Unifying Regret and State-Action Space Coverage for Effective Unsupervised Environment Design

Extended Abstract

Jayden Teoh Jing Xiang*
Singapore Management University
Singapore, Singapore
jxteoh.2023@scis.smu.edu.sg

Wenjun Li*
Singapore Management University
Singapore, Singapore
wjli.2020@smu.edu.sg

Pradeep Varakantham
Singapore Management University
Singapore, Singapore
pradeepv@smu.edu.sg

ABSTRACT

Unsupervised Environment Design (UED) employs interactive training between a teacher agent and a student agent to train generally-capable student agents. Existing UED methods primarily rely on *regret* to progressively introduce curriculum complexity for the student but often overlook the importance of environment novelty — a critical element for enhancing an agent's exploration and generalization capabilities. There is a substantial lack of investigating the effects of environment novelty in UED. This paper addresses this gap by introducing the GMM-based Evaluation of Novelty In Environments (GENIE) framework. GENIE quantifies environment novelty within the UED paradigm by using Gaussian Mixture Models. To assess GENIE's effectiveness in quantifying novelty and driving exploration, we integrate it with ACCEL, the state-of-the-art UED algorithm. Empirical results demonstrate the superior zero-shot performance of this extended approach over existing UED algorithms, including its predecessor. By providing a means to quantify environment novelty, GENIE lays the groundwork for future UED algorithms to unify novelty-driven exploration and regret-driven exploitation in curriculum generation.

KEYWORDS

Unsupervised Environment Design, Novelty Quantification, Gaussian Mixture Model

ACM Reference Format:
Jayden Teoh Jing Xiang*, Wenjun Li*, and Pradeep Varakantham. 2024. Unifying Regret and State-Action Space Coverage for Effective Unsupervised Environment Design: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

To train generally-capable RL agents, a surge of interest recently focused on Unsupervised Environment Design (UED, [2, 4–9, 12, 13]), which formulates a training framework between a teacher agent and a student agent. In UED, the teacher constantly creates new training environments (e.g., mazes with varying positions of obstacles) to improve the student's generalization ability such that it is robust to out-of-distribution (OOD) testing scenarios. UED algorithms have been empirically shown to help RL agents achieve state-of-the-art generalization performance.

We introduce the GMM-based Evaluation of Novelty In Environments (GENIE) framework, a novel approach to quantifying environmental novelty within the UED paradigm. GENIE utilises Gaussian Mixture Models (GMM) to assess the environment's capacity to provide novel experiences for the student agent. It's important to note that unlike previous works on quantifying novelty in UED [8, 12, 13], GENIE's novelty calculation method is scalable, domain-agnostic, and incorporates the agent's policy. We then empirically show the importance of novelty-based objectives for generalization by deploying GENIE on top of state-of-the-art [9].

2 APPROACH: GENIE

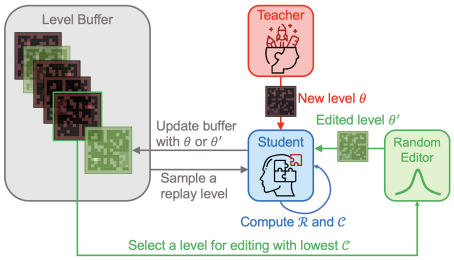

Figure 1: An Overview of GENIE when implemented atop ACCEL, where $\mathcal{R}$ and C are regret and novelty respectively.

2.1 Measuring the Novelty of a Level

We first list the notations in this paper. l_θ refers to the candidate level which we would like to compute the novelty for and it is conditioned by an environment encoding vector θ. We decompose the trajectory, τ_θ, of the agent in l_θ into a set of state-action pairs, i.e., $X_\theta = \{(s, a) \sim \tau_\theta\}$. L is the set of past training levels and $\Gamma = \{x = (s, a) \sim \tau_L\}$ contains all of the state-action pairs collected in levels within L. We treat Γ as the ground truth of the agent's state-action space coverage and compare the difference between Γ and the state-action pairs collected from the candidate level, i.e., X_θ.

