $1/b$ variance scaling is key. Under a unit-variance prior $r(s_+) \sim \mathcal{N}(0, 1)$, the same state splitting shrinks the return variance from $p_1^2 + p_2^2$ to $\frac{1}{2}\big(p_1^2 + p_2^2\big)$. Thus, the constraint that the return distribution is scale invariant implies the above scaling of the reward prior variance.

Now that we have built intuition for the choice of reward prior, we present the proof:

Proof. By Eq. (12), it suffices to lower bound the Gaussian width. The proof consists of two inequalities:

$$J(s_0) \ \overset{\text{step 1}}{\geq}\ \frac{1}{\sqrt{2\pi}} \Big(\mathbb{E}_{p(z|s_0)} \|u_z\|^2\Big)^{1/2} \ \overset{\text{step 2}}{\geq}\ \frac{I(Z; S_+ \mid S_0 = s_0)}{c\sqrt{2\pi}}. \tag{13}$$

Fix the reward r and a skill $z \in \mathcal{Z}$. For the first inequality (step 1), we observe that the best average return obtainable by a skill dominates the marginal skill return:

$$\sup_{z' \in \mathcal{Z}} \mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0, z')}[r(s_+)] \ \geq\ \max\big(\mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0, z)}[r(s_+)],\ \mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0)}[r(s_+)]\big)\,.$$

Take the expectation over rewards on both sides using the identity $\max(x, y) = \frac{1}{2}(x + y) + \frac{1}{2}|x - y|$. Both returns are linear in r, so the $\frac{1}{2}(x + y)$ term vanishes and

$$J(s_0) \ \geq\ \tfrac{1}{2}\, \mathbb{E}_r \big|\mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0, z)}[r(s_+)] - \mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0)}[r(s_+)]\big|\,.$$

The gap in returns is an inner product of the whitened quantities:

$$\mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0, z)}[r(s_+)] - \mathbb{E}_{p^{\pi_{\text{eff}}}(s_+ | s_0)}[r(s_+)] = \langle g,\ u_z \rangle\,.$$

Because the coordinates of g are i.i.d. standard Gaussians, we know that the gap in returns is a Gaussian with variance $\|u_z\|^2$. Hence, for every $z \in \mathcal{Z}$,

$$\begin{aligned} J(s_0) &\geq \tfrac{1}{2}\,\mathbb{E}_r\,|\langle g,\, u_z\rangle| \\ &= \frac{1}{\sqrt{2\pi}}\,\|u_z\|\,. \end{aligned}$$

Averaging the square of this bound over $z \sim p(z \mid s_0)$ (LHS is independent of the skill z) gives step 1 of Eq. (13). For step 2, express the MI as the expected KL divergence between the skill prior and the skill posterior after observing s_+:

$$I(Z; S_+ \mid S_0 = s_0) = \sum_{s_+} p^{\pi_{\text{eff}}}(s_+ \mid s_0)\, D_{\text{KL}}\big(p^{\pi_{\text{eff}}}(z \mid s_0, s_+) \,\big\|\, p(z \mid s_0)\big).$$

Applying Bayes' rule, the χ^2 divergence takes the form of a mean squared deviation of the skill DSOMs from the marginal:

$$\begin{aligned} \chi^2\big(p^{\pi_{\text{eff}}}(z \mid s_0, s_+) \,\big\|\, p(z \mid s_0)\big) &= \mathbb{E}_{p(z|s_0)}\left[\left(\frac{p^{\pi_{\text{eff}}}(s_+ \mid s_0, z)}{p^{\pi_{\text{eff}}}(s_+ \mid s_0)} - 1\right)^2\right] \\ &= \frac{\mathbb{E}_{p(z|s_0)}\big[\big(p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) - p^{\pi_{\text{eff}}}(s_+ \mid s_0)\big)^2\big]}{p^{\pi_{\text{eff}}}(s_+ \mid s_0)^2}. \end{aligned}$$

