UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Exploration in Complex Naturalistic Behavior

Permalink

https://escholarship.org/uc/item/5hp4x26t

Author

Singla, Umesh Kumar

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Exploration in Complex Naturalistic Behavior

Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Umesh Kumar Singla

Committee in charge:

Professor Marcelo Mattar, Chair
Professor Sicun Gao, Co-Chair
Professor Taylor Berg-Kirkpatrick

2022

The thesis of Umesh Kumar Singla is approved, and is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

EPIGRAPH

You have to believe in something,
no matter how stupid it sounds.
— The Servant

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am beyond thankful to Marcelo Mattar for letting me join this remarkably wonderful group of people and for allowing me to freely explore cognitive science in his lab. I am thankful for his enthusiasm, patience, kindness, and dedication, which made working on my thesis a pleasure. I am especially thankful to Matthieu Le Cauchois for his precise and elaborate comments over the duration of this project as well as for engaging me in fascinating discussions about life, research, and philosophy. His commitment to perfection has taught me the value of detail and clarity in science and what it takes to do rigorous science. A deep thank you to my lab colleagues, in particular, Kokila, Homero, Dan, and Kris, who have been a tremendous source of inspiration from the beginning of the project. My special thanks to Aayush Naik for encouraging me to take bolder steps and Ashutosh Narang for being by my side throughout the length of my master's. Finally, I am profoundly thankful to my mother and my sister for their constant encouragement and support, without which none of this would have been possible.

ABSTRACT OF THE THESIS

Exploration in Complex Naturalistic Behavior

by

Umesh Kumar Singla
Master of Science in Computer Science

University of California San Diego 2022

Professor Marcelo Mattar, Chair
Professor Sicun Gao, Co-Chair

Exploration and search are such crucial occurrences in the natural world around us, yet we don't know much about what drives the precise structure we observe. While exploration by animals in discrete choice tasks has been extensively researched, exploration in sequential contexts has received little attention. We take a behaviorally rich dataset of mice exploring a labyrinth by Rosenberg et al. [2021] and model it using search strategies from foraging literature in an RL framework. We discovered that an ecologically inspired Lévy walk model adequately explains the efficiency and preferences of mice exploring the labyrinth. We implemented the model in the temporally extended ϵ-greedy exploration framework, which allowed us to interpret the search strategy using general principles. We found that animals exhibit super-diffusive behavior and leverage temporal persistence to navigate the maze rather than making decisions at each intersection. Our study provides a new perspective on Lévy flight foraging and opens new avenues for investigating the interaction between exploration dynamics and the naturalistic environments.

Chapter 1

Introduction

If you were to experience the complicated landscape of Venice or New York City for the first time, you would likely be struck by the beauty and intricacy of the city. You would likely be amazed by the winding canals, the historic architecture, and the vibrant culture of Venice. You would be amazed by the towering skyscrapers, the bustling streets, and the vibrant energy of the New York City. You might feel a bit overwhelmed at first by the seemingly endless maze of narrow streets and waterways, or the crowds and the fast pace of life in case of NYC, but you would likely be excited to explore and discover the hidden gems of the city. If your sole aim for the evening is to explore the city or look for a fun dinner place, you would start by wandering around the area. Even if it's your first time, you are fairly confident that you will be able to find something to engage yourselves with that evening, and that you will learn more about the city without getting lost. There are a few steps that one can take to help ensure that they do not get lost and be able to discover the sounds and sights of the city. You will likely start by doing some research and familiarize yourself with the layout of the city and the different neighborhoods. This can help you to get a sense of the different areas and how they are connected, and can make it easier to orient yourself once you are there. You are confident that you will not get lost in the city because you have probably done something similar before on an another trip. Alternatively, you *feel* you have a natural understanding of how space is organized, which can help you to orient yourself and find your way around the city. This understanding of spatial design

gives you a sense of confidence and assurance, and is going to help you navigate the city with ease. In any case, your curiosity about things and a capability to move around will most probably lead you to discover something to do in the city that will be memorable for life.

