Learning Diverse Skills with Proximal Policy Optimization

Tomer Ziv

Bachelor of Science in Computer Science and Artificial Intelligence
University of Bath
Academic Year 4

Learning Diverse Skills with Proximal Policy Optimization

Submitted by: Tomer Ziv

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Bachelor of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

This dissertation explores the enhancement of skill acquisition in autonomous agents operating within complex environments, using the Proximal Policy Optimization (PPO) algorithm integrated with Diversity is All You Need (DIAYN), an unsupervised skill discovery method. The primary objective is to investigate whether DIAYN can learn diverse skills to improve the adaptability and performance of agents trained with PPO in partially observable environments of varying complexity, particularly the Bipedal Walker in both its basic and hardcore versions. We implement PPO as a baseline to assess basic locomotion skills and extend it with DIAYN to encourage the development of diverse, task-agnostic skills. Performance evaluations are conducted on intrinsic reward structures, skill discriminators, and captured trajectories. Principal Component Analysis (PCA) is utilised to visualize and interpret the skill exploration within the state space. Experiments are conducted over extensive training periods to investigate skill refinement and generalization capabilities across different environmental settings. Findings indicate that while PPO excels in predictable and less complex environments, it suffers in the hardcore version of the Bipedal Walker due to its inability to effectively navigate novel obstacles and complex terrains. Integrating DIAYN with PPO in a partially observable paradigm, showcases the limitations of this method in skill acquisition and provides a semantic state space interpretation to compare with more advanced methods. PCA confirms that the skills learned, while somewhat diverse, are inapplicable for downstream tasks; thus paving the way for better methods to overcome the challenges discovered. An important step in the development of Hierarchical Reinforcement Learning methods for real-world applications where environmental accessibility is limited. This research has implications in real-life and simulated domains where autonomous exploratory behaviour is required, such as search-and-rescue robots, or self-driving cars.

Key Terms: Proximal Policy Optimization, Diversity is All You Need, Skill Discovery, Skill Acquisition, Autonomous Agents, Reinforcement Learning, Hierarchical Reinforcement Learning, Bipedal Walker, Principal Component Analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Robotic systems have traditionally excelled at performing repetitive tasks in controlled environments, such as manufacturing products on an assembly line. However, creating robots that can adapt and respond to dynamic and unstructured environments, by observing their surroundings and making decisions in real time, remains a challenge (Irpan, Team and Pastor, 2018). The pursuit of this capability underpins the motivation for this research, particularly through the lens of skill discovery within Reinforcement Learning (RL). Skill discovery involves the autonomous identification and learning of diverse actions, or "skills", that an agent can employ to navigate task-agnostic environments effectively.

This concept is particularly crucial for developing versatile autonomous agents capable of completing complex sequences of tasks, without predefined sub-goals. *Sub-goals* are smaller rewarded objectives within a long-horizon overall goal. For instance, consider the sub-task of "driving through traffic" as part of the broader objective of reaching a destination with self-driving cars. This sub-task demands that the agent manoeuvres the vehicle around obstacles, without explicit sub-goals related to each manoeuvre (Pateria et al., 2021). By learning a set of diverse skills autonomously, an agent could potentially access pre-learned skills to navigate its tasks more efficiently, much like using higher-level cognitive strategies to approach problems. For this purpose, a "skill" semantically represents the ability to do something well.

Skill discovery not only aids in performing predefined tasks, but also enhances the agent's ability to generalize these capabilities to new, unseen environments. For example, Sharma, Resident and Research (2020) demonstrated this with a robotic hand learning tasks such as grasping or pointing. Each skill, a composite of numerous actions like adjusting joint torques, substantially reduces the learning burden by abstracting complex action sequences into manageable skills. This method improves sample efficiency by allowing an agent to use pre-learned skills as prior knowledge, before learning to perform a larger task. This method also provides an infrastructure for further learning, which is critical in Hierarchical Reinforcement Learning (HRL) systems, where multiple learned models, or skills, are strategically integrated to perform higher-order tasks. Moreover, skill discovery is beneficial for agents operating in partially observable environments, typical of real-world conditions, where complete environment

information is not always available (Pateria et al., 2021). To better understand this undertaking, a "Reinforcement Learning" (RL) baseline is required.

1.2 Reinforcement Learning

RL, a branch of Artificial Intelligence, addresses the challenge of an agent navigating an unknown environment to achieve a defined goal. This paradigm is based on the hypothesis that all goals can be expressed through the maximisation of the expected cumulative reward for completing a task. The agent learns to perceive and influence a state in the environment by taking actions, aiming to achieve maximum reward. The foundation of RL is rooted in the optimal control theory of Markov Decision Processes (MDP). MDPs provide a mathematical structure for modelling decision-making in scenarios where outcomes are influenced partly by a decision maker, and partly by a degree of randomness. The key components of an RL system, as outlined by an MDP, encompass the following (Bellman, 1957):

- Agent or learner
- Environment with which the agent interacts
- Policy dictating the agent's actions
- Reward signal observed by the agent upon taking actions

An essential component, derived by the reward signal, is the value function, which effectively captures how "good" it is to be in a given state. While the reward signal signifies the immediate benefit of occupying a particular state, the value function denotes the expected discounted return when starting in the given state, and following the policy thereafter. The overall objective of RL algorithms is to learn an action policy (a specific sequence of actions) that maximises the average value of each state within the system. SYNOPSYS (2023).

In the RL paradigm, there are many approaches to solve different problems. Lately, Sharma, Resident and Research (2020) outline the limitations of teaching an agent to perform complex behaviours using well-designed task-specific rewards. Designing reward functions can require significant engineering effort, which becomes unattainable for many tasks. For many practical scenarios, designing a reward function can be complicated, for example, requiring additional instrumentation for the environment such as sensors, which are expensive to implement and maintain. *Supervised RL* employs an extrinsic reward function from the environment, which guides the agent towards the desired behaviours, reinforcing the actions which bring the desired changes in the environment. Considering that the ability to generate complex behaviours is limited by this form of reward-engineering, unsupervised learning presents itself as an interesting direction for RL. *Unsupervised RL* uses an intrinsic reward function (such as curiosity to try different things in the environment) to generate its own training signals to acquire a broad set of task-agnostic behaviours, while being generic and broadly applicable to several agents and problems without any additional design.

POMDPs. Partially Observable MDPs define environments where the whole environment

is not accessible for the agent to observe, so external rewards are not always available to the agent. These are more realistic environments, as in real-world scenarios the environment is almost always larger than the observable subspace available to the agent at any point in time.

The problem investigated has to do with learning diverse, task-agnostic skills in partially observable environments. This has been addressed by a class of methods within the domain of RL known as skill discovery — Though mostly on fully observable environments. The unsupervised skill discovery method explored is called "Diversity is all you need" (Eysenbach et al., 2018b), *DIAYN* in short.

DIAYN. DIAYN represents a significant stride in unsupervised RL, focusing on skill discovery by rewarding agents for developing distinguishable skills based on the diversity of states they visit (Eysenbach et al., 2018b). This approach pivots away from direct action-based differentiation to state-based skill distinction, providing a foundation for exploring diverse strategies. While DIAYN lays the groundwork for further refined approaches, open questions remain regarding its integration with reinforcement learning methods, like PPO (Schulman et al., 2017), as well as extending our understanding of the skills that can be learned in partially observable environments and how these skills could be further used to improve performance on downstream tasks. For this, an appropriate test environment was chosen.

1.3 Bipedal Walker

Bipedal Walker serves as an ideal test environment for this research. It includes a basic version, requiring navigation over uneven terrain, and a more challenging "hardcore" version, with obstacles like ladders and pitfalls. This differentiation allows for examination of skill transferability and generalization from simple to complex environments. The environment's partially observable nature, where agents rely on internal sensors rather than explicit state information, further aligns with real-world application scenarios, making it an appropriate choice for studying skill discovery dynamics (Klimov, 2023). Furthermore, its motor speed actions present an appropriate locomotive control problem for DIAYN. Appendix A details this environment as described by OpenAI's Gymnasium (Klimov, 2023).

1.4 Investigation

This dissertation investigates several key aspects of skill discovery in RL:

- The efficacy of on-policy versus off-policy learning paradigms.
- The ability of agents to generalize skills learned in simple environments to more complex scenarios.
- The nature and utility of skills learned in partial observable environments.

These inquiries are structured to fill existing gaps in the literature and provide a nuanced understanding of the mechanisms underpinning skill acquisition and application in RL.

Following this introduction is a discussion of the research and theoretical advancements relevant to this study. This is followed by the details of the employed experimental design and methodologies. The results and analysis is next, presenting findings and interpreting them in the context of the posed research objectives. A summary and potential directions for future research is then outlined, before finally the appendix, detailing any necessary supplementary information.

Chapter 2

Literature and Technology Survey

2.1 Reinforcement Learning

RL methods can be broadly classified into two categories: model-free and model-based approaches. Model-free algorithms do not construct an explicit model of the environment. These algorithms conduct experiments with the environment through actions, directly deriving the optimal policy. Model-free techniques further divide into value-based and policy-based methods. Value-based algorithms calculate the optimal policy directly from the value function for each state. After the agent samples trajectories of states and rewards to estimate the MDP's value function, determining the optimal policy involves acting greedily with respect to the value function at every state in the process. Contrastingly, policy-based algorithms directly estimate the optimal policy without modelling the value function. These approaches parameterize the policy directly using learnable weights. However, issues with policy-based approaches stem from high variance, leading to instabilities during the training process. And, value-based approaches, while more stable, are less suited for modelling continuous action spaces. A solution that overcomes the limitations of both value-based and policy-based approaches is the actor-critic algorithm. In this hybrid method, both the policy (actor) and the value function (critic) are parameterized, facilitating the effective utilization of training data with stable convergence. This allows for a more robust and versatile learning process (SYNOPSYS, 2023).

On vs. Off Policy. On-policy algorithms perform the basic RL procedure, trying to optimize a policy and value function, however, off-policy algorithms split the responsibilities of the policy function into an *update policy* and *behaviour policy*. The update policy determines how the agent learns the optimal policy, and the behaviour policy determines how the agent acts in the environment (Mao, 2019).

Consider the grid world shown in Figure 2.1, as described by Sutton and Barto (2018). This is a standard undiscounted, episodic task, with start and goal states, and the usual discrete actions for movement up, down, right, and left. The reward is 1 for all transitions except those into the region marked "The Cliff". Stepping into this region incurs a reward of 100 and sends the agent instantly back to the start. Take two methods with ϵ-greedy action selection (GeeksforGeeks, 2023), with $\epsilon = 0.1$, one on-policy and one off-policy. ϵ—*greedy* is a simple method to balance exploration and exploitation by choosing between exploration

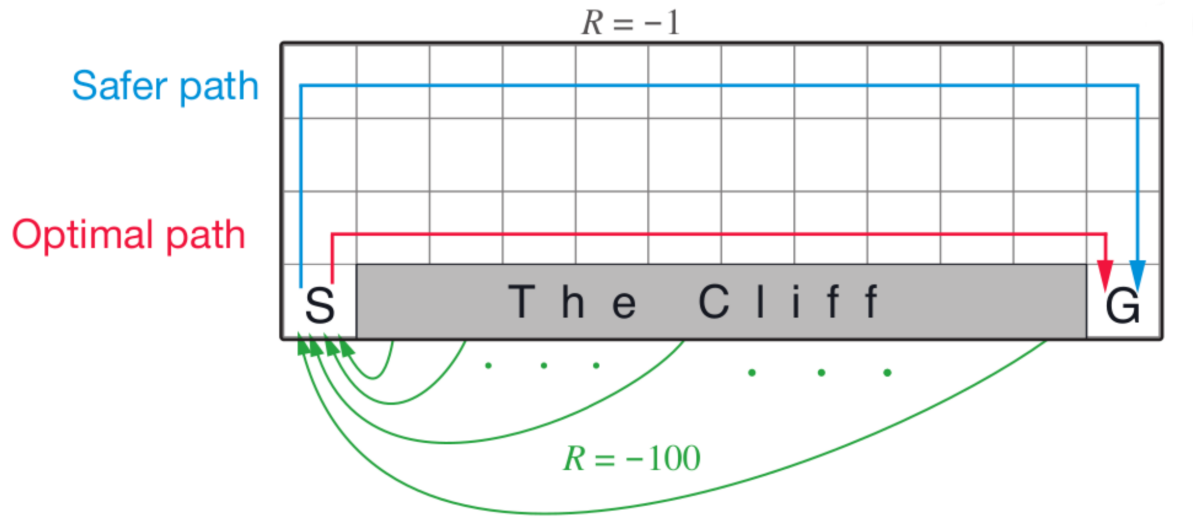

Figure 2.1: Cliff Walking Grid World Environment

and exploitation randomly with probabilities ϵ and $1 - \epsilon$, respectively. The on-policy algorithm learns values for the optimal policy, that which travels right along the edge of the cliff, and results in the agent occasionally falling off the cliff due to the ϵ-greedy action selection. The off-policy algorithm, on the other hand, takes the action selection into account and learns the longer but safer path through the upper part of the grid. Although the on-policy algorithm actually learns the values of the optimal policy, its on-line performance is worse than that of the off-policy algorithm, which learns the safer, roundabout policy. Naturally, if ϵ were gradually reduced, then both methods would converge to the optimal policy.