Then, through χ^2, we can upper bound the MI with the norm of the occupancy gap vector:

$$\begin{aligned} I(Z; S_+ \mid S_0 = s_0) &\leq c \sum_{s_+} p^{\pi_{\text{eff}}}(s_+ \mid s_0) \sqrt{\chi^2\big(p^{\pi_{\text{eff}}}(z \mid s_0, s_+) \,\big\|\, p(z \mid s_0)\big)} \qquad \text{(Theorem A.3)} \\ &= c \sum_{s_+} \sqrt{\mathbb{E}_{p(z|s_0)}\big[\big(p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) - p^{\pi_{\text{eff}}}(s_+ \mid s_0)\big)^2\big]} \\ &= c \sum_{s_+} \sqrt{b(s_+)} \sqrt{\frac{\mathbb{E}_{p(z|s_0)}\big[\big(p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) - p^{\pi_{\text{eff}}}(s_+ \mid s_0)\big)^2\big]}{b(s_+)}} \\ &\leq c \left(\sum_{s_+} \frac{\mathbb{E}_{p(z|s_0)}\big[\big(p^{\pi_{\text{eff}}}(s_+ \mid s_0, z) - p^{\pi_{\text{eff}}}(s_+ \mid s_0)\big)^2\big]}{b(s_+)}\right)^{1/2} \\ &\qquad \text{(Cauchy-Schwarz)} \\ &= c \left(\mathbb{E}_{p(z|s_0)} \|u_z\|^2\right)^{1/2}. \end{aligned}$$

The first line applies Theorem A.3 with the bound $\ln(1+x) \leq c\sqrt{x}$ to each divergence. The second line expands χ^2 and cancels the marginal weight. The third and fourth lines apply Cauchy–Schwarz. The last equality is the definition of u_z, proving step 2 of Eq. (13) which finishes the proof. □

We provide some didactic tabular experiments that demonstrate the behavior and scaling of the bound over different MDPs. In Fig. 9, we demonstrate the relationship between the degree of commitment within an MDP – the controllability of the MDP – and the bound for a fixed set of states. As controllability of the MDP increases, the bound tightens. In Fig. 10, we demonstrate the relationship between the number of blocks in an MDP (e.g. number of relative states) and the bound under fixed commitment. As the number of blocks grows, the bound loosens.

G When Information Geometry is *Insufficient* for Adaptation

The adaptation result above is in the tabular setting. For the following results, we also consider MDPs whose state space $\mathcal{S}$ is a metric space with arbitrary metric d. Importantly, these state spaces may be continuous.

Proposition G.1 (MDPs can admit arbitrarily many distinct skills)**.** *Fix a discount $\gamma \in (0, 1)$ and a skill space $\mathcal{Z} = [K]$ with the uniform prior $Z \sim \text{Unif}([K])$. There exists an MDP with a start state $s \in \mathcal{S}$ and K skills that attains the potential empowerment*

$$\mathcal{E}_{\text{pot}}(s) = \max_{\pi_{\text{eff}}} I^{\pi_{\text{eff}}}(Z; S_+ \mid S_0 = s) = \gamma \log K.$$

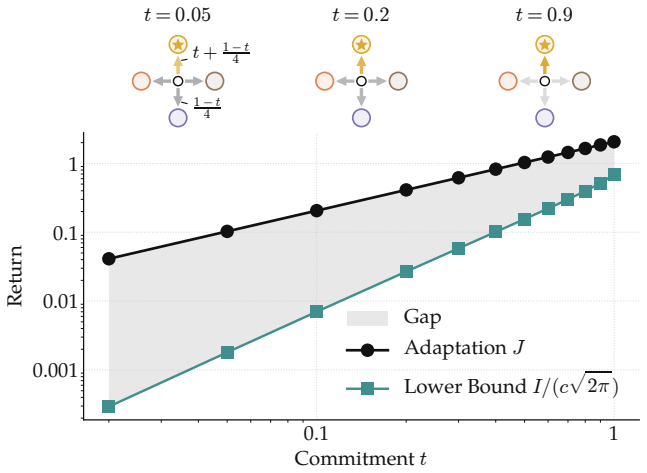


Figure 9: **MI bound as a function of skill commitment** (Theorem 4.3)**.** Consider the MDP consisting of 5 states with a central starting state, where any left/right/up/down action sends probability mass to absorbing left/right/up/down states. Actions left/right/up/down have a level of commitment t, which measures the additional mass above uniform assigned to the corresponding state. When commitment is small, probability is almost spread equally over all four states regardless of action so there is little control. In this regime, the adaptation J dominates the lower bound. As the commitment increases, actions/skills commit to states and the bound tightens.