First, we fit an initial GMM on the ground truth data Γ as such

$$P(\Gamma|\lambda_\Gamma) = \prod_{j=1} \sum_{i=1}^{K} \alpha_k \mathcal{N}(x_j|\mu_i, \Sigma_i), \quad (1)$$

*Equal contribution.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

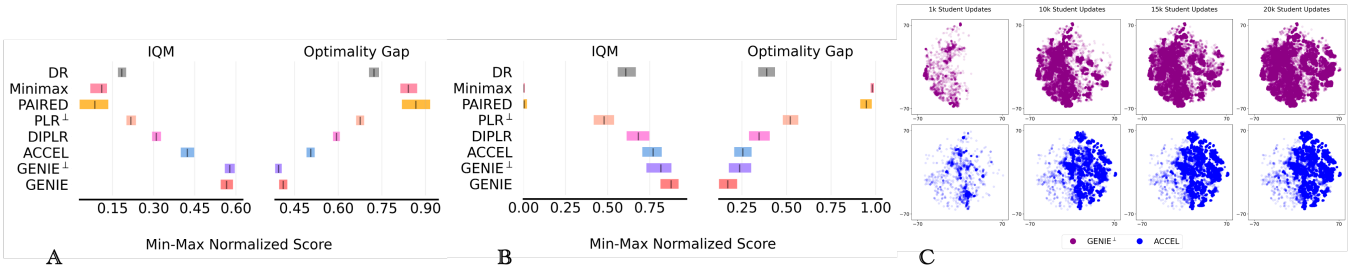

Figure 2: (A) Aggregate zero-shot performance in Bipedal-Walker Domain. (B) Aggregate zero-shot performance in Minigrid Domain. (C) Evolution of the state-action space coverage of GENIE$^{\perp}$ and ACCEL at 1k, 10k, 15k, and 20k policy updates.

where K is the number of Gaussians of the mixture with α_i representing the weight of each Gaussian component. $\mathcal{N}(x_n|\mu_i, \Sigma_i)$ is the multi-dimensional Gaussian function with mean vector μ_i and covariance matrix Σ_i, and $\lambda_\Gamma = \{(\alpha_1, \mu_1, \Sigma_1), ..., (\alpha_K, \mu_K, \Sigma_K)\}$ represents the initial set of parameters of the mixture model.

We optimise our GMM's parameters using the Expectation Maximization (EM) algorithm [3, 10]. The EM algorithm uses the initial λ_Γ to estimate a new λ'_Γ such that $P(\Gamma|\lambda'_\Gamma) > P(\Gamma|\lambda_\Gamma)$, and iterates this process until convergence to a small threshold.

Next, we estimate the difference of the state-action space coverage between the sample points X_θ and ground truth data Γ by computing the likelihood that the sample points in X_θ are observed under our GMM model. The log-likelihood function can be written as

$$\log \mathcal{L}(X_\theta|\lambda_\Gamma) = \sum_{j=1} \log p(x_j|\lambda_\Gamma) = \sum_{j=1} \log(\sum_{i=1}^{K} \alpha_k \mathcal{N}(x_j|\mu_i, \Sigma_i)) \tag{2}$$

We take the negative mean log-likelihood of X_θ as the novelty score of the candidate level, i.e.,

$$Novelty(l_\theta) = -\frac{1}{|X_\theta|} \log \mathcal{L}(X_\theta|\lambda_\Gamma). \tag{3}$$

A higher novelty score means that X_θ covers more novel state-action space compared to the ground truth data and therefore the candidate level induces more novel experiences for the agent. Consequently, we can compare the novelty of different levels using this metric.