2.1.1 Policy Gradient Methods

These methods do not use value functions. Instead, they search in spaces of policies defined by a collection of numerical parameters, and estimate the directions the parameters should be adjusted in order to most rapidly improve a policy's performance. This produces estimates while the agent is interacting with its environment, and so can take advantage of the details of individual behavioural interactions. These have proven useful in many problems, and some of the simplest RL methods fall into this category. Some of these methods might take advantage of value function estimates to improve their gradient estimates (Sutton and Barto, 2018). Policy gradient methods are well suited for action-continuous environments, like our chosen bipedal-walker, because it is not plausible to compute a state value considering the number of possible actions in any given state. A better way is to update the parameters directly based on the policy and reward. Proximal Policy Optimization (PPO) and Soft-actor critic (SAC) are 2 of the best on and off-policy gradient actor-critic algorithms, respectively.

PPO. PPO (Schulman et al., 2017) represents a significant advancement in policy gradient methods for RL, offering an efficient and reliable approach for large-scale applications. In addition to an effective variance reduction scheme for policy gradients, called Generalized Advantage Estimation (GAE), as proposed by Schulman et al. (2015); PPO's key contribution lies in its optimization technique, which seeks to maintain a balance between exploration and

exploitation by limiting the updates to the policy to a small range. This is achieved through the use of a clipped objective function, ensuring that the policy does not deviate excessively from its previous iteration, thereby improving the stability and robustness of the learning process. PPO's ability to handle multiple epochs of data within a single update, combined with its straightforward implementation, makes it a versatile choice for a wide range of RL tasks.

SAC. SAC is an off-policy actor-critic deep RL algorithm, that incorporates the principle of entropy regularization into the RL paradigm, aiming to achieve a balance between exploration and exploitation. Unlike PPO, which operates on an on-policy basis, requiring fresh samples from the policy being learned, SAC can utilise a replay buffer to learn from experiences more efficiently. This fundamental difference allows SAC to improve sample efficiency and learn more stable and robust policies. SAC optimizes a stochastic policy in an off-policy manner, making it well-suited for tasks requiring continuous action spaces and providing an inherent exploration mechanism through entropy maximisation (Haarnoja et al., 2018).

Differences Between PPO and SAC. While both PPO and SAC are state-of-the-art RL algorithms, they exhibit distinct differences in their approach to policy optimization. PPO's on-policy nature requires new data for each update, focusing on making small, safe updates to the policy, whereas SAC's off-policy approach and use of a replay buffer allow for more extensive use of previously collected data, promoting sample efficiency. Additionally, SAC's incorporation of entropy regularization inherently encourages exploration by rewarding policies for uncertainty, contrasting with PPO's clipped objective function designed to prevent drastic policy updates. These differences make PPO and SAC suitable for various applications, with PPO being favoured for its simplicity and stability, and SAC for its efficiency and effectiveness in complex environments with continuous action spaces.

2.2 Skill Discovery

A prominent challenge faced by all these RL methodologies arises when confronted with environments featuring sparse or no extrinsic rewards within a goal-oriented context; When, only upon the completion of the ultimate goal, reward is given to the agent by the environment. This often leads to a situation where the agent engages in random movements and speculative actions without receiving any goal-related rewards. This poses a substantial issue for RL agents, as the delayed and infrequent feedback makes learning and optimizing behaviour significantly more challenging. As a consequence, there is a crucial need to address the challenges associated with sparse or delayed reward structures, enhancing the agent's learning capabilities and accelerating the convergence towards optimal policies (Pathak et al., 2017).

In response to these challenges, an open research domain is that of skill discovery in RL, an avenue of exploration aimed at identifying and achieving sub-goals within larger goal-oriented environments. This approach leverages the concept of intrinsic motivation.

Intrinsic Motivation. Intrinsic motivation is a mechanism for instilling artificial agents with inherently rewarding behaviours such as exploration and curiosity. Similarly in human psychol-

ogy, intrinsic motivation is defined as the drive to engage in an activity for the sheer satisfaction, fun, or challenge it presents, independent of external rewards or goals. For intelligent agents, this manifests when the information content alone or the experience resulting from an action becomes the motivating factor for the agent to take the action (Oudeyer and Kaplan, 2008).

Information Theory. Information theory defines the quantification of uncertainty as a foundation for measuring information content. One of the key metrics in information theory is entropy, which quantifies the amount of uncertainty associated with the value of a random variable or the outcome of a random process. For instance, the higher the uncertainty in a variable or process, the higher entropy it exhibits (Shannon and Weaver, 1998).

Mutual Information. Furthermore, the concept of mutual information (MI) (i.e. information gain) measures the mutual dependence between two random variables. Specifically, MI quantifies the "amount of information" obtained about one random variable by observing the other. This measure is closely linked to the entropy of a random variable as a fundamental measure that quantifies the expected amount of information held in a random variable (Shannon, 1948; Kreer, 1957).

Through navigating the landscape of skill discovery and intrinsic motivation methods, in the subsequent sections, it is shown that the fusion of information-theoretic principles and RL methods is essential for addressing the challenges associated with sparse reward environments. In the context of this project, an exploration into the realm of skill discovery is undertaken, specifically investigating the skills learned by DIAYN (Eysenbach et al., 2018b). Both DIAYN and VIC (Gregor, Rezende and Wierstra, 2016) operate based on mutual information, employing it as a foundational principle for intrinsic motivation, to develop skills that accomplish sub-goals within their respective tasks. The comparative analysis extends to the respective mutual information-based methods employed by DIAYN and VIC, analysing their efficacy in skill acquisition for autonomous agents.

2.2.1 Empowerment

The concept of Empowerment is foundational within RL, aiming to quantify an agent's degree of freedom and formalize this as an agent's preparedness, as shown in the paper "Empowerment — An Introduction" (Salge, Glackin and Polani, 2013). It is hypothesized that preparedness is a valuable indicator, distinguishing promising from unpromising regions in the state-space without the need to evaluate the entire state-space. Empowerment is defined through potential information flow or channel capacity, representing the agent's capacity to reliably influence the world. This quantification is achieved by assessing the information flow, between the agent's previous (actuator) state and its next (sensor) state, offering a means to measure the agent's ability to influence the world. This measure is widely used in further research toward solving reward-sparse environments in RL.

2.2.2 Mutual Information Based Methods

The use of mutual information, as proposed in the paper "Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning" (Mohamed and Rezende, 2015), presents mutual information as a metric for empowerment in environments characterized by sparse rewards. Here, the fundamental principle involves utilizing mutual information to quantify internal drives, enabling the agent to discern the value of information within action-observation sequences. This study addresses the challenge of formally defining internal drives, with unsupervised approaches, that empower the agent to evaluate the significance of information within its experiences. As a result, mutual information emerges as an essential measure, facilitating this type of reasoning, and forms the foundation for the intrinsic reward metric, empowerment.

Within RL, the exploration-exploitation trade-off is a well known challenge, demanding effective strategies to balance the agent's experimentation with novel approaches and the maximisation of rewards through known successful behaviours. Addressing this issue, the study on "VIME: Variational Information Maximising Exploration" (Houthooft et al., 2016) contributes a novel approach stemming from a curiosity-driven exploration strategy. This strategy uses information gain (i.e. mutual information) concerning the agent's internal belief of the dynamics model as a driving force. Drawing inspiration from the human psychological concepts of curiosity and surprise, the framework encourages actions that lead to states deemed surprising, resulting in a substantial update to the environment dynamics model distribution.

The practical implementation of this approach involves measuring information gain using Variational Inference (Ganguly and Earp, 2021), a method of estimating complex probability densities. This strategy stands out by addressing the exploration problem in continuous control tasks, using neural networks to represent the agent's understanding of the environment dynamics. Moreover, the model introduces a unique interpretation of curiosity, as measuring compression improvement, demonstrating scalability to continuous state and action spaces.

Options. An option or *macro-action* is a generalization of the concept of action, as proposed by (Sutton, Precup and Singh, 1999), to capture the idea that certain actions are composed of other sub-actions. This is a tuple composed of an initiation set, a policy, and a termination condition. An option would constitute picking up an object, going to lunch, or travelling to a distant city, as opposed to primitive actions such as muscle twitches and joint torques.

Variational Intrinsic Control (VIC)

VIC (Gregor, Rezende and Wierstra, 2016) is a method designed for discovering the set of intrinsic options accessible to an agent without explicit supervision. The key contribution lies in maximising the mutual information between the set of options and option termination states. VIC utilises two distinct policy gradient-based algorithms: one creating an explicit embedding space of options; and the other representing options implicitly. VIC specifically maximises empowerment over closed loop policies rather than open-loop policies, considered in much of the previous work. *Closed-loop policies* are policies that take an input and condition their output on feedback from the environment, unlike *open-loop policies*, which simply take an input and produce an output, irrespective of any feedback. This approach demonstrates

scalability with function approximation, showcasing its applicability across various tasks.

The main objective of this study is to answer the question: What intrinsic options are available to an agent in a given state? Specifically, what consequences these options have on the environment. Unlike traditional option learning, focusing on the identification of few task-specific options, VIC operates on a larger space of intrinsic options. This scope is justified by the potential benefit of exploring a broad array of possibilities, enabled by learning good embeddings, representative of these options.

Addressing Sparse Rewards and Long Task Horizons

In the pursuit of addressing the challenges posed by sparse rewards or extended task horizons in deep RL, this paper on "Stochastic Neural Networks for Hierarchical Reinforcement Learning" (Florensa, Duan and Abbeel, 2017) proposes a comprehensive framework. The background of this work displays the limitations of conventional exploration strategies, particularly in tasks with sparse rewards or extended time horizons. The key contribution lies in the integration of intrinsic motivation and hierarchical methods to facilitate skill learning. This approach involves pre-training the agent to acquire a diverse set of skills in a controlled environment, using Stochastic Neural Networks (SNNs) and an information-theoretic regularizer, based on Mutual Information. This is done prior to the training of the agent in a given environment. The design incorporates a single proxy reward, requiring minimal domain knowledge about downstream tasks. A high-level policy is subsequently trained on the acquired skills, leading to a substantial enhancement of exploration capabilities and improved performance in tasks with sparse rewards. The framework presents a promising solution to issues posed by naive exploration strategies, offering an efficient approach to learning interpretable skills and achieving substantial performance gains in various challenging tasks.

2.3 Diversity is All You Need (DIAYN)

DIAYN (Eysenbach et al., 2018b), as explored in this project, builds on VIC and introduces an original method for unsupervised skill discovery, grounded in three fundamental principles.

- Firstly, the approach defines the utility of skills by ensuring that each skill dictates the states the agent visits, making the skills distinguishable by the states they incur.
- Second, the method utilises states, rather than actions, for skill distinction, because actions without observable environmental impact are not externally discernible.
- Lastly, the paper incentivizes diverse skills by encouraging exploration and learning skills that exhibit maximal entropy (i.e. the most available subsequent options).

In terms of the objective, the study maximises the mutual information between skills and states, signifying that skills should control the states the agent visits. The resulting relationship ensures that skills are distinguishable from the states visited. To prioritize states over actions for skill distinction, the mutual information between skills and actions given the state is

minimized. The paper considers all skills as a mixture of policies, and thus maximises the entropy of this mixture policy. The implementation of DIAYN uses a soft actor-critic algorithm to learn a policy and thus maximise the policy's entropy over actions. This methodology proposes a means of learning useful skills in the absence of rewards, as well as demonstrates the unsupervised emergence of diverse skills, such as running and jumping, across various simulated robotic tasks. Thus, DIAYN presents an avenue for unsupervised skill discovery, with potential applications in autonomous robotic systems.