Why is it that we tend to keep exploring despite any certainty of getting a reward? And how is it that we are able to cover a large portion of space in so little time without getting stuck and without getting overwhelmed by the sheer number of possible paths and options to choose from? Berlyne's book on 'Conflict, Arousal and Curiosity', that influenced the development of research into animal exploratory behavior for decades, was the first to make a distinction between extrinsic and intrinsic exploration [Berlyne, 1960]. Extrinsic exploration was now seen as behavior focused on achieving a specific goal or fulfilling a particular requirement. In contrast, intrinsic exploration behavior is characterized by the investigation of the stimuli and is driven by curiosity and an interest in the stimuli itself. But, while the distinction between the two types of exploration is extremely useful for gaining insights into the brains and behavior, in practice it can be difficult to disentangle the two types of exploration since the two generally go hand in hand. Even if the motives or the consequences of the two differ, the responses can be identical [Hughes, 1997]. Building upon this motivation, we seek to tease apart and understand various components that could underlie exploratory behavior: the ones that are driven by external factors, or by intrinsic motivation, or simply are an outcome of the neuronal noise, and lastly, possibly originated as a part of millions of years of the natural selection process [Viswanathan et al., 2011]. The problem of foraging aims to identify the complex mixture of behavior and physical structure of animals that gives rise to their efficiency in gathering food in a variety of environments. Identifying characteristics of efficient search and exploration is one important component in this problem [Krebs and Stephens, 2019].

1.1 Exploration

Biological intelligence is characterized by its ability to quickly learn new concepts from only a handful of examples, adapt to changing surroundings, and use previous knowledge to make predictions about the environment [Hassabis et al., 2017]. Efficient exploration in the real world is one hallmark of biological intelligence. Exploration is fundamental for survival of animals and human beings. Animals need to keep looking for sources of food and new places to shelter to keep themselves safe from getting predated by avoiding being in the same place for long. Human exploration in the modern world is important for similar reasons but also to keep looking for new things and experiences in life to avoid the stress of monotonicity and keep the will to live alive.

Understanding the exploratory behavior of humans and animals is one of the central goals of behavioral ecology and neuroscience. Exploratory behavior is a natural expression of spatial learning [O'keefe and Nadel, 1979]. Exploration in animals has been widely studied in open field foraging tasks [Tchernichovski et al., Tchernichovski and Benjamini] and in discrete settings such as multi-arm bandits [Costa et al., 2019]. The dynamics of exploration have also been studied at multiple time scales - from characterizing twists and turns in a novel environment to capturing movement patterns in mazes or farmlands [Atkinson et al., 2002]. Modeling animal exploration can help us capture normative principles behind their strategies and provide a low-dimension description of behavior, which can further be correlated with the neural activity to gain insights into the brain.

To further motivate the question of why the exploratory behavior are not fully random, we can look into a recent example from the neuroscience literature. Coen et al. [2014] analyzed the type of courtship songs in Drosophila which showed that the type of song can be predicted by postural cues from the female, previously thought to vary at random. By analyzing more than 100,000 love songs and carefully monitoring the location of the courting couple, the authors suggested that a logic and order exist in the apparent musical randomness. Further, this behavior is also better predicted by a model that takes into account an estimate of the singing male's internal state. A thorough and detailed analysis of complex and

unpredictable behavioural patterns resulted in the identification of simple underlying rules, showing behavioural variability and complexity can have an underlying structure and can help us with understanding nervous-system function and identify the computational problems brains are trying to solve [Ölveczky, 2014]. There's no reason to expect a structure in the animal movement behavior but there is also no reason to not expect.

1.2 Naturalistic Behavior

While a large portion of the literature has focused on exploration in simplistic settings of bandits tasks or open fields, only a few studies have tried to model behavior in larger, complex environments. The experiments in neuroscience do not come close to the complexity, the sequential nature and naturalistic quality of real life settings. They also involve significant amount of human interference and tend to have very short periods of exploration. In biological organisms, it is clear that exploration is not optimal and animals resort to heuristics, possibly due to limited cognitive resources. In more realistic settings which are generally sequential in nature, algorithms ought to be much more complex. Recent developments in computer vision and machine learning tools have made it possible to track and monitor animals with unprecedented accuracy. We are now seeing an upsurge in the use of complex mazes for studying animal behavior, which contain many choice points and much larger spatial area [Vallianatou et al., 2021, Uster et al., 1976, Alonso et al., 2020, Grobéty and Schenk, 1992, Nagy et al., 2020]. One such recent experiment developed by Rosenberg et al. [2021] involves animals exploring a complex labyrinth for hours without any human interference whatsoever. The animals have access to sufficient food and water in the home cage and the maze offers no explicit external reward. But, as noted in Rosenberg et al. [2021], we see animals continue to explore the maze throughout the night. While this behavior supports the presence of some kind of intrinsic motivation in the animals that is making them explore, the structure and efficiency exhibited by animals in itself is quite remarkable and makes up for a perfect example of complex yet naturalistic

exploratory behavior that remains poorly understood.