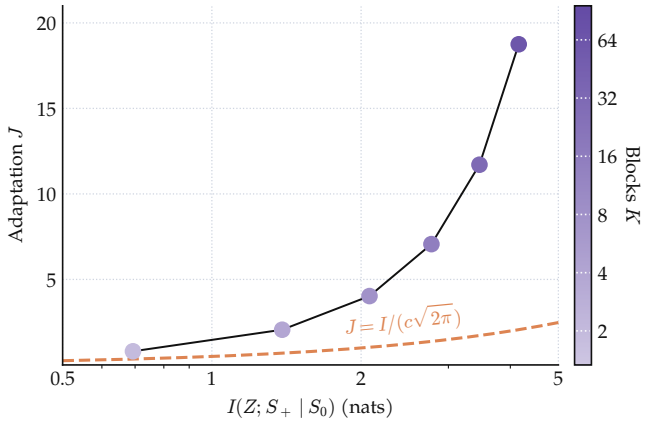


Figure 10: **MI bound as a function of distinct state regions** (Theorem 4.3)**.** Consider the MDP consisting of 128 absorbing states split into K distinct blocks. Each of the K actions sends probability uniformly over the Kth block. Thus, K counts the maximum possible number of total skills. The scaled MI (orange) correctly lower bounds the adaptation J (purple). The bound weakens as the number of blocks in the MDP grows.

If $\mathcal{S}$ is continuous, there exists an MDP where a continuum of these skills gives unbounded potential empowerment $\mathcal{E}_{\text{pot}}(s) = +\infty$.

Proof. Discrete case. We construct MDPs that admit $\mathcal{E}_{\text{pot}}(s) \geq \gamma \log K$ for all K.

Consider an MDP that admits a large number of distinguishable occupancies. Then, we can assign each skill a distinct occupancy at some target future state. In this limit, the skill becomes readily identifiable from the future state.

For finite K, consider a discretized state space $\mathcal{S}$ with states $s, s_1, \ldots, s_K \in \mathcal{S}$ where all states $s_1, \ldots, s_K$ are absorbing. Let skill z immediately transition from start state s to another target state. Then, the induced DSOM is

$$p^{\pi_{\text{eff}}}(s_+ \mid s, z) = (1-\gamma)\,\delta_s + \gamma\,\delta_{s_z},$$

where δ is the Kronecker delta. Here, the skills share the start state s with probability mass $1-\gamma$ and are otherwise disjoint over $\mathcal{S} \setminus \{s\}$. Observing $S_+ \neq s$ exactly determines the skill while observing $S_+ = s$ gives no information. Thus, the MI is

$$I^{\pi_{\text{eff}}}(Z; S_+ \mid S_0 = s) = \gamma\, H(Z) = \gamma \log K,$$

which obtains the potential empowerment.

Continuous case. We similarly construct MDPs that admit unbounded $\mathcal{E}_{\text{pot}}(s)$. When $\mathcal{S}$ is continuous, we can index skills by a continuous parameter rather than a finite set and map $z \in [0,1]$ to $S \setminus \{s\}$. The induced DSOM is

$$p(s_+ \mid s, z) = (1-\gamma)\,\delta_s + \gamma\,\delta_{\phi(z)}$$

where δ is now the Dirac delta. As before, any final state that is not the start state identifies the skill. If we cut $[0,1]$ into n equal segments, the future state S_+ reveals at least $\gamma \log n$ bits about the skill. Since this construction holds for all n, we have that

$$I(Z; S_+ \mid S_0 = s) \geq \sup_n \gamma \log n = +\infty.$$

Thus, we have constructed an MDP where the MI is unbounded for every $\gamma \in (0,1)$ for continuous state spaces. □

We can utilize this proposition to prove that potentially infinite bits can fit within a tiny metric ball:

Proposition G.2 (Many bits can fit in a tiny metric ball)**.** *Fix a start state s and a ball $B(\bar{s}, \epsilon)$ of radius ϵ with infinitely many states, centered at an arbitrary state $\bar{s} \in \mathcal{S}$. For every K, there exists an MDP with K skills as in Proposition G.1 whose target states s_z all lie inside the ball, so*

$$I(Z; S_+ \mid S_0 = s) = \gamma \log K,$$

yet every bounded 1-Lipschitz reward r gives all skills nearly the same return:

$$\sup_{z,z'} \left| \mathbb{E}_{p(s_+|s,z)}[r(S_+)] - \mathbb{E}_{p(s_+|s,z')}[r(S_+)] \right| \leq 2\gamma\epsilon \leq 2\epsilon.$$

With a continuum of skills inside the ball, the MI can be unbounded while the upper bound on the return difference between skills is unchanged.

