# Efficient Skill Discovery via Regret-Aware Optimization

He Zhang [12] Ming Zhou [2] Shaopeng Zhai Ying Sun [1] Hui Xiong [13]

[1]Thrust of Artificial Intelligence, HKUST(GZ) [2]Shanghai AI Lab [3]The Hong Kong University of Science and Technology Hong Kong SAR

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

ICML International Conference On Machine Learning

## ABSTRACT & INTRODUCTION

Unsupervised skill discovery aims to learn diverse and distinguishable behaviors in open-ended reinforcement learning. For existing methods, they focus on improving diversity through pure exploration and mutual information optimization, but they remain limited in terms of efficiency, especially for high-dimensional situations. Previous work has generally overlooked the critical role of skill selection in improving learning efficiency and enhancing the diversity of skills.

To address these challenges, we frame skill discovery as a min-max game of skill generation and policy learning. We propose a regret-aware method that expands the discovered skill space along the direction of upgradable policy strength. The key insight is that skills with weak strength should be further explored, while less exploration is needed for skills with converged strength.

To implement this, we propose a method that scores the degree of strength convergence with regret and guides the skill discovery with a learnable skill generator. To avoid degeneration, skill generation comes from an up-gradable population of skill generators. Empirical results show that our method outperforms baselines in both efficiency and diversity. Moreover, our method achieves a 15% zero-shot improvement in high-dimensional environments compared to existing methods.

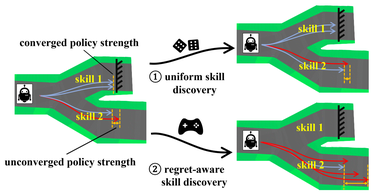



Figure 1. Comparison between uniform skill discovery and regret-aware skill discovery. Blue lines indicate skills with converged strength (low regret). Red lines indicate skills need further explore (high regret). By re-balancing exploration based on regret signals, the regret-aware method exhibits improved efficiency.

## PROBLEM FORMULATION

**Conditioned MDP.** Following prior work, we formulate the unsupervised skill discovery problem within a conditioned Markov Decision Process (MDP) framework. Conditioned MDP is formulated as $M = \langle S, A, P, R, \mu_0, Z \rangle$. The state space is $S \subseteq \mathbb{R}^n$, with $s \in S$ denoting an individual state with $n$ dimensions. The action space is $A \subseteq \mathbb{R}^m$, where $a \in A$ represents an individual action with dimensions $m$. The transition dynamics is given by the transition function $P(s'|s, a) \in [0, 1]$. $\mu_0(s)$ denotes the initial state distribution over $S$. And $Z = \{z | z \in \mathbb{R}^d\}$ denotes the latent space of the skill. Given a skill $z$, the skill-conditioned policy can be defined as $\pi(a \mid s, z) \in [0, 1]$ that satisfies $\sum_{a \in A} \pi(a \mid s, z) = 1$. With the time horizon denoted as $T$, the trajectory derived from $\pi$ can be formulated as: $\tau = \{(s_t, a_t, s_{t+1})\}_{t=0}^{T-1}$. If $p(z)$ denotes the distribution of skills, we have

$$p^\pi(\tau \mid z) = \mu_0 p(z) \prod_{t=0}^{T-1} \pi(a_t \mid s_t, z) P(s_{t+1} \mid s_t, a_t),$$

which indicates the probability of $\tau$ derived from $\pi$ and $z$. In the context of conditioned MDP, the learning objective of the unsupervised skill discovery problem is to find a policy $\pi$ that maximizes the expectation of the cumulative return over the whole skill space $Z$, which is formulated as :

$$V_\pi(s) = \mathbb{E}_{\substack{a_t \sim \pi(\cdot|s_t, z) \\ z \sim p(z) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left( \sum_t \gamma^t r(s_t, s_{t+1}, z) \mid s_0 = s \right), \tag{1}$$

where $r(s_t, s_{t+1}, z) \in R$ and $R : S \times S \times Z \to [-1, 1]$.

**Temporal Representation for Mutual Information.** Previous works achieve skill discovery by maximizing the mutual information (MI) between states and skills:

$$I(S; Z) = \mathbb{E}_{p(s,z)} \left[ \log \frac{p(s, z)}{p(s)p(z)} \right].$$

Based on the MI objective, METRA leverages Wasserstein Dependency Measure (WDM) to actively maximize the representation distance between different skill trajectories. The objective can be expressed using Kantorovich-Rubenstein duality as follows:

$$I_{\mathcal{W}}(S; Z) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p(s,z)}[f(s, z)] - \mathbb{E}_{p(s)p(z)}[f(s, z)], \tag{2}$$

where $f$ scores the similarity between each $s$ and $z$, and is formulated as an inner production of the state representation $\phi(s)$ and $z$, i.e., $f(s, z) = \phi(s)^\top z$. With further derivation, Eq. (2) is equivalent to maximize the following objective:

$$I_{\mathcal{W}}(s_T; Z) \approx \sup_{\|\phi\|_L \leq 1} \mathbb{E}_{p(\tau,z)} \left[ \sum_{t=0}^{T-1} (\phi(s_{t+1}) - \phi(s_t))^\top z \right],$$
$$\text{s.t.} \quad \sum_{t=0}^{h-1} \|\phi(s_{t+1}) - \phi(s_t)\| \leq 1. \tag{3}$$

Thus, maximization of $I_{\mathcal{W}}$ can be achieved by learning a temporal representation $\phi(\cdot)$. With that, METRA further gives a unified reward function for policy learning, i.e.,

$$r(s_t, s_{t+1}, z) = (\phi(s_{t+1}) - \phi(s_t))^\top z. \tag{4}$$

## METHOD

Our proposed method, Regret-aware Skill Discovery (RSD), frames unsupervised skill discovery as a min-max adversarial optimization problem between an agent policy and a skill generator policy. The method consists of three core components:

**1. Min-Max Optimization Framework** The overall objective is to find an equilibrium between an agent policy ($\pi_{\theta_1}$) that tries to minimize regret by mastering skills, and a skill generator policy ($\pi_{\theta_2}$) that tries to maximize regret by proposing new, challenging skills. Regret is defined as the improvement in the agent's value function at each learning stage $k$. This adversarial process is captured by the following objective:

$$\min_{\theta_1} \max_{\theta_2} \mathbb{E}_{z \sim P_z}[V^k_{\pi_{\theta_1}}(s_0|z) - V^{k-1}_{\pi_{\theta_1}}(s_0|z)]$$

**2. Agent Policy Learning in a Bounded Space** The agent policy, $\pi_{\theta_1}$, learns to master skills by maximizing a specially designed intrinsic reward. To ensure skills are distinguishable and to stabilize learning, we introduce a bounded temporal representation space where the state representation $\phi(s)$ and the skill vector $z$ live. The agent is rewarded for reducing the distance between its current state representation and the target skill vector representation. The intrinsic reward is formulated as:

$$r_\phi(s_t, s_{t+1}, z) = \|z - \phi(s_t)\| - \|z - \phi(s_{t+1})\|$$

**3. Regret-aware Skill Generator** The skill generator, $\pi_{\theta_2}$, is optimized to generate skills that are challenging for the current agent policy, thereby maximizing its regret. To prevent the generator from proposing skills that are impossible to learn or redundant, we introduce two regularizers. The final objective for the generator balances maximizing the expected regret with maintaining skill diversity ($d_z$) and staying close to the frontier of the agent's capabilities ($d_\phi$). The objective is:

$$\max_{\theta_2} J_{\theta_2}(z) = \mathbb{E}_{z \sim \pi^k_{\theta_2}}[Reg_k(z)] + \alpha_1 d_z + \alpha_2 d_\phi$$
$$d_z = D_{KL}\left(p(z|P^k_z) \middle\| \pi^k_{\theta_2}(z)\right)$$
$$d_\phi = \max_{\phi_{seen} \sim \mathcal{B}_k} \log \pi_{\theta_2}(\phi_{seen})$$

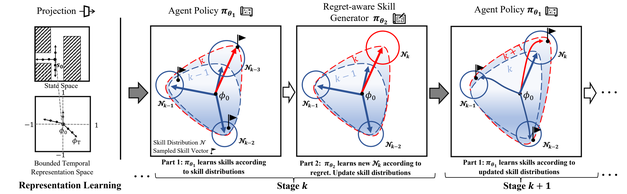
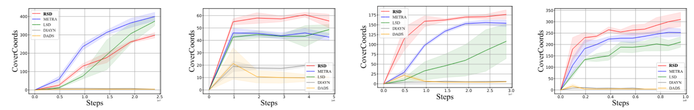



Figure 2. The leftmost illustrates how the state space is projected onto a bounded temporal representation space. The right side shows learning process of the skill generator: circles denotes skill distributions, dashed regions indicate coverage, the larger the more diverse, solid lines represent trajectories, and flags correspond to skills sampled from these distributions. Collectively, these circles form the population of skill generators for the agent's policy. In Part 1, the agent policy expands its coverage by mastering the skills (as seen by the red region outgrowing the blue region). In Part 2, the trains a new $\pi_{\theta_2}$ to generate high-regret skills, depicted by a red circle replacing a blue one. Here, $\mathcal{N}_k$ represents the skill distribution of the current $\pi_{\theta_2}$.

## EXPERIMENTS




(a) Ant. (b) Maze2d-large. (c) Antmaze-medium. (d) Antmaze-large.

Figure 3. Comparison between RSD and baselines, represented as red curves in different environments. The x-axis shows timesteps of interaction, while the y-axis represents Unique Coordinates, which measure state coverage achieved through sufficiently sampled skills.

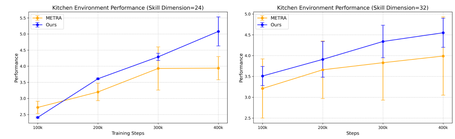



Figure 4. Our method consistently demonstrated superior sample efficiency and final performance compared to METRA. Particularly at a skill dimension of 32, the performance gains were increasingly pronounced as training progressed (e.g., at 300k steps was +0.51 compared to +0.36 at dimension 24). These improvements are clearly visible in performance curve visualizations, underscoring our method's robustness and scalability.

- **Efficiency and Diversity**: In the simple, skill-symmetric Ant environment, the proposed RSD method demonstrated performance competitive with top baselines like METRA and LSD. In the more complex, skill-asymmetric maze environments, RSD significantly outperformed all baselines in both learning efficiency and final state coverage. This is because baseline methods tend to explore primary directions akin to PCA, which is less effective in complex environments. In contrast, RSD's non-uniform sampling and non-unit skill vectors allow it to discover more granular and diverse skills.
- **Zero-shot Performance**: In zero-shot navigation tasks requiring no additional training, RSD achieved the best results on both success rate (AR) and final distance to goal (FD). The improvement in success rate was particularly pronounced in the higher-dimensional Antmaze-large environment.
- **Additional Experiments**: The method's effectiveness was further validated in the challenging, pixel-based Kitchen environment to demonstrate its capabilities in skill-asymmetric scenarios.

In Proceedings of the $42^{nd}$ International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

The authors' email: zhanghe18@tsinghua.org.cn, yings@hkust-gz.edu.cn, xionghui@ust.hk