There are a few open problems suggested by the authors (Eysenbach et al., 2018a), ranging from simple modifications, like changing the environment, to hard modifications, such as using pre-trained skills hierarchically. Some options of note include

- Applying DIAYN to a new environment, such as Bipedal Walker.
- Investigating the effect on the skills learned of using PPO have rather than using SAC to maximise the action entropy.
- Investigating how the skills learned can be better used to solve downstream tasks.

These will be further explored throughout this research.

DIAYN's objective limits the utility of the skills learned by the agent. Sharma et al. (2020b) demonstrate that diversity is not sufficient for the learned skills to be useful for downstream tasks due to the high variance of the DIAYN skills limiting their temporal compositionality: the composition of behaviours from different timesteps within a trajectory. As such, DIAYN serves as a stepping stone to newer, more practical methods.

2.4 Further Improvements

Decoding From Trajectories

The following study introduces an improvement to the foundational skill discovery methods of VIC (Gregor, Rezende and Wierstra, 2016) and DIAYN (Eysenbach et al., 2018b), as presented above. This paper introduces these methods as instances of decoding skills from states in the environment. The primary contributions include the introduction of "Variational Autoencoding Learning of Options by Reinforcement" (VALOR) (Achiam et al., 2018), a novel method, decoding trajectories to encourage the learning of dynamic modes over goal-attaining modes. Here, a curriculum learning approach is proposed. The curriculum learning approach is designed to enhance the agent's exposure to different contexts as its performance improves, as measured by the decoder. This results in the visible scope of the environment available to the agent, to increases as performance on the current scope improves. This is comparable to how a child might learn a simplified version of a problem before being exposed to the full problem.

A comparative analysis undertaken in this study includes, VALOR, VIC, and DIAYN, with and without the curriculum approach and evaluated across various robotics environments. The results indicate that the three methods perform similarly, however, VALOR demonstrates qualitatively different behaviour due to its trajectory-centric approach, and DIAYN exhibits

quicker learning attributed to its denser reward signal. Also, the curriculum trick is shown to stabilize and accelerate learning for all three methods.

Addressing Generalization

At this point, the current state-of-the-art for RL with unsupervised pre-training comes from algorithms with the objective of behavioural mutual information (BMI) maximisation. This is defined as, the maximisation of the mutual information, independent of reward, between latent variable policies and their state visitation behaviour. These objectives, as used by VIC (Gregor, Rezende and Wierstra, 2016) and DIAYN (Eysenbach et al., 2018b), yield policies which exhibit great diversity in behaviour. However, these methods suffer from poor generalization and a slow inference process when the reward signal is introduced.

Next, the challenges of generalization and slow inference are addressed by using a novel approach with unsupervised pre-training. "Fast Task Inference with Variational Intrinsic Successor Features" (VISR) (Hansen et al., 2020) introduces an algorithm that combines two techniques, BMI maximisation and successor features (SF), to overcome the limitations of each method. Successor features (SF) enable fast transfer learning between tasks that differ only in their reward function, assumed to be linear for some features. Prior to this work, the automatic construction of these reward function features was an open research problem (Barreto et al., 2018).

Continuous Spaces and Multiple Samples

The next advancement is introduced by the "Dynamics-Aware Discovery of Skills" (DADS) Sharma et al. (2020b), an unsupervised RL framework. The primary goal of DADS is to learn low-level skills in continuous spaces autonomously, without the need for supervision, goals, or reward functions, making model-based control more accessible. The acquired skills and their predictive models are learned before the agent is tasked with any specific goal or reward, establishing a repertoire that can be utilised for various downstream tasks.

The main contribution of DADS lies in its mutual-information-based exploration strategy. The proposed objective enables the embedding of learned primitives in continuous spaces, facilitating the learning of a diverse set of skills. DADS also learns to model the dynamics of these skills, enabling the possibility to leverage model-based planning algorithms for downstream tasks. This exploits the adaptability of conventional model predictive control algorithms to plan, solving downstream tasks without additional training.

However, DADS, as well as other reward-free, mutual-information-based learning algorithms, suffers from sample-inefficiency. Thus, an off-policy version of DADS, namely "off-DADS" (Sharma et al., 2020a) is proposed, with the primary objective of enhancing sample efficiency in real-world robotics training. The off-DADS algorithm is designed to efficiently collect data from multiple robots, using its off-policy nature to make training more practical and feasible in real-world robotics. This multi-robot data collection aligns with the goal of enhancing the scalability and applicability of the approach in real-world scenarios.

Combining Methods

The final research explored is presented as a state-of-the-art unsupervised RL method, by a novel approach called "Active Pre-training with Successor Features" (APS) (Liu and Abbeel, 2021a). APS aims to optimize the mutual information between tasks and states during the unsupervised reward-free pre-training phase. The key innovation is specifically the lower bound they've chosen, using APT for the state entropy term and VISR for the state-skill conditional entropy. The proposed APS method combines variational successor features (Hansen et al., 2020) with non-parametric entropy maximisation (Liu and Abbeel, 2021b) to explore the environment effectively. By addressing the limitations of existing mutual information maximisation and entropy maximisation-based unsupervised RL approaches, APS integrates the strengths of both paradigms. It employs non-parametric entropy maximisation for environmental exploration, and the gathered data is leveraged to learn behaviour through variational successor features.

The main contribution of the paper is to address the challenges posed by existing methods, specifically APT (Liu and Abbeel, 2021b), and VISR (Hansen et al., 2020) by combining them in a novel way. APS introduces an alternative direction by maximising mutual information between states and task variables, leveraging state entropy for exploration and conditional entropy to encourage the learning of task-conditioned behaviours. This paper demonstrates state-of-the-art performance on the Atari benchmark.

Recent Innovations

Explore, Discover, Learn

The approach of maximising mutual information between states and latent variables in skill discovery methods often hinges on approximations that use policy-induced distributions. Traditional methods, by focusing on reinforcing behaviours dictated by the initial random policy, inadvertently restrict the breadth of options an agent can explore. This often leads to a reinforcement of already known behaviours rather than a true exploration of new possibilities, as the agent's exploration capabilities are limited by the early stages of a randomly initialized policy.

In response to these limitations, a novel approach named Explore, Discover, and Learn (EDL) (Campos et al., 2020) is proposed to more effectively model unknown distributions and facilitate genuine option discovery. Unlike conventional methods that depend heavily on the state distribution skewed by high-reward states, EDL employs a fixed distribution over states, which remains unaffected by the reward structures. This detachment from reward-focused policy effects is designed to circumvent the pathological learning dynamics in existing methods.

EDL is structured around three distinct phases: exploration, skill discovery, and skill learning. Each phase has its own set of strategies and can be tailored and optimized based on the specific requirements of the task at hand. By compartmentalizing these stages, EDL provides a framework where each component can be independently developed and enhanced. The inclusion of a fixed state distribution is a critical feature that sets EDL apart, offering a more stable and consistent foundation for skill discovery and learning. This approach not only promises a deeper and more systematic exploration of the skill space but also enhances the

agent's ability to generalize these skills across various tasks and environments.

Contrastive Intrinsic Control

More recently, the introduction of Contrastive Intrinsic Control (CIC) (Laskin et al., 2022) represents further evolution in skill discovery within complex environments for RL. Unlike traditional competence-based methods that maximise the mutual information between states and skills, such as DIAYN (Eysenbach et al., 2018b), or knowledge-based methods that rely on maximising predictive model errors like Curiosity (Pathak et al., 2017), CIC introduces a novel approach that addresses the limitations found in environments with large spaces of potential behaviours.

Current competence-based methods often fall short in complex settings where the diversity of potential behaviours exceeds the capacity of conventional discriminators. These methods are constrained because they require an exponentially large number of diverse data samples to maintain accuracy, a condition often not met in highly dynamic and varied environments. CIC addresses this challenge by implementing a new type of discriminator, a contrastive density estimator, designed to approximate the conditional entropy effectively.

This novel discriminator leverages contrastive learning principles, similar to those used in visual representation learning, but uniquely applies them to state transitions and skill vectors in RL contexts. By doing so, CIC can enhance the representation learning capabilities within the domain of unsupervised skill discovery, making it more suited for environments that support a broad spectrum of skills. This approach not only increases the efficacy of learning in large skill spaces but also introduces the potential to bring advanced representation learning techniques from the vision field into RL, thereby broadening the scope and applicability of skill discovery methods in more complex scenarios.

2.5 Hierarchical Reinforcement Learning

Hierarchical Reinforcement Learning (HRL) represents an advanced framework in the domain of artificial intelligence that enhances the traditional RL processes. By adopting a divide-and-conquer strategy, HRL simplifies complex tasks by breaking them down into manageable sub-tasks, each individually solvable and potentially reusable across various problems (Hutsebaut-Buysse, Mets and Latré, 2022). This approach leverages *temporal abstraction*, enabling the system to perform operations across different temporal scales, which enhances decision-making in long-horizon tasks.

Developmental psychology (Spelke and Kinzler, 2006) provides evidence supporting the natural inclination of humans, from infancy through adulthood, to engage in hierarchical problem-solving. This evidence is seen in behaviours ranging from toddlers stacking blocks to adults planning projects, where sub-goals are defined within foundational knowledge domains such as space, obstacles, agents, and social interactions. When these principles of temporal abstraction are integrated into RL, they evolve the process into HRL (Flet-Berliac, 2019), thus extending

RL's capacity for handling complex decision-making processes.

HRL addresses several limitations inherent in traditional RL (Flet-Berliac, 2019):

Sample Efficiency. Traditional RL methods are often criticized for their heavy data requirements. HRL enhances efficiency by allowing learned sub-tasks to be applied across different contexts within the same domain, facilitating transfer learning.

Scalability and Generalization. Classic RL struggles with large action or state spaces due to the curse of dimensionality. HRL counters this by decomposing complex problems into smaller, more tractable sub-problems, thus streamlining the learning process. These sub-problems, by producing not overspecialized skills, help the agent to adapt to new, similar environments, an issue with state-of-the-art RL algorithms.

Abstraction. By employing state and temporal abstractions, HRL simplifies the RL problems, promoting a better representation of knowledge.

Skill discovery methods are used to learn skills of varying granularity to be further used in HRL. A recent contribution by Co-Reyes et al. (2018) introduces a HRL algorithm called SeCTAR (Self-Consistent Trajectory Autoencoder), a practical implementation of this paradigm on continuous control tasks.

SeCTAR

SeCTAR stands out by creating a continuous latent space of skills rather than relying on a discrete set of options or behaviours, coupled with a probabilistic latent variable model that learns to execute and predict the outcomes of these skills. This approach allows a high-level controller to operate within a diverse behaviour space without the constraints imposed by a limited set of predefined actions. The temporally extended nature of these behaviours enables effective temporal abstraction; by using outcome predictions from the learned model, SeCTAR facilitates model-based control at the higher level, merging the benefits of model-free methods (direct interaction with environments through learned behaviours) and model-based strategies (use of predicted outcomes for planning).

This method utilises a trajectory-level Variational Autoencoder (VAE) by Kingma and Welling (2013), as it supports the simultaneous acquisition and prediction of skill outcomes. This dual capability circumvents the need for conventional RL at the higher level, as well as addresses one of the major shortcomings of model-based RL: The difficulty of accurately predicting low-level events at a granular temporal resolution.

This section outlined the research domain of the project, explaining methods and concepts foundational to RL, skill discovery, DIAYN, HRL, as well as further research methods in each of these domains. The following section delves into our methodology for this investigation.

Chapter 3

Methodology

Bipedal Walker. In our proposed paradigm, Bipedal Walker is used as the environment for our analysis. The basic version is used to determine baselines and run ablation tests, while the hardcore version is used for testing the limits of our methods and evaluating their ability to learn skills that transfer to harder environments.

3.1 Proximal Policy Optimization

Our implementation of PPO adheres to the specifications outlined by Huang et al. (2022a) and Huang et al. (2022b), incorporating 13 core and 9 continuous action domain-specific details. This setup achieves performance on par with the established Raffin et al. (2021) baseline. Key modifications and optimizations include:

Entropy Bonus. PPO's loss function is designed as defined by Schulman et al. (2017):

$$L_t(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)], \tag{3.1}$$

where $L_t^{CLIP}(\theta)$, $L_t^{VF}(\theta)$ are the clipped actor loss and critic loss, respectively; c_1, c_2 are coefficients; and S denotes an entropy bonus.