Proof. We can apply Proposition G.1 to the ball $B(\bar{s}, \epsilon)$ and construct valid MDPs with the above MIs. For any 1-Lipschitz reward, the skill returns of $z \neq z'$ in these MDPs differ by

$$\begin{aligned} \left| \mathbb{E}_{p(s_+|s,z)}[r(S_+)] - \mathbb{E}_{p(s_+|s,z')}[r(S_+)] \right| &= \gamma\, |r(s_z) - r(s_{z'})| \\ &\leq \gamma \left(d(s_z, \bar{s}) + d(\bar{s}, s_{z'}) \right) \\ &\leq 2\gamma\epsilon \end{aligned}$$

which proves the result. □

For any 1-Lipschitz reward, it is always possible for a set of skills to have high MI (potentially unbounded) while providing no adaptation advantage for continuous rewards beyond a single policy. Thus, we should *not* extrapolate the tabular adaptation bound (Theorem 4.3) to continuous spaces with smooth rewards.

We show an example of this collapse in Fig. 6. The reachable future state occupancies pack closer together in continuous space from left to right. In the limit, the difference in values between skills shrinks to zero while the MI diverges, so the discrete adaptation bound becomes trivial. We note that prior works explicitly point out and propose solutions to the problem of diverging channel capacities in continuous spaces (Tiomkin et al., 2017; Salge et al., 2013), though the works do not formally connect empowerment and adaptation.

H Anisotropic Reward Priors

This section relaxes the reward isotropy assumption in the adaptation result Theorem 4.3. Unfortunately, using an anisotropic reward prior with covariance Σ scales the lower bound by $\sqrt{\lambda_{\min}(\Sigma)}$ (Proposition H.1). In the limit, skills with arbitrarily large mutual information can exhibit no adaptation even with fixed reward power constraints (e.g. fixed $\operatorname{Tr}\Sigma$; see Proposition H.2).

We keep the discrete setting of Theorem 4.3 and only change the reward prior. We sample rewards with $\sqrt{b}\,r \sim \mathcal{N}(0,\Sigma)$, so the law of the uniform baseline-whitened reward g in Eq. (12) becomes $\mathcal{N}(0,\Sigma)$. An arbitrary anisotropic prior may concentrate variance on directions that the empowerment-maximizing skills choose to ignore. The marginal's return remains mean zero for any Σ, so the adaptation objective takes the form

$$J^{\Sigma}(s_0) \triangleq \mathbb{E}_{g\sim\mathcal{N}(0,\Sigma)}\left[\sup_{z\in\mathcal{Z}}\langle g,\ u_z\rangle\right].$$

Proposition H.1 (Anisotropic reward prior scales bound)**.** *In the setting of Theorem 4.3, sample rewards with $\sqrt{b}\,r \sim \mathcal{N}(0,\Sigma)$ instead. Then, the MI bound rescales by $\sqrt{\lambda_{\min}(\Sigma)}$:*

$$J^{\Sigma}(s_0) \ \geq\ \sqrt{\lambda_{\min}(\Sigma)}\ \frac{I(Z;S_+ \mid S_0=s_0)}{c\sqrt{2\pi}}.$$

The isotropic prior $\Sigma = \mathrm{Id}$ recovers Theorem 4.3 exactly.

Proof. The proof of Theorem 4.3 uses the reward prior to express a lower bound on $J(s_0)$ in terms of the norm of the occupancy gap vector. Under the $\mathcal{N}(0,\Sigma)$ prior, the occupancy gap vector variance scales by the anisotropic covariance. Then, the result from Theorem 4.3 suggests the following modified bound:

$$J^{\Sigma}(s_0) \overset{\text{step 1}}{\geq} \frac{1}{\sqrt{2\pi}}\left(\mathbb{E}_{p(z|s_0)}\left\|\Sigma^{1/2}u_z\right\|^2\right)^{1/2} \overset{\text{step 2}}{\geq} \sqrt{\lambda_{\min}(\Sigma)}\ \frac{I(Z;S_+ \mid S_0=s_0)}{c\sqrt{2\pi}}. \tag{14}$$

Now, we carefully derive this bound. The reduction of J to the gap in returns in Theorem 4.3 only depends on a centered reward. Thus, as before, we have that

$$J^{\Sigma}(s_0) \ \geq\ \tfrac{1}{2}\,\mathbb{E}_g\left|\langle g,\ u_z\rangle\right|.$$

