A Toolkit for Encouraging Safe Diversity in Skill Discovery

Doctoral Consortium

Maxence Hussonnois
A^2I^2, Deakin University
Geelong, Australia
m.hussonnois@deakin.edu.au

ABSTRACT

Diversifying agents' skills has proven to be critical for adapting to a wide range of tasks. However, continuously promoting diversity can have catastrophic effects, such as accumulating unsafe, ineffective or misaligned skills. To avoid such outcomes, providing agents with the ability to modulate diversity in skill discovery remains a largely unexplored research area. In my research, I aim to design agents that can control and adapt their diversity to fit any context. Integrating context into skill discovery was my initial approach to controlling diversity. This was done by allowing the agent to use human preferences to identify regions of the environment where diversity is most likely to be desired. However, to modulate skill diversity, an agent has to be able to not only identify where to demonstrate diversity, but also comprehend how it affects the environment and others around it to decide when to be more (or less) diverse. To achieve this, an agent needs more tools, such as observing its own diversity and methods of adjusting it. The incorporation of controlled diversity will, I believe, make agents with multiple behaviors more flexible, reliable, and robustly applicable in a wide variety of contexts.

KEYWORDS

Skill Diversity; Human Preferences; Reinforcement Learning

ACM Reference Format:
Maxence Hussonnois. 2023. A Toolkit for Encouraging Safe Diversity in Skill Discovery: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Deep Reinforcement learning (DRL) [8] is a powerful computational approach for solving sequential decision making tasks by maximizing prespecified rewards over time. Despite its proven success in a number of applications ranging from Atari games to robotics [7, 8], DRL still fails to perform long-horizon tasks or to generalise to novel tasks. Hierarchical reinforcement learning approaches[12, 14, 15] have been developed to address these challenges by discovering sub-behaviors (skills) and learning to compose them. While discovering those skills, it is however, imperative to ensure that they are diverse enough to cover a large range of complex behaviors. To this end, prior works have proposed information theory-based objectives as an intrinsic motivation to ensure diversity in skill discovery [1, 4, 10].

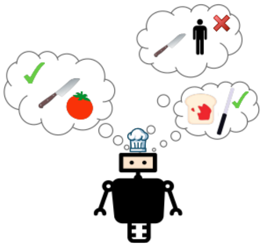

Figure 1: With unconstrained skill disovery, a cooking robot may discover undesirable skills (such as harming humans) using a kitchen knife. (Figure reproduced from [5]).

However, while such methods can produce promising results, blindly promoting diversity may lead to the acquisition of useless, dangerous, or misaligned skills. For example, as depicted in Figure 1, a robot tasked with learning diverse skills with a kitchen knife may learn undesirable skills such as harming a human. In addition, promoting diversity regardless of its effects on the environment or other agents is inadequate, dangerous and inefficient. In a multi-agent setting, excessive diversity in agents' behaviors can lead to confusion, and can negatively impact the performance of a team. For example, when crossing the road, if one agent's crossing strategy is difficult to predict, it could confuse and disrupt the actions of other agents, which could lead to a higher risk of accidents or collisions. This highlights the importance of balancing diversity with coherence and predictability in the actions of agents. In other words, agents should be able to adjust their level of diversity depending on the context.

Using reinforcement learning (RL)[11], my thesis proposes to develop tools for agents to modulate diversity in skill discovery. To modulate diversity, we address two problems: how to avoid encouraging distinct undesirable behaviors to be discovered and how to avoid encouraging diversity when the context doesn't favor it. To tackle the first issue, we identified that undesirable behaviors can occur because the agent lacks context about the real world. Without context, the agent views all aspects of the environment as equally relevant. Therefore, it learns to correlate its skills with any part of the environment regardless of its importance, relevance or safety. Our first step towards controlling diversity in skill discovery was to incorporate context through human feedback. We used this additional information to identify regions of the state space where diverse behaviors are desirable. We then used existing skill discovery methods and guided them to promote diversity in those preferred regions. By allowing the agent to discover diverse skills only in preferred regions, we can ensure that we are not promoting undesirable skills.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

To address the second problem, we identified that an agent is blind to its own diversity and how it can affect the environment. In addition, it is not in control of how much diversity is promoted. Therefore, we aim to train an agent to understand the relation between its diversity and a given context, to adapt it to either promote or reduce diversity. Consequently, we plan to give the agent the ability to estimate the diversity of its own skills and to adjust it in accordance with any context.

In summary, the main objectives of this research are to develop tools to modulate diversity in skill discovery. This can be described as agents who can: promote diversity among desired skills, observe its own diversity and understand context, that is, its effect on the environment and vice-versa and adapt finnaly its diversity to fit any context to enhance performance.

2 METHODS

2.1 Controlled Diversity with Preference

First, as outlined in Hussonnois et al. [5], we focus on the problem of controlling diversity in unsupervised skill discovery to avoid encouraging distinct undesirable skills.

In this work, we contend that the agent can learn more desirable skills through guidance provided by humans in the loop during the learning of skills. The key idea behind our approach was that we framed the problem of *controlling skill diversity* as finding regions of the environment where skill discovery will more likely produce desirable skills.