2.2 State-Action Space Coverage Directed Training Agent

Now that we have established a method to compute the novelty of levels, we show its generality and effectiveness by deploying it on top of the state-of-the-art, i.e., the ACCEL algorithm, to motivate the student agent policy to cover more state-action space. For convenience, in subsequent sections of this paper, we will refer to this GENIE-augmented methodology of ACCEL simply as GENIE. The ACCEL algorithm performs mutation on levels with the lowest learning potential, essentially moving levels back to the learning frontier once their learning potential has been reduced. ACCEL relies solely on the regret of the level to determine its learning potential. To study the effectiveness of the proposed novelty metric, we use novelty alone to evaluate the learning potential of a level for mutation, i.e., $\alpha = 0$ in Equation (4).

$$LearningPotential = (1-\alpha) \cdot Novelty + \alpha \cdot Regret \tag{4}$$

Figure 1 provides a visual overview of the GENIE-augmented methodology of ACCEL. Furthermore, to better understand the coexistence of novelty and regret, we conducted an extended study, which we will refer to as GENIE$^{\perp}$, that uses an equal weightage of novelty and regret to select levels for mutation, i.e., set $\alpha = 0.5$.

3 EXPERIMENTS

We compare GENIE and GENIE$^{\perp}$ to a set of leading UED algorithms including ACCEL on two distinct domains, Minigrid and Bipedal-Walker. Minigrid is a partially-observable navigation problem under discrete control with sparse rewards, while Bipedal-Walker is a partially-observable walking task under continuous control with dense rewards. To make the comparison more reliable and straightforward, we employ the standardized DRL evaluation metrics [1], with which we show the aggregate inter-quartile mean (IQM) and optimality gap plots.

Figure 2A shows that both GENIE and GENIE$^{\perp}$ outperform all other benchmarks by a substantial margin, with their performance surpassing the next best algorithm, ACCEL, by over 30%. Similarly, Figure 2B demonstrates the superior perforance of both GENIE and GENIE$^{\perp}$ over the baseline approaches in the Minigrid domain. The remarkable performance of GENIE and GENIE$^{\perp}$ in comparison to its predecessor underscores the significance of novelty in enhancing agents' out-of-distribution performance. Additionally, considering the resemblance in the performance of both GENIE and GENIE$^{\perp}$, it suggests that novelty takes precedence over regret as a criterion for level mutation in both dense rewards and sparse rewards domains.

Finally, to better reveal how the incorporation of a novelty objective affects the curriculum generation, we also tracked the evolution of the agents' state-action space coverage during the training. State-action pairs encountered by the agent during training are collected for ACCEL and GENIE$^{\perp}$ in the Bipedal-Walker domain and projected onto two-dimensional manifold using t-SNE [11]. Figure 2C illustrates that GENIE$^{\perp}$ exhibits significantly broader coverage of the state-action space compared to the predecessor, ACCEL. This observation further substantiates the effectiveness of low-novelty level editing within the GENIE framework in generating curricula that enhance exploration capabilities for the agent.

REFERENCES

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems* 34 (2021), 29304–29320.

[2] Abdus Salam Azad, Izzeddin Gur, Jasper Emhoff, Nathaniel Alexis, Aleksandra Faust, Pieter Abbeel, and Ion Stoica. 2023. CLUTR: Curriculum Learning via Unsupervised Task Representation Learning. In *International Conference on Machine Learning*. PMLR, 1361–1395.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* 39, 1 (1977), 1–22.

[4] Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. 2020. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems* 33 (2020), 13049–13061.

[5] Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2021. Replay-guided adversarial environment design. *Advances in Neural Information Processing Systems* 34 (2021), 1884–1897.

[6] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. 2021. Prioritized level replay. In *International Conference on Machine Learning*. PMLR, 4940–4950.

[7] Dexun Li, Wenjun Li, and Pradeep Varakantham. 2023. Diversity Induced Environment Design via Self-Play. *arXiv preprint arXiv:2302.02119* (2023).

[8] Wenjun LI, Pradeep VARAKANTHAM, and Dexun LI. 2023. Generalization through diversity: Improving unsupervised environment design. (2023).

[9] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2022. Evolving Curricula with Regret-Based Environment Design. *arXiv preprint arXiv:2203.01302* (2022).

[10] Richard A Redner and Homer F Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* 26, 2 (1984), 195–239.

[11] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[12] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. 2019. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753* (2019).

[13] Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeffrey Clune, and Kenneth Stanley. 2020. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *International Conference on Machine Learning*. PMLR, 9940–9951.