For $g \sim \mathcal{N}(0,\Sigma)$, the return gap $\langle g,\ u_z\rangle$ is Gaussian with variance $\left\|\Sigma^{1/2}u_z\right\|^2$. Hence, for every $z \in \mathcal{Z}$,

$$J^{\Sigma}(s_0) \ \geq\ \frac{1}{\sqrt{2\pi}}\left\|\Sigma^{1/2}u_z\right\|.$$

Averaging the square of this bound over $z \sim p(z \mid s_0)$ (the left side does not depend on z) gives step 1. We can lower bound the norm:

$$\|\Sigma^{1/2}u_z\|^2 = \langle u_z, \Sigma u_z\rangle \ \geq\ \lambda_{\min}(\Sigma)\,\|u_z\|^2.$$

Combining with the step 2 bound (does not change with the reward prior) gives the desired result. □

Indeed, we can construct an example where a state and set of policies with high potential empowerment exhibit zero adapation to anisotropic rewards:

Proposition H.2 (Large MI with zero anisotropic adaptation)**.** *Fix a power budget $T > 0$. For every K, there exists a discrete MDP with skills achieving $I(Z;S_+ \mid S_0=s_0) = \gamma\log K$ and a reward prior $\sqrt{b}\,r \sim \mathcal{N}(0,\Sigma)$ with $\operatorname{Tr}\Sigma = T$ such that $J^{\Sigma}(s_0) = 0$.*

Proof. The discrete construction in Proposition G.1 attains $I(Z;S_+ \mid S_0=s_0) = \gamma\log K$ with K occupancy vector offsets u_z that span a subspace of dimension at most $K-1$ of the $(K+1)$-dimensional whitened reward space.

Choose a unit vector v orthogonal to every u_z and set the reward variance $\Sigma = T\,vv^{\top}$. This choice places all variance of the reward prior in the space orthogonal to the span of the occupancy gap vectors. By the proof of Proposition H.1, each return gap $\langle g, u_z\rangle$ is Gaussian with variance $\|\Sigma^{1/2}u_z\|^2 = 0$, so $J^{\Sigma}(s_0) = 0$ while the MI is unchanged. □

Thus, while the MI alone can provide nontrivial lower bounds on adaptation in tabular settings for an isotropic reward prior, the same no longer holds true for anisotropic reward priors. An empowerment-maximizing agent could learn skills with high control over observations and ignore other learnable skills that control the rewarding degrees of freedom.

I Strengthening the Adaptation Bound for General Priors

J Tabular Experiment Details

J.1 Tabular 5×5 grid experiments

Table 1 details the 5×5 grid MDP parameters used for Fig. 2.

Skill-learning via convex optimization. In the state centrality experiments, we used a convex solver to enumerate vertices of the reachable discounted state occupancy measure (DSOM) polytope for all possible starting states. An MDP with $|\mathcal{S}|$ states and $|\mathcal{A}|$ actions can have up to $|\mathcal{A}|^{|\mathcal{S}|}$ vertices. Thus, enumerating all vertices is challenging (Eysenbach et al., 2021).

However, we can use the fact that reward-maximizing DSOM always sit at the vertices of the convex polytope to conduct vertex search. Specifically, we (1) sample a random reward vector $\theta \in \Delta(\mathcal{S})$ in the state simplex and (2) use a convex solver to maximize DSOM alignment with the reward vector subject to probability flow constraints (Puterman, 2014). Since a linear objective over the reachable state polytope is maximized at an extreme point, and a generic random objective exposes a vertex almost surely, this procedure retrieves valid vertices of the polytope conditioned on specific starting states. This procedure generally does not retrieve the complete set of candidate vertices over which to specify skills, but it provides a good approximation.

Policies that optimize

$$\max_{\pi(a|s,z;s_0),\, p(z|s_0)} I(Z; S_+ \mid S_0 = s_0),$$

seek out the vertices of the polytope (see Theorem 4.1). Thus, we fix the approximate feasible set of optimal policies to the set of found vertices and break symmetries among policies with equivalent vertex DSOMs by selecting the policy with minimal entropy. This procedure constructs a one-to-one mapping between policies and vertex DSOMs and fixes the encoder in the information channel. Then, getting the empowerment-maximizing policy with known dynamics is simple: it is well known that for a fixed channel encoder $p^\pi(s_+ \mid s_0, z)$, the MI is concave in the source distribution $p(z \mid s_0)$ and convex in the channel (Theorem A.1; Cover and Thomas, 2006). Thus, finding the optimal source distribution $p(z \mid s_0)$ is reduced to a convex optimization problem solvable with standard convex optimization packages (see the Blahut–Arimoto algorithm; Arimoto, 1972; Blahut, 1972).