Due to the difficulty of identifying such regions without human-provided context, we proposed leveraging recent work in learning from human preferences [3, 6, 13] to infer preferred regions in the environment. Intuitively, these are regions of the environment that are generally associated with favorable agent behaviors. We posited that such regions also correspond to suitable regions for learning a diverse set of skills. In practice, we defined a preferred region as those regions associated with high estimated preference rewards, where the preference reward was learnt using the preference-based RL framework[3, 6, 13]. Once such regions were identified, we integrated them into the EDL framework [1] for a more efficient exploration of the preferred region.

Furthermore, by learning a representation of the state space from human preferences, we showed that our approach scales to higher dimensional problems and learns skills that are discernibly diverse to human eyes. Specifically, we introduced the preferred latent representation by simply using the output of the last hidden layer of the reward model learnt from human preferences. The intuition was that the last hidden layer of the neural network that models the internal reward function of a human would learn a latent state representation that captures features that matter for human preferences.

Finally, using a 2D navigation environment and Mujoco environments, we demonstrated our approach's capability to discover diverse, yet desirable skills.

2.2 Learning when to promote diversity

After focusing on controlling diversity with preferences, we now intend to explore how the agent can determine when to adjust its level of diversity in regards to a context.

In this regard, we place ourselves in hierarchical RL settings, where there is an extrinsic reward and we are simultaneously learning a high level policy and low level policies (the skills). The high level policy determines which skill to use in a given state, for a specific duration of time, and is rewarded with an extrinsic reward. Skills are responsible for taking low-level actions, and are rewarded with an extrinsic reward and an intrinsic reward that promotes diversity. In this work, we wish the high level policy to have the ability to control the weighting of the intrinsic reward relative to the extrinsic reward, thereby enabling the extent of diversity.

To achieve this goal, we aim to develop a measure of diversity describing how diverse an agent is at a specific time and state. With this measure of diversity, we could provide the agent with additional information about its own diversity. This will enable the agent to base its decision on whether or not it needs to promote diversity. That is, whether the weighting of the intrinsic reward should be increased (to promote diversity) or decreased (to suppress diversity). We contend that this measure could be defined as the expected sum of intrinsic rewards promoting diversity for each skill. It can be seen as a distribution of how distinct each skill can be, starting from the current state. Then, we can choose to let a high level policy adjust the diversity by predicting the weight of the intrinsic reward that promotes diversity for the selected skills.

The implementation of high-level control mechanisms in the promotion of diversity within an agent's decision-making process may yield a range of potential outcomes and implications in the field of reinforcement learning. In cooperative settings, such as the one described in the introduction, adapting agents's diversity to reduce confusion might be appropriate. A similar experiment can be conducted in an adversarial setting, where maximizing the confusion of the other agent can be advantageous. Alternatively, this work could find applications in safe reinforcement learning settings. In such a setting, the agent would consider the level of safety in its surroundings and adjust its level of diversity accordingly.

3 FUTURE WORKS

Recent work [9] in skill discovery has been working to promote more diverse, dynamic, and far-reaching skills. These methods cover the state space of the environment well, but not the behavior space. They typically discover monotonic skills that move toward a particular direction, but never discover motions such as moving in circles or zig-zagging. While promoting more diversity among dynamic, and far-reaching skills may be beneficial for locomotion tasks, it may not be sufficient for manipulation tasks, where object-oriented diversity has been shown to be more performant [2]. In other words, the desired characteristics of skills may differ depending on the nature of the future task, thus necessiting the need for different diversity objectives. As coming up with novel objectives for promoting diversity may be challenging, it might be more practical to learn those objectives through human interaction. As shown in our first paper [5], promoting diversity in the preferred feature of the state space helps to discover more diverse and task-relevant skills. Future work may benefit from exploring further these issues of task-aligned diversity.

REFERENCES

[1] Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giro i Nieto, and Jordi Torres. 2020. Explore, Discover and Learn: Unsupervised Discovery of State-Covering Skills. In *ICML*.

[2] Daesol Cho and Jigang Kim. 2022. Unsupervised Reinforcement Learning for Transferable Manipulation Skill Discovery. *IEEE Robotics and Automation Letters* 7 (07 2022), 1–1. https://doi.org/10.1109/LRA.2022.3171915

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[4] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Diverse Skills without a Reward Function. (2018).

[5] Maxence Hussonnois, Thommen Karimpanal George, and Santu Rana. Accepted. Controlled Diversity with Preference : Towards Learning a Diverse Set of Desired Skills. *AAMAS, 2023* (Accepted).

[6] Kimin Lee, Laura Smith, and Pieter Abbeel. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. *International Conference on Machine Learning* (2021).

[7] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *CoRR* abs/1509.02971 (2016).

[8] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.

[9] Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. 2021. Lipschitz-constrained Unsupervised Skill Discovery. In *International Conference on Learning Representations*.

[10] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. 2020. Dynamics-Aware Unsupervised Discovery of Skills. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJgLZR4KvH

[11] Richard S. Sutton and Andrew G. Barto. 2005. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks* 16 (2005), 285–286.

[12] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1 (1999), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

[13] Aaron Wilson, Alan Fern, and Prasad Tadepalli. 2012. A Bayesian Approach for Policy Learning from Trajectory Preference Queries. In *NIPS*.

[14] Jiachen Yang, Igor Borovikov, and Hongyuan Zha. 2019. Hierarchical Cooperative Multi-Agent Reinforcement Learning with Skill Discovery. In *Adaptive Agents and Multi-Agent Systems*.

[15] Jesse Zhang, Haonan Yu, and Wei Xu. 2021. Hierarchical Reinforcement Learning By Discovering Intrinsic Options. (01 2021).