While entropy is intended to enhance exploration by promoting a more uniform action distribution, Raffin et al. (2021) determined that for continuous control environments like Bipedal Walker, the entropy term does not significantly impact performance, and thus, the entropy coefficient is initially set to zero. However, Eysenbach et al. (2018b) change this to increase exploration of diverse skills with DIAYN.

Reward Normalization. Although discount-based scaling of rewards is common, it proved ineffective with the Bipedal Walker's dynamics as per Raffin et al. (2021), and was omitted from our implementation.

Reward Clipping. The scaled rewards were originally clipped between $[-10, 10]$, however, this range is environment depended, and through experimentation was found to be optimal at $[-400, 400]$ for Bipedal Walker. This improves the algorithm's stability.

Learning Rate Annealing. A linear decay of the learning rate over all training steps has shown to improve the episodic returns significantly, indicating more effective learning

earlier in the training process. However, this may prove to have effects on DIAYN, thus an ablation study on this implementation in PPO's optimizer was performed to determine its effect on skill acquisition.

3.2 Learning Skills with DIAYN

The DIAYN framework defines a skill as a latent-conditional policy $\pi(a|s,z)$ that consistently alters the state of the environment. Here, z is a latent variable sampled from a prior distribution $p(z)$, embodying different potential skills. The formulation of DIAYN focuses on three main objectives (Chen, 2019):

State-Skill Dependency. Maximising the mutual information between the states visited S and the skills Z ensures that different skills lead to distinguishable state visitation patterns.

Action Irrelevance. Minimizing the mutual information between actions A and skills Z given the state S, ensuring that states are used to distinguish skills, not actions.

Exploration Maximisation. Encouraging skills to be as diverse as possible by maximising the entropy $H(A|S)$ of the collective policy distribution over all skills, thus developing exploratory behaviour.

The combined objective is expressed as:

$$\begin{aligned}\mathcal{F}(\theta) &:= I(S;Z) + [A \mid S] - I(A;Z \mid S) \\ &= (H[Z] - H[Z \mid S]) + H[A \mid S] - (H[A \mid S] - H[A \mid S,Z]) \\ &= H[Z] - H[Z \mid S] + H[A \mid S,Z] \end{aligned} \tag{3.2}$$

Maximising $H[Z]$ encourages our prior distribution $p(z)$ to have high entropy, minimizing $H[Z \mid S]$ suggests that it should be easy to infer the skill from the current state, and maximising $H[A \mid S,Z]$ indicates that each skill should act as randomly as possible. We maximise $H[A \mid S,Z]$ using PPO, and entropy bonus with coefficient $c_2 = 0.1$ (Equation 3.1).

We approximate this objective using a variational lower bound (Eysenbach et al., 2018b; Barber and Agakov, 2003), a defining characteristic of this family of methods:

$$\begin{aligned}\mathcal{F}(\theta) &= H[A \mid S,Z] - H[Z \mid S] + H[Z] \\ &= H[A \mid S,Z] + \mathbb{E}_{z\sim p(z),s\sim\pi(z)}[\log p(z \mid s)] - \mathbb{E}_{z\sim p(z)}[\log p(z)] \\ &\geq H[A \mid S,Z] + \mathbb{E}_{z\sim p(z),s\sim\pi(z)}[\log q_\phi(z \mid s) - \log p(z)] := \mathcal{G}(\theta,\phi) \end{aligned} \tag{3.3}$$

For a fixed set of skills $Z \sim p(z)$, we set the prior $p(z)$ to be a *discrete uniform distribution*, guaranteeing that it has maximum entropy. The constant $\log p(z)$ in the reward function helps encourage the agent to stay alive, removing $\log p(z)$ results in negative rewards, which tempts the agent to end the episode as quickly as possible. Finally, it is intractable to integrate over all states and skills to compute $p(z \mid s)$ exactly, so this posterior is approximated with a learned discriminator $q_\phi(z \mid s)$, serving as a variational lower bound $\mathcal{G}(\theta,\phi)$ on our objective $\mathcal{F}(\theta)$. This leads to the intrinsic pseudo-reward $r_z(s,a)$ passed to the PPO agent at each time-step.

$$r_z(s, a) := \log q_\phi(z \mid s) - \log p(z)$$

This reward formulation incentivizes the agent to maintain skill diversity and avoid premature termination of episodes.

DIAYN's implementation details, combined with inspiration from Alirezakazemipour (n.d.) and Huang et al. (2022a), were executed as described in the provided pseudocode, Algorithm 3.2. Several experiments were conducted to validate this framework.

Algorithm 3.2: PPO based DIAYN

while *not converged* **do**
  Sample skill $z \sim p(z)$ and initial state $s_0 \sim p_0(s)$
  for $t \leftarrow 1$ to *steps_per_episode* **do**
    Sample action $a_t \sim \pi_\theta(a_t|s_t, z)$ from skill
    Step environment: $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    Compute $q_\phi(z|s_{t+1})$ with discriminator
    Set skill reward $r_t = \log q_\phi(z|s_{t+1}) - \log p(z)$
    Update policy (θ) to maximise r_t with PPO
    Update discriminator (ϕ) with Gradient-descent
  end for
end while

3.3 Experiments

PPO Test. A base implementation of PPO is tested on the basic version of Bipedal Walker.

PPO Limitation Test. The same PPO agent is further tested on the hardcore version of Bipedal Walker

Novel DIAYN Method Test. A novel implementation of PPO based DIAYN is evaluated on Bipedal Walker by an analysis of performance graphs, agent trajectory snapshots, and PCA (Jolliffe and Cadima, 2016) for dimensionality reduction to facilitate visualization.

DIAYN Skill Variety Test. We evaluated the impact of varying the number of skills (5, 10, and 20) on the skills themselves, using performance graphs and PCA.

DIAYN Learning Rate Annealing Ablation. An ablation study was performed for annealing the learning rate in PPO's optimizer.

PPO Environment Generalization. Skills were trained in the basic Bipedal Walker environment and subsequently in the hardcore version to explore their transferability and ability to generalize.

DIAYN Time-step Length Variety Test. To investigate the implications on the skills acquired by the length of training, DIAYN is trained for 40 million time-steps, in contrast to all the other experiments that train for a mere 15 million time-steps.

Fine-Tuning. Further experimentation and fine-tuning was done on the hyperparameters of our model, helping to empirically converge the hyperparameters to optimal values for

the experiments, as described in Appendix B.

Hierarchical Integration of Learned Skills. The ultimate goal was to integrate the skills learned via DIAYN into a hierarchical PPO framework. These skills can be used in various ways, foundational to HRL (Chapter 2), such as for facilitation of priors, action space simplification, or to be used by a higher-level controller that selects appropriate skills at fixed intervals during task execution. However, this part of the project remains for future research due to time constraints.

The outcomes of these approaches and their implications are detailed in the following sections, presenting a comprehensive analysis of the experimental results.

Chapter 4

Results and Analysis

4.1 Performance of PPO in Basic and Hardcore Environments

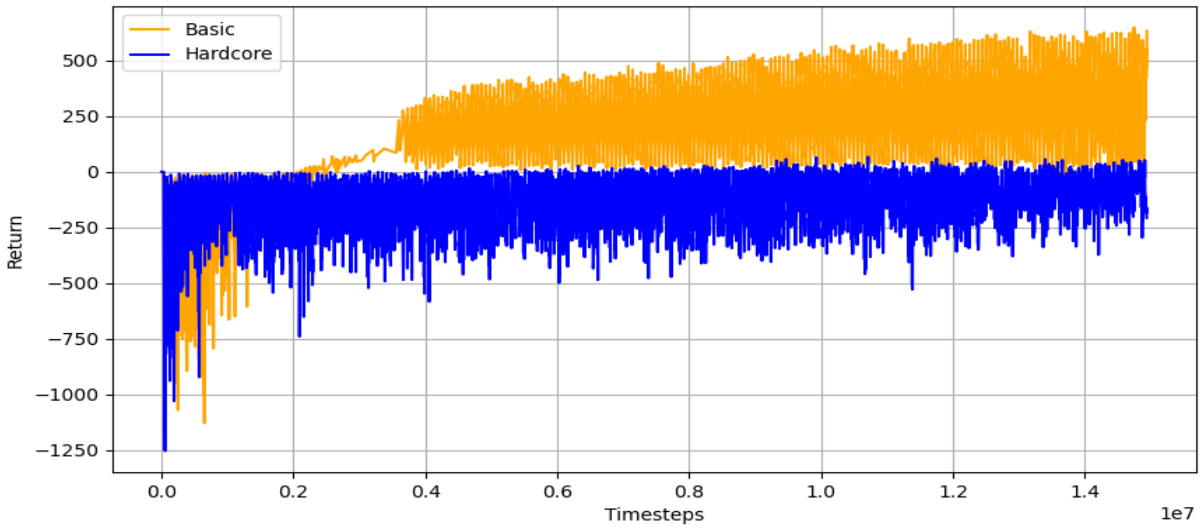

Figure 4.1: Return over 15 million time-steps for PPO on Bipedal Walker basic (Orange) and hardcore (Blue) environments.

Our implementation of PPO is extensively evaluated within two distinct settings of the Bipedal Walker environment: the basic and the hardcore versions. The episodic returns of external rewards from these experiments are depicted in Figure 4.1, where the performance of PPO is represented in two colour-coded trends: orange for the basic environment and blue for the hardcore environment.

In the basic environment, as shown by the orange line in Figure 4.1, the PPO agent demonstrates a robust ability to navigate and complete the terrain. Notably, the agent consistently achieves scores exceeding 300 points, suggesting a successful learning outcome. The episodic length decreases from approximately 32 to 20 seconds over training iterations, and the scores progressively increase, reaching upwards of 600 points. This improvement indicates not only a mastery of the task but also an optimization in speed, highlighting the agent's potential to

further excel, possibly by training for additional timesteps (e.g., up to 40 million timesteps) to enhance its sprinting capability. However, this potential is inherently bounded by the physical constraints within the environment. The initial phase of the learning curve is marked by exploration, with the agent experimenting with different strategies before and stabilizing on a successful trajectory, as illustrated in the early timesteps of the graph. This phase transitions into a more stable learning period where the agent refines its approach to optimize performance.

Contrastingly, in the hardcore environment, depicted by the blue line in Figure 4.1, the agent struggles significantly. It fails to achieve consistently positive rewards and shows no progression towards overcoming this challenge throughout the training period. Video recordings of the agent's interactions reveal that it frequently encounters obstacles such as blocks, stairs, and pits, which not only impede its progress but often lead to termination of the episode. The agent's inability to recognize or adapt to these obstacles, compounded by the complex and varied impacts they have on performance, underscores the hardcore environment's difficulty and the limitations of the current PPO implementation in managing such challenges. ppo (n.d.) show better results on the hardcore environment after training for 100,000,000 timesteps, suggesting that algorithmic improvements and adjustments to hyperparameters may allow PPO to solve the environment.

In conclusion, while the PPO algorithm excels in the structured and predictable basic environment of the Bipedal Walker, achieving high scores and demonstrating efficient learning, it encounters significant barriers in the more unpredictable and complex hardcore setting. These findings suggest that while PPO can serve as a solid baseline, there is substantial room for enhancement, particularly through the integration of hierarchical or skill-based methods such as DIAYN, which could potentially equip the agent with the tools needed to tackle the diverse and dynamic challenges observed in the hardcore environment. This sets a clear direction for future improvements and methodological refinements.

4.2 Evaluating DIAYN Across Different Configurations

4.2.1 Overview of DIAYN Models

In this section, we delve into the performance of our novel DIAYN implementation modified for use with PPO. The original research (Eysenbach et al., 2018b) produces analysis that is mostly inapplicable to our environment, as for positional states, the method directly partitions the environment by diversifying state visitation, whereas for our partially observably Bipedal Walker, the state-space does not relate to the positional scope of the environment. This opens a new avenue for analysis of DIAYN on locomotive POMDPs.