An alternative to sampling random reward vectors to search for policies is to utilize an iterative form of the B-A algorithm that alternates source optimization with a more principled vertex search. We additionally implement a version of Huang and Meyn (2005) that uses cutting plane linearization to find occupancy vertices that increase the MI. While this method is more involved, the takeaways of the tabular experiments are the same whether utilizing the random reward vertex search or a more principled method for capacity maximization and vertex search.

The results in Fig. 2 show that the latent z-conditioned policies that start at $s_0 = (2, 2)$ cover all states within the MDP.

Learning empowerment-maximizing policies. The above procedure optimizes the channel capacity over the sampled vertex set, yielding a lower bound on

$$\mathcal{E}_{\text{pot}}(s_0) = \max_{\pi(a|s,z;s_0),\, p(z|s_0)} I(Z; S_+ \mid S_0 = s_0)$$

that tightens as the candidate vertex set grows. We use standard Q-learning methods to obtain the greedy policy that maximizes

$$Q^{\pi_{\text{pot}}}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{E}_{\text{pot}}(s_t) \middle| s_0 = s, a_0 = a\right],$$

which we refer to as the potential empowerment Q-value. We anneal the epsilon greedy action rate from $1.0 \to 0.01$ from the start to the end of Q-learning. The optimal policy shown in Fig. 2(c) moves to the center of the room, maximizing reachability to all other states in the MDP.

J.2 Bottleneck grid experiments

Figure 4 uses 8×8 grids with center and corner doors, $\gamma \in \{0.5, 0.95\}$, and no random actions. We use Blahut–Arimoto source optimization with multi-start policy search; solver settings are fixed across discounts.

This MDP demonstrates the bottleneck effect discussed in §. 4.2. In the open grid, the center state has maximal potential empowerment. With the added walls, the bottleneck openings at the gaps in the walls have high empowerment. At these states, the agent's choice in skill substantially changes its temporal distance to future states.

Table 1: Hyperparameters for the 5×5 gridworld experiments (Fig. 2)

Parameters	**Value**
Environment size (n $\times$ m)	5×5
Environment actions	{Left, Right, Up, Down, Stay}
Periodic boundary conditions	No
P(random action)	0.0
Gamma for DSOM	0.9
# uniformly sampled vertex directions per s_0	500
Convex solver	Clarabel (Goulart and Chen, 2024)
# max iterations	`50_000`
Error tolerance	$1e{-6}$
Vertex seed	0
Epsilon schedule	Linear interpolation from $1 \to 0.01$
Gamma for potential policy	0.9
# Q-learning iterations	2×10^5

J.3 KEY-FINDING EXPERIMENTS

Table 2 gives the key-finding experiment hyperparameters.

We use a numerical experimental setup identical to that of the 5×5 room with walls. In this MDP, the agent can pick up and drop off a key at the square $(0, 2)$. Picking up this key allows access to the blue-shaded regions of the MDP; without the key, the agent cannot enter these spaces in the grid. Fig. 11 shows the resulting empowerment-maximizing policy that ascends the empowerment landscape. Starting at state $s_0 = (4, 2)$ without a key, the agent learns to navigate to the key, pick it up, and navigate to the center square $(2, 2)$. This experiment shows that states with maximal empowerment capture an information-geometric notion of centrality—rather than simply navigating to and staying at the *physically* center state, as in the 5×5 grid, the agent recognizes that the most central state is the one where it has obtained the key.