We compare four distinct configurations of the DIAYN model to assess how varying the number of skills and the learning rate decay impacts performance. These configurations are designated as follows for clarity and consistency throughout our analysis:

DIAYN_PPO. The control experiment, which uses the base version of our DIAYN implementation featuring 20 skills with learning rate annealing and an entropy coefficient of

0.1, as described by Table B.2.

DIAYN_PPO_S5. A variant of the base model trained with only 5 skills.

DIAYN_PPO_S10. Similar to DIAYN_PPO, but with 10 skills.

DIAYN_PPO_ABL. The base model without learning rate decay, serving as an ablation study for this feature. This mirrors the original method by Eysenbach et al. (2018b).

4.2.2 Advanced Skill Analysis

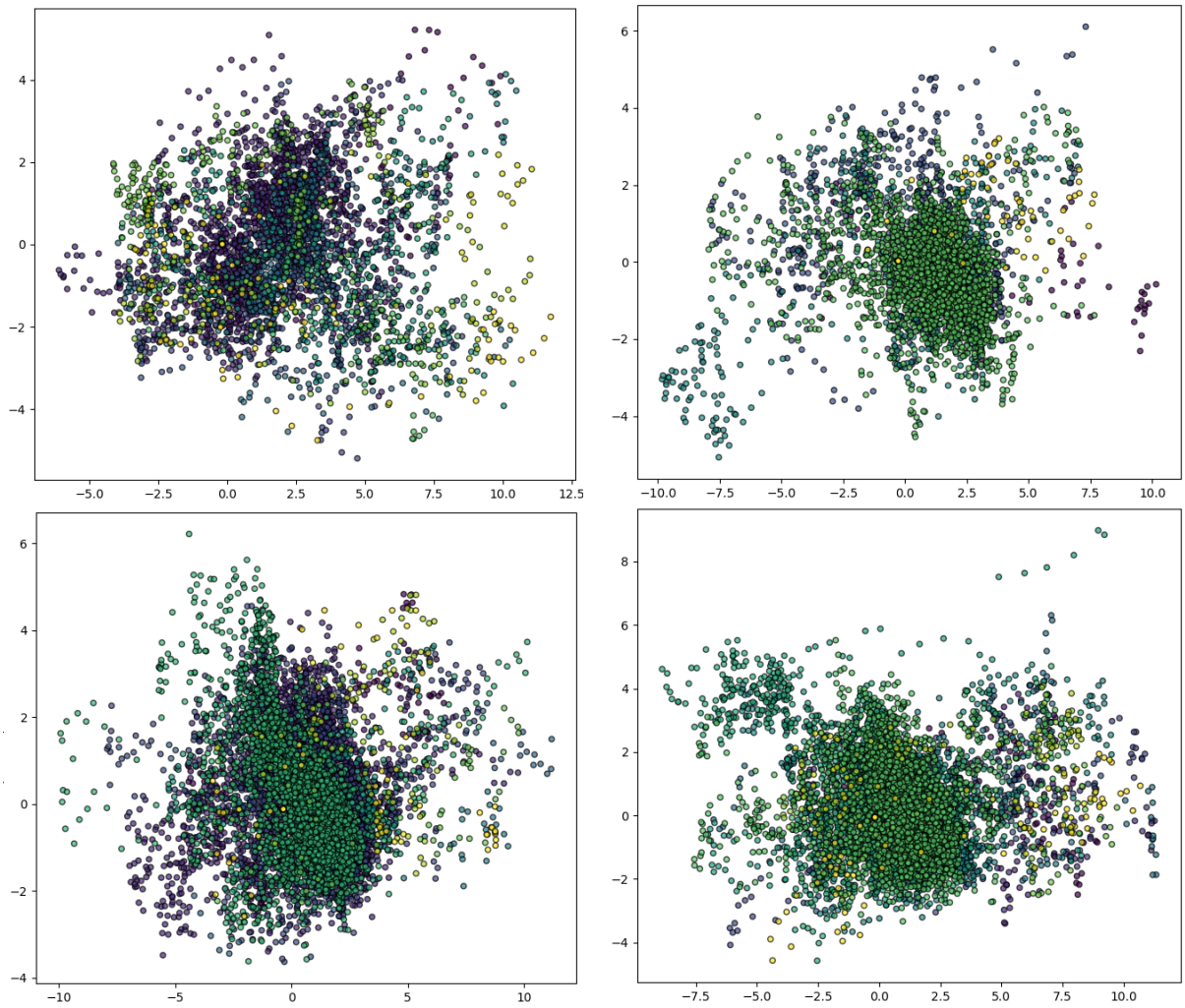

Figure 4.2: PCA visualizations of 2,500 observed states based on skills for DIAYN_PPO (Top-left), DIAYN_PPO_S5 (Top-right), DIAYN_PPO_S10 (Bottom-left), DIAYN_PPO_ABL (Bottom-right). Each skill trajectory is represented by a group of data points of a specific colour.

In this analysis, we apply PCA to visualize the state visitation patterns of each skill developed through the DIAYN framework, as depicted in Figure 4.2. These visualizations are pivotal for understanding how different skills explore the state space of the Bipedal Walker environment.

Central Common and Peripheral Rare Regions. Each graph in Figure 4.2 highlights two main areas: a dense central region and sparser peripheral regions. The dense central area, or the "common region," represents states where the agent maintains balance and stability, which are crucial for prolonged survival in the environment. These correspond to states that achieve low intrinsic reward in Figure C.1 and result in the discriminator performing worse in Figure 4.5. In contrast, the peripheral "rare region" includes states that lead to unstable positions, potentially causing the agent to fall and terminate the episode. These regions underscore the dual nature of the agent's latent exploration: maintaining commonality in indiscriminate states, while venturing into rarer, discriminable states that offer better diversity. These behaviours, supported by corresponding figures and videos captured during the experiments, are a direct result of the state-space of Bipedal Walker being used to discriminate skills. The objective of DIAYN is to best partition the state-space based on skills, this clearly shows an inherent flaw with DIAYN, as the skills do not confidently partition the semantic state-space.

Skill Trajectories and Their Limitations. Analysing the skill colouration in Figure 4.2 reveals that skills start from the common region and progressively explore outward into the rare zones. This exploration pattern is indicative of the agent testing the limits of its diversity and discovering new behavioural niches within the environment. By cross-referencing the agent videos and their semantic representation, it can be seen that diversity of a skill is directly correlated to the instability of the agent in the resultant states. In this case, specific trajectories, such as those leading to the top right corner in the bottom-right graph or the right edge in the top-right graph, correspond to specialized skills that have evolved to unstable positions. As in Eysenbach et al. (2018b); Alirezakazemipour (n.d.), the learned skills achieved static poses. These, while useful for analysis, serve no purpose for the objective of DIAYN in Bipedal Walker; the skills learned cannot be used to navigate the environment. However, these types of skills could be useful in environments where static poses are the objective, such as exploring all possible single positions available to an agent in a partially observable environment, i.e. the agents' range of motion. Newer methods, explored in Chapter 2, fix this issue with DIAYN using various strategies. Given more time, this research would be extended to learn from observation trajectories with the following hypothesis: "If learning from individual states develops static poses, then learning from a set of consecutive states should develop dynamic trajectories, considered to be complex movement. These should better incorporate the area of the environment, and partition the state-space.

Visualization Improvement. Generally these illustrations provide some basis for semantic reasoning, however, leave much room to be improved to better abstract the observations recorded. To further separate skills in semantic space, some consistent or performance dependent bias could be introduced. Once such case could calculate an average trajectory of for states of the same skill, then transform all corresponding states toward this function by some consistent bias. Alternatively, for a more interpretable semantic space, a higher dimensional space could be used, where each axis corresponds to some hand-picked set of observation parameters, such as one set of all motor torques, and one of all lidar readings. This leaves potential for deeper, more interpretable trajectory analysis.

Semantic State Space Interpretation. The semantic division of the state space, illustrated in Figure 4.3, provides a structured way to interpret these explorations. The left half of the

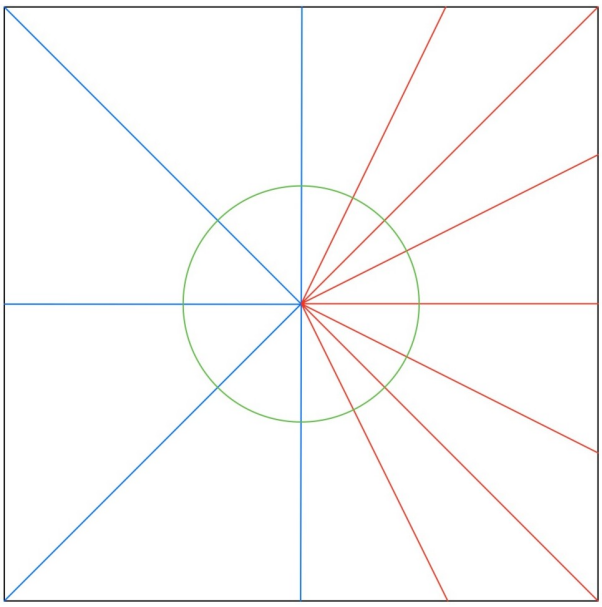

Figure 4.3: Proposed region split for PCA semantic state space.

PCA plot is divided into segments that represent different fundamental postures or actions, such as standing on one leg or transitioning between various dynamic movements. As skills increase from 8 to 16, we see a finer segmentation of the space, suggesting that each skill becomes more specialized, refining its ability to consistently exploit a specific set of states, that have further implications for task-specific purposes. This presents a semantic interpretation of the number of skills trained: "The more skills learned, the higher granularity skills emerge." So to learn more sophisticated skills, more skills must be learned. On the other hand, the observed states are less concentrated, with more skills, showing the agent to learn skills that although more sophisticated, are less discriminable from each other, thus utilise many similar physical postures. This supports the corresponding episodic returns and discriminator losses in Figures C.1 and 4.5, respectively.

Skill Specificity and Learning Rate Effects. The progression from fewer to more skills shows no inherent pattern; intuitively, more skills should partition the semantic state-space with higher granularity, however, this is not shown. Contrastingly, Figure 4.5 shows that fewer skills are more discriminable, an inherent characteristic where fewer skills overlap fewer times over a finite set of states. This shows clear issues with DIAYN in this environment. Interestingly, the learning rate decay does not visibly affect the distribution of states explored, which aligns with findings from discriminator performance in Figure 4.5, suggesting that the fundamental character of skills, once developed, remains robust to changes in learning dynamics.

Method Validation. It can be seen on each graph in Figure 4.2 that each skill starts in the common region, explores some distinct subsection of this region, and further explores the rare region in some specific direction. Coupled with the emergent static poses that the agents learn, this validates our implementation of DIAYN_PPO; more immediately visible in Figure 4.8, as double the number of visited states by each skill are recorded.

The trajectory snapshots of the agent (Figure 4.4), captured from the beginning of each

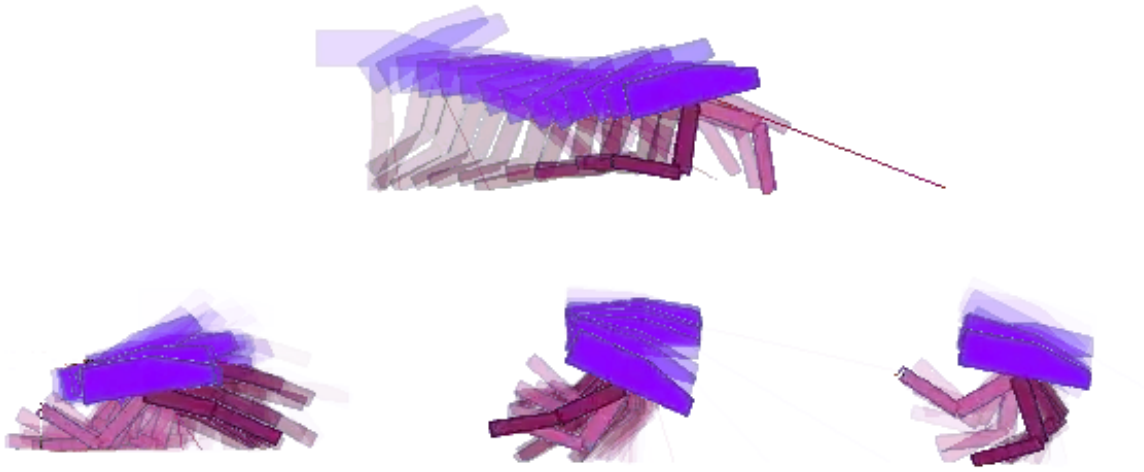

Figure 4.4: Captured agent trajectories for PPO (Top) and 3 different skills by DIAYN_PPO (Bottom) on the basic Bipedal Walker environment.

run, further validate the PCA findings. The top snapshot shows a PPO agent purposefully running forward toward the goal, utilizing skills honed for speed and efficiency, these manifest by increasing power and stride length, respectively.

The bottom snapshots depict various DIAYN skills in action, highlighting diverse strategic postures such as balancing, leaping, and cautious manoeuvring. The left agent balances, resembling the PPO agent above, except showing no intention of forward progression. The middle agent leaps forward, demonstrating its emergent behaviour; Similarly, the right agent resembles the PPO agent, expressing its own emergent behaviour: "balancing on the front leg." Each skill illustrates different angular positions of the leg joints, head, and lidar sensors. The consistent characteristic for all DIAYN_PPO skills is the emergent "static" pose, where the agent refines and stabilizes some physical position; this is expected by the intrinsic nature of the agent's feedback loop, as skills are distinguished by singular states and no external rewards or environmental coordinates are used by the agents, thereby removing the incentive for traversing the environment. This validates our implementation of DIAYN, supports our interpretation of the semantic state space shown in Figures 4.2 and 4.8, as described by Figure 4.3, and shows the inherent limitation of DIAYN in Bipedal Walker.

4.2.3 Discriminator Performance

Discriminator efficacy, as depicted in Figure 4.5, directly shows how well the DIAYN model can identify skills from state observations. The graph shows negative losses, where values closer to zero, or higher on the graph, indicate better discriminator performance.

A clear trend emerges: fewer skills result in more consistent and stable discriminator performance, underscoring the challenges of distinguishing a larger number of skills within the same action and state spaces. This is outlined by the range of values in Figure 4.5 for each model configuration, mirroring this latent behaviour by Figure C.1 in Appendix C. Furthermore, the

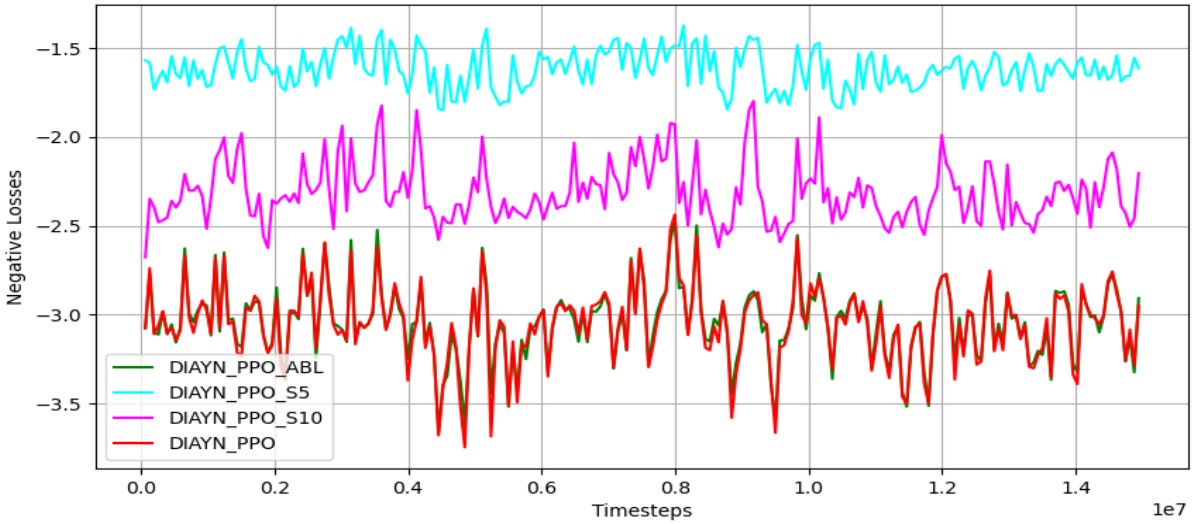

Figure 4.5: Negative losses over 15 million-time-steps of the skill discriminators for DIAYN_PPO (Red), DIAYN_PPO_S5 (Blue), DIAYN_PPO_S10 (Magenta), DIAYN_PPO_ABL (Green).

absolute performance shown by Figure 4.5, outlines that the discriminator performs much better on all skills for fewer skills, suggesting that fewer skills can spread further from each other over the state-space. This is interpreted as the agent visiting fewer overlapping states. With a finite set of "stay alive" states, fewer skills can more easily partition these states for maximum diversity.

The stability of the discriminator is not affected by the learning rate decay (DIAYN_PPO_ABL), suggesting that learning at a constant rate without decaying does not enhance the ability of the discriminator to differentiate between skills. This observation aligns with theoretical expectations that skill discriminability should primarily depend on the diversity of state visitations rather than the specifics of the learning algorithm. Continual learning doesn't affect the static position the agent learns; this is due to the inherent flaw with DIAYN, where using states to discriminates skills limits the learned skills to static poses in partially observable environments.

4.2.4 Policy Loss and Action Entropy

The relationship between policy loss, action entropy, and the learning rate decay is explored through Figure 4.6.

Maintaining a constant learning rate, as expected, is shown to maximise action entropy S in Equation 3.1 by (Figure 4.6, left), and consequently by the policy loss in (Figure 4.6, right). These show that the PPO agent functions properly and learns refined policies for each skill, this affirms that the issue with this method is the objective that the PPO is given (Equation 3.2). Despite this, the figures show that when fewer skills are trained, the agent displays a marginal increase in decision-making confidence, evident by slightly lower action entropy and improved policy performance. This finding suggests that a condensed skill set may allow for more focused learning and optimization on those specific skills, potentially enhancing overall agent efficiency and effectiveness in the skills it learns. The skills themselves may not be useful, however, this characteristic is expected to extend to useful skills.

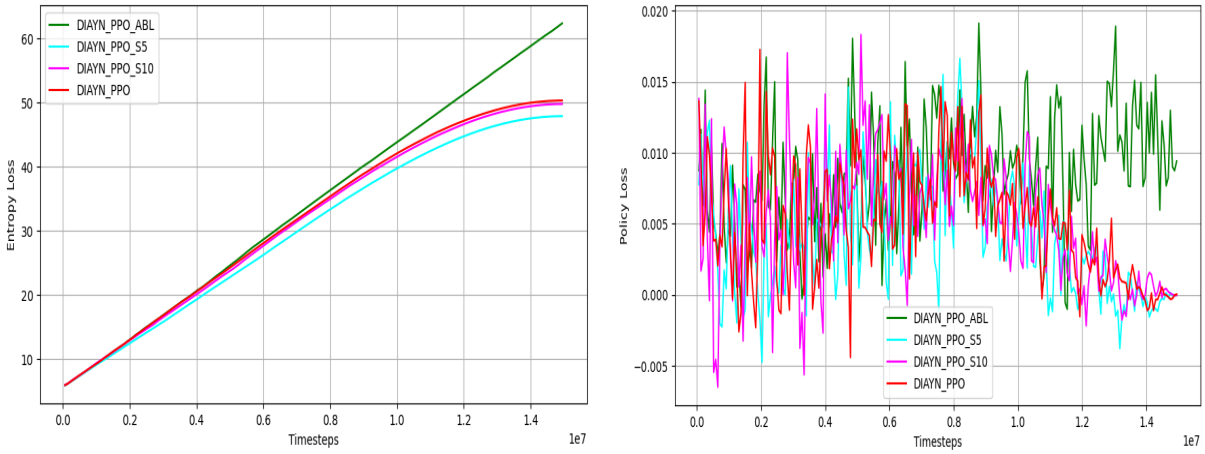

Figure 4.6: PPO's action entropy (Left) and policy loss (Right) over 15 million time-steps for DIAYN_PPO (Red), DIAYN_PPO_S5 (Blue), DIAYN_PPO_S10 (Magenta), DIAYN_PPO_ABL (Green).

4.3 Extended Training

Extended training sessions provide deeper insights into the skill maturation process over longer time frames, as shown in Figures 4.7 and 4.8.

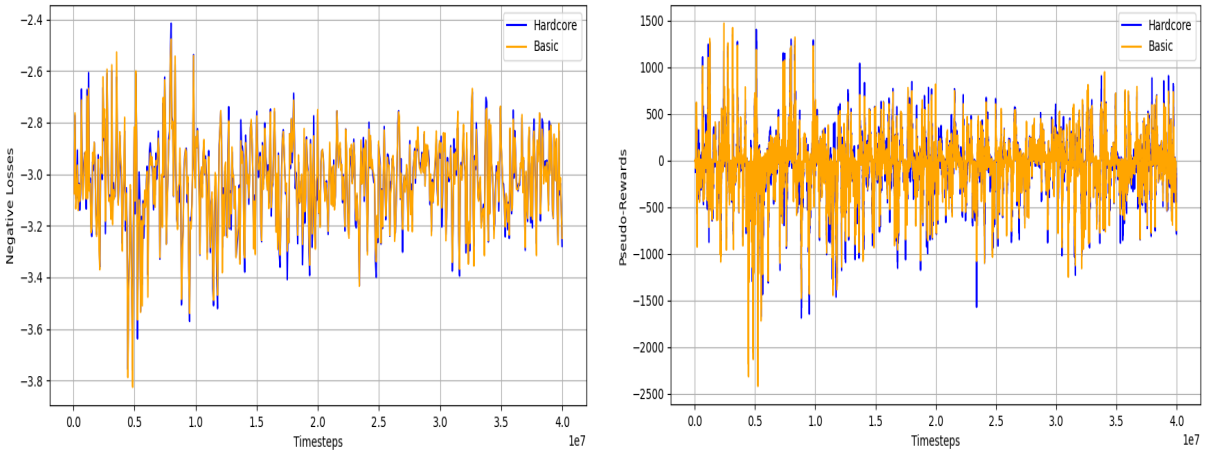

Figure 4.7: Negative losses of the skill discriminators (Left) and intrinsic rewards (Right) over 40 million time-steps, for DIAYN_PPO on the Bipedal Walker basic (Orange) and hardcore (Blue) environments.

Discriminator Stability and Skill Refinement. Over 40 million timesteps, the discriminator's performance stabilizes (Figure 4.7, left), indicating that the skills have become well-defined and are consistently recognizable by the model. The episodic returns (Figure 4.7, right) mimic patterns from shorter training regimes, suggesting a plateau in learning efficacy and an optimal level of skill differentiation. This inspires the following deduction: "DIAYN_PPO has the ability to discover a diverse set of physical postures in Bipedal Walker with little training." This is promising for the applicability of a pre-learned set of skills by DIAYN_PPO for downstream

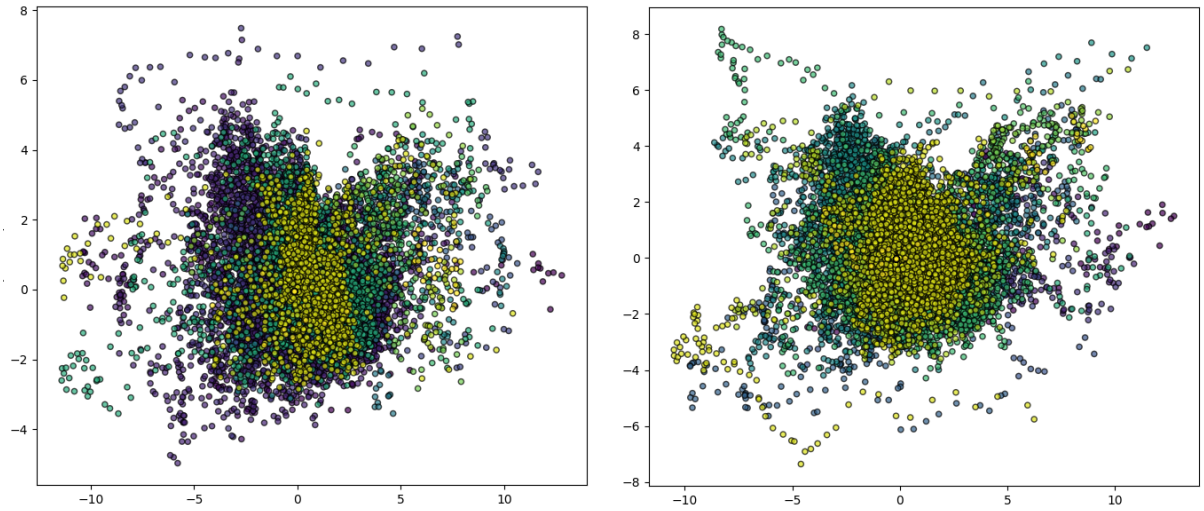

Figure 4.8: PCA visualizations of 5000 observed states based on skills for DIAYN_PPO after 40 million training time-steps on the Bipedal Walker basic (Left) and hardcore (Right) environments. Each skill trajectory is represented by a group of data points of a specific colour.

posing tasks, with improved sample efficiency.