Table 2: Hyperparameters for the 5×5 gridworld with walls and a key

Parameters	**Value**
Environment size (n $\times$ m)	5×5
Environment actions	{Left, Right, Up, Down, Stay, Pick up key at $(0, 2)$, Drop key at $(0, 2)$}
Periodic boundary conditions	No
P(random action)	0.0
Gamma for skill-learning	0.99
# uniformly sampled vertex directions per s_0	1000
Convex solver	Clarabel (Goulart and Chen, 2024)
# max iterations	`50_000`
Error tolerance	10^{-6}
Vertex seed	0
Epsilon schedule	Linear interpolation from $1 \to 0.01$
Gamma for potential policy	0.99
# Q-learning iterations	2×10^5

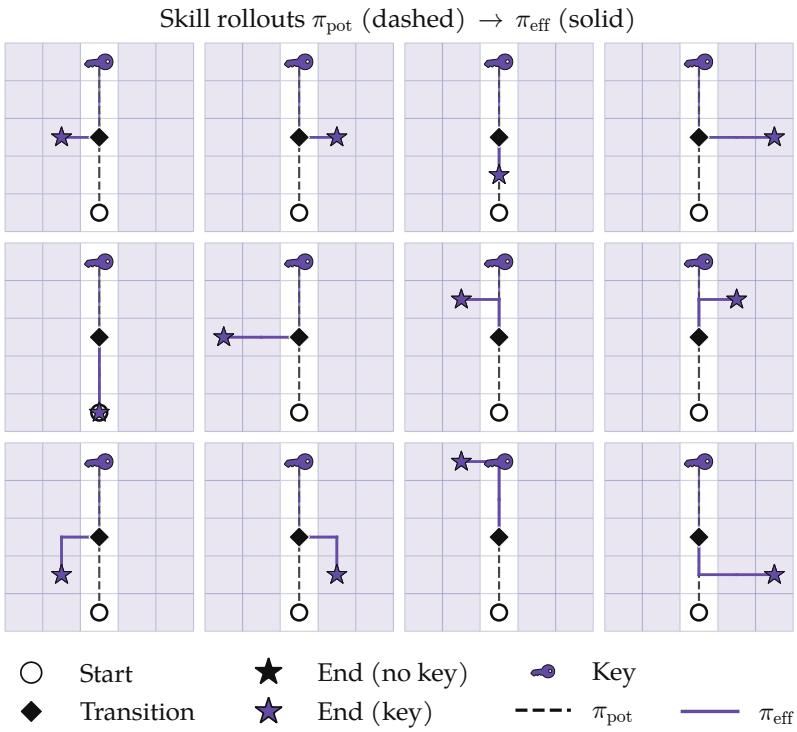


Figure 11: **High empowerment finds central states, where skills are diverse from the central states** (Theorem 4.1)**.** Potential policy navigates to central states. Learned skills spread out from the central state and measure the degree of centrality of the state (in information geometry). The dotted line denotes the potential empowerment policy's trajectory for $\frac{1}{1-\gamma}$ steps. Diamond denotes the handoff state s_0, where the agent randomly samples $z \sim p(z \mid s_0)$ and rolls out skill policy $\pi(a \mid s, z; s_0)$.

J.4 Adaptation geometry examples

In the example MDP of Fig. 5, every skill transitions in one step from the starting state to an absorbing target state. Thus, a skill's DSOM, normalized over the targets, is the transition distribution. Table 3 lists the resulting occupancy distributions.

In Example X (Fig. 5(a)), there are $N_T = 2$ targets, and each of two skills transitions to its target with probability one. In Example Y (Fig. 5(d)), all three skills reach $N_T = 3$ targets and no others. Each skill transitions to its target with probability $\frac{23}{30}$ and to each of the other two targets with probability $\frac{7}{60}$.

Figure 5(b, e) plot the polytope of the whitened state occupancy offsets u_z in Eq. (12) with the uniform baseline $b = 1/N_T$ over the targets.

The overlap between Example Y's skills leads to low channel capacity. However, the spread over the target states gives a larger Gaussian width. The orderings of the adaptation and the obtained MI therefore flip: X attains the higher mutual information and Y the larger width.

In the Gaussian width panels (Fig. 5(b, e)), the offsets in Eq. (12) span at most two dimensions, and the axes (u_1, u_2) give the coordinates of u in an orthonormal basis of this span.

At most two offsets are linearly independent in each example in Fig. 5. The occupancy offsets in the top row are negatives of each other under the uniform skill prior and the bottom row's three offsets satisfy $u_1 + u_2 + u_3 = 0$. Thus, we can draw a 2D diagram of these geometries (Fig. 5(b, c, e, f)); the insets in Fig. 5(f) visualize the occupancy distributions. Example X's antipodal pair spans only one degree of freedom, so Fig. 5(b) shows a segment. In addition to the possible occupancies, the coordinates in this space can also denote reward vectors, and the background color indicates the skill that adapts to that reward. In the information panels, each colored endpoint is a possible skill's DSOM, at the vertices of a reachable polytope in Fig. 5. The skill DSOMs and their mixture are shown explicitly in the inset bar charts. Arrows point from the capacity-achieving mixture $q^\star = b$ (black dot) to every possible skill's DSOM, with equal positive source weights. The solid black contour in Fig. 5(f) traces the KL level set at the capacity. In Fig. 5(c), this level set is just two endpoints.