Generalization to Hardcore Environment. Sometimes the agent will experience some effect of the obstacles in the hardcore environment, this will either obstruct the agent's movement, or be detected by its lidar sensors. The agent would sometimes terminate early due to an unexpected collision with an obstacle at the beginning of the run, this explains why Figure (4.7, right) shows some lower rewards for the hardcore environment in some cases. On the other hand, for most of the training time, both graphs affirm that the discriminator performs comparably on both environments. This suggests that the agent manages to learn the similar skills whether faced with obstacles or not, when trained long enough, implying that the agent learns skills based on the physical positions it manages to achieve, with little to no effect by external factors in the environment. This demonstrates that to further test DIAYN_PPO on different environments, it may not be necessary to simplify the environment (e.g. remove obstacles) to learn a semantically similar set of skills. The agent could learn directly on the complex environment, indicating robust skill transferability and generalization. This has great implications for future experiments being cheaper to run in terms of time, space, environmental resources, and compute power.

PCA Enhancements Post-Extended Training. The PCA plots with 5000 observed states (Figure 4.8) post-extended training exhibit clearer and more distinct clusters. In the basic environment, skills create "egg-shaped" clusters that venture from common to rare zones, illustrating a sophisticated navigation of the state space. This further affirms our analysis from Figure 4.2 and the applicability of DIAYN_PPO on complex environments as shown by the clear diversity of latent state visitation in Figure (4.8, right), reaching different corners of semantic space. Improvements to the objective function would better partition this space, and a more informative semantic state-space could be used to better visualize the distribution of visited states.

Overall, the analysis undertaken in this section validates our implementation of DIAYN_PPO.

The limitations of our method are shown to originate from the objective function (Equation 3.2, using states to discriminate between skills with no incentive to move forward, which leads to the agent learning static poses, thus not advancing through the terrain of the environment. This is further validated by the analysis of intrinsic rewards in Appendix C. Similar analysis and conclusions have been reached by more recent works, as explored in Chapter 2 (Achiam et al., 2018; Sharma et al., 2020b; Laskin et al., 2022; Campos et al., 2020). This has implications for this class of methods on partially observable environments, even with more sophisticated objectives. Deeper analysis of the skills learned in Bipedal Walker is demonstrated through the use of PCA, video capturing and trajectory snapshots, and different hyperparameter tests. The conclusions and implications of this analysis are explored in the next section.

Chapter 5

Conclusions

5.1 Review of Findings

This research has methodically explored the performance of PPO and DIAYN within simple and complex (hardcore) Bipedal Walker environments. Our findings present a deep understanding of both algorithms' capabilities and limitations, as reviewed below.

PPO Performance. PPO demonstrated robust performance in the basic environment, efficiently navigating and completing the terrain with increasing scores and decreasing episode lengths. This success shows PPO's capability in environments with predictable dynamics and clear objectives. In contrast, PPO struggled significantly in the hardcore environment, frequently failing to navigate obstacles effectively. This limitation was primarily due to the algorithm's inability to adapt its policy in response to new and unexpected challenges, improvements have been explored with modifications to the original method, as well as other actor-critic methods (Kingma and Welling, 2013).

DIAYN Enhancement. DIAYN_PPO showed expected behaviour as demonstrated by Alirezakazemipour (n.d.), and clearly demonstrated the limitation of the method in partially observable Bipedal Walker. Static poses were produced as a result of discriminating skills by states that provide no incentive to navigate the environment, thereby highlighting the limitation of DIAYN_PPO in POMDPs; this is illustrated by our analysis of the discriminator, agent state visitation (Chapter **??**), and intrinsic rewards (Appendix C).

When varying the number of skills learned, more skills showed finer granularity, but also resulted in higher performance volatility. The learning rate decay in DIAYN did not visibly impact skill discriminability, affirming that the intrinsic properties of skill learning are robust to certain hyperparameter modifications of the base RL algorithm. PCA revealed distinct clusters of skill utilization, illustrating how skills developed by DIAYN_PPO navigate between common and rare states (physical postures) in POMDPs. This state-space interpretation (Figure 4.3) serves as a promising framework to further refine and develop skills for agents navigating partially observable environments, for more advanced methods (Co-Reyes et al., 2018; Laskin et al., 2022; Campos et al., 2020; Sharma et al., 2020b).

5.2 Practical Implications

The practical implications of these findings are promising for the development of autonomous systems and the application of RL in both simulated and real-world scenarios. We demonstrate the limitations of DIAYN in POMDPs, and offer a baseline analysis for comparison with more advanced methods. The significance, is extending this class of methods to partially observable paradigms, effectively, simulating more realistic environments. By solving the limitations of our method, as achieved by a myriad of studies (Co-Reyes et al., 2018; Laskin et al., 2022; Campos et al., 2020; Sharma et al., 2020b), the implication of this area of research is vast.

Enhanced Navigation in Robotics. The ability to generate diverse skills that can be further used for effective navigation in complex environments suggests utility in robotics, particularly for tasks that require adaptive behaviour in dynamic and unpredictable settings, such as search-and-rescue missions or exploratory robots in hazardous environments. The efficacy of DIAYN in partially observable environments opens an avenue for more complex, realistic environments for the application of DIAYN.

Sample Efficiency. The findings that extended training does not necessarily improve skill discriminability imply that RL models can be designed to achieve optimal performance with shorter training cycles. This can significantly reduce computational costs and speed up the development cycle of autonomous systems.

Robustness to Environmental Changes. The transferability of skills learned in one environment to another, especially from basic to hardcore scenarios, demonstrates agents' robustness, crucial for developing systems that must operate under varied and unpredictable conditions without needing retraining from scratch.

AI Safety and Ethics. As RL applications grow, especially in contexts involving interaction with humans (e.g., autonomous vehicles and healthcare), the ability to predict and control the behaviour of AI systems through approaches like DIAYN becomes increasingly important. Ensuring these systems can adapt safely to new environments without catastrophic failures is fundamental to their widespread use. The results from this study demonstrate the ability of autonomous agents to learn very complex, and possibly dangerous skills such as "kicking", and also open up new avenues for applying these skills in practical applications. If used in real-world scenarios, a robust framework for controlling and monitoring the agents is essential.

5.3 Future Research

This study serves as a solid bedrock for future research. Our semantic state-space representation and skill analysis provides a baseline for more advanced methods to be tested in partially observable paradigms. This opens many avenues for further investigation, some suggestions are explored.

5.3.1 Experiment Improvement

Introducing some bias, such as a consistent transformation towards a moving average of semantic states for each skill, has potential to improve the segmentation of semantic state space. Alternatively, introducing a hand-picked structure for dimensionality reduction, so that each axis relates to some interpretable set of observations, such as all motor torques or lidar readings, can help better interpret the latent skills learned by the agent.

5.3.2 Hierarchical Integration

Action Space Simplification. The action space of the environment can be replaced for a PPO agent with a set of pre-learned skills by DIAYN_PPO.

Skill Priors. DIAYN_PPO agent is trained to pre-learn a set of skills, the weights of the agent's actor and critic networks are used as priors for a different PPO agent on a set of downstream tasks.

Meta-Controller. A meta-controller is trained to learn which skill to be used for the next k (consistent) number of timesteps, after pre-learning a set of skills, as described by Eysenbach et al. (2018a). Bayesian hyperparameter tuning can be used to optimize the frequency the agent changes skill. Another approach could be to change the skill when the agent reaches a state within some accepted subsection of the "common" region of observations, after having previously left this region; this can be achieved with a simple boolean control parameter. A further improvement to the meta-controller could be to train another network to learn the hyperparameter k, to chose which skill to use and how long to use it for.

5.3.3 Algorithmic Improvements

Learning Skill Granularity. The discriminator network could be extended to learn the optimal number of skills to learn in order to best partition the state-space. This would allow for the optimization of skill granularity for each environment.

Learning Skills From Trajectories. Further implement PPO with DADS (Sharma et al., 2020b,a) by learning skills based on state trajectories instead of individual observations, test on a partially observable environment, and compare the analysis with that of DIAYN_PPO.

5.3.4 Environments

Bipedal Walker Hardcore. For any further experimentation of DIAYN or DIAYN_PPO on Bipedal Walker, it is suggested to use the hardcore version directly.

More Complex Environment. Implement DIAYN_PPO on a more complex locomotive environment, such as increasing the degrees of freedom of the agent in its physical space, or introducing harsher dynamics. These will likely need to be custom environments, as these

requirements exceed the complexity of many widely used baselines.

5.4 Conclusion

The analysis in this research defines a stepping-stone for more advanced methods to operate in partially observable paradigms, an important avenue for autonomous system engineering in real-world environments. The limitations on Bipedal Walker and analysis of the skills learned, provide a baseline interpretation of agent's behaviour to compare with newer methods in this paradigm. As we continue to push the boundaries of what these algorithms can achieve, it is imperative to consider their broader implications, including the ethical and safety considerations that come with deploying AI systems in real-world settings.

In conclusion, this dissertation not only advances our understanding of complex RL methods but also sets the stage for future innovations in AI that are adaptable, efficient, and more adept to the dynamic nature of real-world scenarios. The journey of AI towards achieving true autonomy and reliability is ongoing, and the insights gained from this study contribute a step toward better understanding this challenging frontier.

Number of words until this point, excluding front matter: 10,963.

Bibliography

n.d. Available from: https://huggingface.co/sb3/ppo-BipedalWalkerHardcore-v3.

Achiam, J., Edwards, H., Amodei, D. and Abbeel, P., 2018. Variational option discovery algorithms. 1807.10299.

Alirezakazemipour, n.d. diayn-pytorch: Diversity is all you need: Learning skills without a reward function in pytorch. Available from: https://github.com/alirezakazemipour/DIAYN-PyTorch [Accessed 2024-04-08].

Barber, D. and Agakov, F., 2003. The im algorithm: a variational approach to information maximization. *Proceedings of the 16th international conference on neural information processing systems*. Cambridge, MA, USA: MIT Press, NIPS'03, p.201–208.

Barreto, A., Dabney, W., Munos, R., Hunt, J.J., Schaul, T., Hasselt, H. van and Silver, D., 2018. Successor features for transfer in reinforcement learning. 1606.05312.

Bellman, R., 1957. A markovian decision process. *Indiana univ. math. j.*, 6, pp.679–684.

Campos, V., Trott, A.R., Xiong, C., Socher, R., Nieto, X.G. i and Torres, J., 2020. Explore, discover and learn: Unsupervised discovery of state-covering skills [Online]. *International conference on machine learning*. Available from: https://api.semanticscholar.org/CorpusID:211068721.

Chen, S., 2019. Diayn: Diversity is all you need. Available from: https://pub.towardsai.net/diayn-diversity-is-all-you-need-23aaa6532e84.

Co-Reyes, J.D., Liu, Y., Gupta, A., Eysenbach, B., Abbeel, P. and Levine, S., 2018. Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings [Online]. *International conference on machine learning*. Available from: https://api.semanticscholar.org/CorpusID:46977753.

Eysenbach, B., Gupta, A., Ibarz, J. and Levine, S., 2018a. Diversity is all you need: Learning diverse skills without a reward function. *Github* [Online]. Available from: https://github.com/haarnoja/sac/blob/master/DIAYN.md [Accessed 2024-04-08].

Eysenbach, B., Gupta, A., Ibarz, J. and Levine, S., 2018b. Diversity is all you need: Learning skills without a reward function. 1802.06070.

Flet-Berliac, Y., 2019. The promise of hierarchical reinforcement learning. Available from: https://thegradient.pub/the-promise-of-hierarchical-reinforcement-learning/#fn1.

Florensa, C., Duan, Y. and Abbeel, P., 2017. Stochastic neural networks for hierarchical reinforcement learning. 1704.03012.

Ganguly, A. and Earp, S.W.F., 2021. An introduction to variational inference. 2108.13083.

GeeksforGeeks, 2023. Epsilon-greedy algorithm in reinforcement learning. Available from: https://www.geeksforgeeks.org/epsilon-greedy-algorithm-in-reinforcement-learning/.

Gregor, K., Rezende, D.J. and Wierstra, D., 2016. Variational intrinsic control. 1611.07507.

Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor [Online]. In: J. Dy and A. Krause, eds. *Proceedings of the 35th international conference on machine learning*. PMLR, *Proceedings of Machine Learning Research*, vol. 80, pp.1861–1870. Available from: https://proceedings.mlr.press/v80/haarnoja18b.html.

Hansen, S., Dabney, W., Barreto, A., Wiele, T.V. de, Warde-Farley, D. and Mnih, V., 2020. Fast task inference with variational intrinsic successor features. 1906.05030.

Houthooft, R., Chen, X., Chen, X., Duan, Y., Schulman, J., De Turck, F. and Abbeel, P., 2016. Vime: Variational information maximizing exploration [Online]. In: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds. *Advances in neural information processing systems*. Curran Associates, Inc., vol. 29. Available from: https://proceedings.neurips.cc/paper_files/paper/2016/file/abd815286ba1007abfbb8415b83ae2cf-Paper.pdf.

Huang, S., Dossa, R.F.J., Raffin, A., Kanervisto, A. and Wang, W., 2022a. The 37 implementation details of proximal policy optimization [Online]. *Iclr blog track*. Https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. Available from: https://github.com/vwxyzjn/ppo-implementation-details.

Huang, S., Dossa, R.F.J., Raffin, A., Kanervisto, A. and Wang, W., 2022b. The 37 implementation details of proximal policy optimization [Online]. *Iclr blog track*. Https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. Available from: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

Hutsebaut-Buysse, M., Mets, K. and Latré, S., 2022. Hierarchical reinforcement learning: A survey and open research challenges. *Mach. learn. knowl. extr.* [Online], 4, pp.172–221. Available from: https://api.semanticscholar.org/CorpusID:246984577.

Irpan, A., Team, G.B. and Pastor, P., 2018. Scalable deep reinforcement learning for robotic manipulation. Available from: https://research.google/blog/scalable-deep-reinforcement-learning-for-robotic-manipulation/.

Jolliffe, I.T. and Cadima, J., 2016. enPrincipal component analysis: a review and recent developments. *Philos trans a math phys eng sci*, 374(2065), p.20150202.

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. *Corr* [Online], abs/1312.6114. Available from: https://api.semanticscholar.org/CorpusID:216078090.

Klimov, O., 2023. Bipedal walker - gymnasium documentation. Available from: https://gymnasium.farama.org/environments/box2d/bipedal_walker/.

Kreer, J., 1957. A question of terminology. *Ire transactions on information theory* [Online], 3(3), pp.208–208. Available from: https://doi.org/10.1109/TIT.1957.1057418.

Laskin, M., Liu, H., Peng, X.B., Yarats, D., Rajeswaran, A. and Abbeel, P., 2022. Cic: Contrastive intrinsic control for unsupervised skill discovery. *Arxiv* [Online], abs/2202.00161. Available from: https://api.semanticscholar.org/CorpusID:246442181.

Liu, H. and Abbeel, P., 2021a. Aps: Active pretraining with successor features. 2108.13956.

Liu, H. and Abbeel, P., 2021b. Behavior from the void: Unsupervised active pre-training. *Arxiv* [Online], abs/2103.04551. Available from: https://api.semanticscholar.org/CorpusID:232146715.

Mao, L., 2019. On-policy vs off-policy in reinforcement learning. Available from: https://leimao.github.io/blog/RL-On-Policy-VS-Off-Policy/.

Mohamed, S. and Rezende, D.J., 2015. Variational information maximisation for intrinsically motivated reinforcement learning. 1509.08731.

Oudeyer, P.Y. and Kaplan, F., 2008. How can we define intrinsic motivation ? [Online]. Available from: https://api.semanticscholar.org/CorpusID:14217330.

Pateria, S., Subagdja, B., Tan, A.h. and Quek, C., 2021. Hierarchical reinforcement learning: A comprehensive survey. *Acm comput. surv.* [Online], 54(5). Available from: https://doi.org/10.1145/3453160.

Pathak, D., Agrawal, P., Efros, A.A. and Darrell, T., 2017. Curiosity-driven exploration by self-supervised prediction. 1705.05363.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M. and Dormann, N., 2021. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of machine learning research* [Online], 22(268), pp.1–8. Available from: http://jmlr.org/papers/v22/20-1364.html.

Salge, C., Glackin, C. and Polani, D., 2013. Empowerment – an introduction. 1310.1863.

Schulman, J., Moritz, P., Levine, S., Jordan, M.I. and Abbeel, P., 2015. High-dimensional continuous control using generalized advantage estimation. *Corr* [Online], abs/1506.02438. Available from: https://api.semanticscholar.org/CorpusID:3075448.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., 2017. Proximal policy optimization algorithms. 1707.06347.

Shannon, C.E., 1948. A mathematical theory of communication. *The bell system technical journal* [Online], 27(3), pp.379–423. Available from: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shannon, C.E. and Weaver, W., 1998. eng*The mathematical theory of communication*. Urbana: University of Illinois Press Urbana.

Sharma, A., Ahn, M., Levine, S., Kumar, V., Hausman, K. and Gu, S., 2020a. Emergent real-world robotic skills via unsupervised off-policy reinforcement learning. 2004.12974.

Sharma, A., Gu, S., Levine, S., Kumar, V. and Hausman, K., 2020b. Dynamics-aware unsupervised discovery of skills. 1907.01657.

Sharma, A., Resident, A. and Research, G., 2020. Dads: Unsupervised reinforcement learning for skill discovery. Available from: https://research.google/blog/dads-unsupervised-reinforcement-learning-for-skill-discovery/.

Spelke, E.S. and Kinzler, K.D., 2006. Core knowledge. *Developmental science* [Online], 10(1), p.89–96. Available from: https://doi.org/10.1111/j.1467-7687.2007.00569.x.

Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction* [Online]. 2nd ed. The MIT Press. Available from: http://incompleteideas.net/book/the-book-2nd.html.

Sutton, R.S., Precup, D. and Singh, S., 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* [Online], 112(1), pp.181–211. Available from: https://doi.org/https://doi.org/10.1016/S0004-3702(99)00052-1.

SYNOPSYS, 2023. What is reinforcement learning? – overview of how it works | synopsys [Online]. Available from: https://www.synopsys.com/ai/what-is-reinforcement-learning.html#1 [Accessed 2023-12-02].

Appendix A

Bipedal Walker Environment

Win Condition. To solve the normal version, you need to get 300 points in 1600 time steps. To solve the hardcore version, you need 300 points in 2000 time steps.

Actions. Actions are motor speed values in the [-1, 1] range for each of the 4 joints at both hips and knees.

States. A state consists of: hull angle speed, angular velocity, horizontal speed, vertical speed, position of joints and joints angular speed, legs contact with ground, and 10 lidar rangefinder measurements. There are no coordinates in the state vector.

Rewards. Reward is given for moving forward, totalling 300+ points up to the far end. If the robot falls, it gets -100. Applying motor torque costs a small amount of points. A more optimal agent will get a better score.

Start. The walker starts standing at the left end of the terrain, with the hull horizontal, and both legs in the same position with a slight knee angle.

Termination. The episode will terminate if the hull gets in contact with the ground, or if the walker exceeds the right end of the terrain length.

Appendix B

Experimental Hyperparameters

Table B.1 details the default hyperparameters used in our experiments with PPO, as recommended by Raffin et al. (2021) and empirically explored and optimized.

Table B.1: Hyperparameters for PPO

Hyperparameters	Values
Learning Rate	0.0003
Total Timesteps	15,000,000
No. Environments	32
Max. Episode Steps	2048
Learning Rate Annealing	True
GAE	True
Gamma	0.99
GAE Lambda	0.95
No. Mini-batches	64
Update Epochs	10
Normalize Advantage	True
Clip Coefficient	0.18
Clip Value Loss	True
Entropy Coefficient	0.0
Value Function Coefficient	0.5
Max. Gradient Normalization	0.5
Target KL	None
Batch Size	65,536
Mini-batch Size	1,024
Observation Normalization	True
Observation Clip Range	[-10,10]
Reward Normalization	False
Reward Clip Range	[-400,400]

For further experimentation, Table B.2 details additional and adjusted parameters to TableB.1, as required by DIAYN (Eysenbach et al., 2018b). These are used for the control experiment

with DIAYN.

Table B.2: Hyperparameters for DIAYN control experiment

Additional/Modified Parameters	Values
Entropy Coefficient	0.1
No. Skills	20
No. PCA Observations	2500

Appendix C

Intrinsic Reward

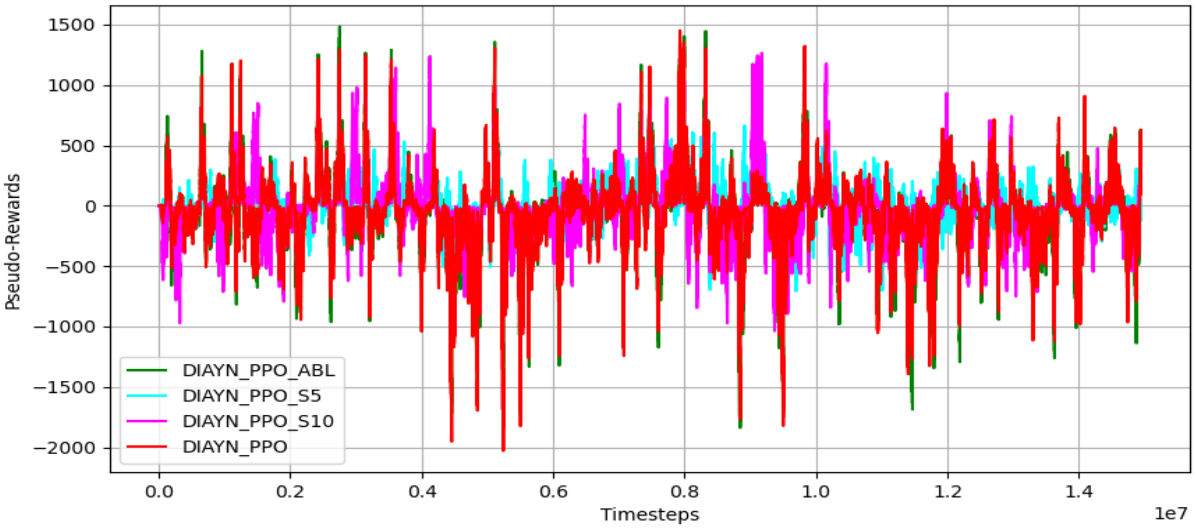

Figure C.1: Intrinsic reward over 15 million time-steps for DIAYN_PPO (Red), DIAYN_PPO_S5 (Blue), DIAYN_PPO_S10 (Magenta), DIAYN_PPO_ABL (Green).

The episodic returns of intrinsic rewards across these configurations are visualized in Figure C.1. These rewards, calculated as the scaled log probabilities that the discriminator correctly identifies the skills based on the states visited by the agent, provide insights into the discriminability of each skill. A high return indicates that the current skill is easily distinguishable based on the states it visits, whereas low or negative values suggest poor discriminability.

The results demonstrate that all configurations exhibit volatile performance, with scores swinging dramatically between high and low extremes, showing no increasing trend. This contrasts the expectation of increasing reward, proving this method fails to learn diverse skills. This pattern is influenced by the partial observability of the Bipedal Walker environment, where the agent's physical stance, not its exact location, determines state observation; implying that the agent repeatedly traverses through a limited set of physically feasible positions ("stay alive" states), which affects the diversity of skills. This demonstrates the limitation of DIAYN_PPO in partially observable environments using states to discriminate skills, which could be solved by a multitude of newer methods to this end, as described in Chapter 2.

Interestingly, configurations with a greater number of skills (DIAYN_PPO) show more pronounced extremes in scores compared to those with fewer skills (DIAYN_PPO_S5 and DIAYN_PPO_S10). This suggests that having more skills increases the granularity with which skills can be discriminated based on state observations, although it also introduces greater variability and potential overlap in the skills' state-space coverage.