J.5 Worked example of centrality in information geometry

We illustrate the geometry of empowerment in Fig. 3a. To start, consider learning skills that start at state s_2. From this state, there are three actions which deterministically transition to s_3, s_4, or s_5. We consider learning three skills, so $z \in \{1, 2, 3\}$. The problem of optimizing mutual information at state s_2 corresponds to learning the policy $\pi(a \mid s = s_2, z)$ so that the skill index z conveys as many bits as possible about the next state. A poor choice for this policy would be the random policy, wherein each skill selects randomly among the three actions. Each of these three skills induces the same distribution over future states s_3, s_4, s_5, conveying zero bits of information. We visualize the induced state occupancies in Fig. 3a. Conversely, if we assign each skill a unique action (e.g., $z = 1$ always takes the leftmost action), then the state distributions the skills induce become markedly different. As visualized in Fig. 3a, this choice of skills covers a much larger region.

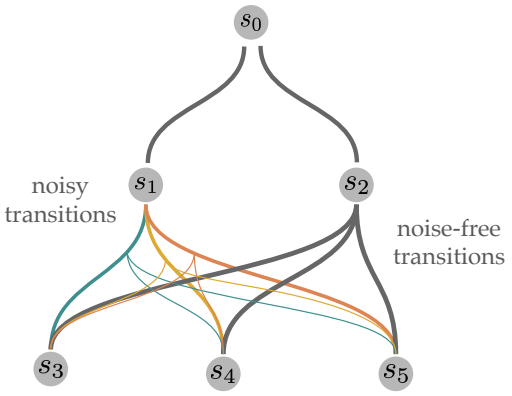


Figure 12: **Worked Example of Empowerment.**

Table 3: Skill occupancy distributions of the example MDPs. (a) Example X with mass on the two reachable targets. (b) Example Y with mass on the three reachable targets.

(a) Figure 5(a) MDP

	s_1	s_2
$z = 1$	1	0
$z = 2$	0	1

(b) Figure 5(d) MDP

	s_1	s_2	s_3
$z = 1$	$23/30$	$7/60$	$7/60$
$z = 2$	$7/60$	$23/30$	$7/60$
$z = 3$	$7/60$	$7/60$	$23/30$

In state s_1, the transitions are stochastic. If we similarly learned a set of skills to maximize mutual information, they would arrive at occupancy measures denoted by the orange dots in Fig. 3a. We can now interpret this example through the lens of empowerment. Is the empowerment $\mathcal{E}_{\text{pot}}(s)$ higher in state s_1 or s_2? Looking at Fig. 3a, we see that the skills at state s_2 are more "spread out"; formally, they convey more bits to the future. Thus, an empowerment-maximizing policy would prefer to navigate towards and stay at state s_2.

GLOSSARY

Gaussian width $w(\Theta) \triangleq \mathbb{E} \sup_{\theta \in \Theta} \langle \theta, Z \rangle$, $Z \sim \mathcal{N}(0, I_d)$, for $\Theta \subset \mathbb{R}^d$ (Polyanskiy and Wu, 2022, Eq. (27.19)) . 7, 8

$\mathcal{E}_{\text{eff}}$ the effective empowerment as a function of the policy . 4

$\mathcal{E}_{\text{pot}}$ the potential empowerment as a function of the state . 1, 3–6, 14–16, 21, 23, 25, 29

π_{eff} an effective empowerment maximizing policy . 1, 4, 6, 13, 16, 20, 21, 23

π_{pot} a potential empowerment maximizing policy . 4, 6, 25

ACRONYMS

DSOM discounted state occupancy measure. 3, 5, 7, 8, 13, 16, 17, 19–21, 23, 25, 26, 28

KL Kullback-Leibler. 5, 15, 28

MDP Markov decision process. 2–4, 6–9, 13, 16, 18–26, 28

MI mutual information. 4, 7–9, 13, 16, 18, 19, 21–25, 28

MISL mutual information skill learning. 2–4