1.3 Our Contribution

While exploration by animals in discrete choice tasks has been extensively researched, exploration in sequential contexts in biological agents has received little attention. During cognitive activities such as spatial navigation, animals demonstrate complex yet highly structured behavior. Despite their structure, the concepts underpinning exploratory patterns are poorly understood. To our knowledge, no existing MDP algorithm has attempted to capture exploratory patterns in animal movement in complex real-life settings. In this work, we emphasize on the sequential settings from the natural world that offer sparse or zero external rewards. The motivation behind this and the future work is to identify the structure in the natural world around us, attempt to learn the origins behind it, and computationally model it to help build better and efficient autonomous agents. Once we have separated out the seemingly random component of exploration and have replicated it, we can move on to other aspects of the problem such as learning and adaptation. That is, adding the effects of a "brain" to a "brainless" model.

We take a behaviorally rich dataset of mice exploring a labyrinth by Rosenberg et al. [2021] and model it using search strategies from foraging literature in a reinforcement learning framework. We develop a set of evaluation metrics through careful data analysis. We discovered that an ecologically inspired Lévy search model adequately explains the efficiency and preferences of mice exploring the labyrinth. We implemented the model in temporally-extended ϵ-greedy exploration [Dabney et al., 2020] framework, allowing us to interpret the search strategy using general principles: animals exhibit super-diffusive behavior and leverage temporal persistence to move in the maze rather than making decisions at every junction. Our study provides a new perspective on Lévy flight foraging and opens new avenues for investigating the interaction between exploration dynamics and the naturalistic environments. There are a number of possibilities for interesting theoretical and experimental research on the complex dynamics of biological

interactions. Future studies may uncover new and unexpected insights into the mechanisms and behaviors that govern these interactions.

1.4 Outline

In this work, we review a set of algorithms proposed in the Reinforcement Learning literature on exploration in sequential settings in section 3.1, with the goal of identifying algorithms that could plausibly describe animal behavior. In chapter 2, we also review the literature on anomalous diffusion and a number of random models that have been proposed in movement ecology to describe animal foraging patterns. We then analyze in chapter 4 a rich dataset of animals exploring a labyrinth, a fairly complex naturalistic setting and build a set of evaluation metrics. And finally in chapter 5, we conclude by proposing a new candidate algorithm inspired by ecology and implemented in a reinforcement learning framework that seems to capture much of the animal behavior we witnessed in our dataset.

Chapter 2

Preliminaries

This chapter introduces core concepts that will be used throughout the thesis. We briefly present the full reinforcement learning problem in section 2.1, its formulation using the markov decision process framework in subsection 2.1.1, and introduce the problem of exploration as tackled in artificial intelligence in bandit or sequential settings (subsection 2.1.3). We then introduce the relatively recent work in RL on capturing the behavior at multiple time scales in section 2.2, briefly reviewing the theory behind semi-MDPs and options in subsection 2.2.1.

2.1 Reinforcement Learning

Reinforcement Learning (RL) provides a framework to study how an agent can learn to choose actions that maximize the *reward* signal. When an RL agent starts acting in an environment, it typically does not have any knowledge about the task or the environment it needs to tackle. RL agents are interactive: they have a certain goal, can sense parts of their environments and choose actions to influence their environment. The agent has to operate despite significant uncertainty about the environment it is in. The agent interacts with the environment, observes the consequences of its actions in the form of transitions and feedback, and then is expected to use these observations to perform better. The interactivity is the distinguishing element of an RL agent. Reinforcement learning is the closest to the kind of learning that humans and other animals do, out of all the forms of machine

learning approaches. Many of the core algorithms of reinforcement learning are inspired by biological learning systems. Various RL algorithms operate in mildly different ways, although most of them do some variation of state value estimation at their core [Sutton and Barto, 2018].

Model-Based vs Model-Free RL When it comes to learning about the environment, it is important to distinguish algorithms that are model-based and that are model-free. When the agent is supposed to learn the model of the environment, models can be used for planning a course of action by considering possible future states and transitions. While this seems obviously useful, it also requires more computational resources as well as adds work for us to find its biological implementation. On the other hand, we have model-free methods that are trial-and-error learners - which could be viewed as almost the opposite of planning [Sutton and Barto, 2018].

2.1.1 Markov Decision Processes

Reinforcement learning can be set within the Markov Decision Process (MDP) formalism [Puterman, 1994]. An MDP is defined by the tuple (S, A, P, r, γ) where:

- $s \in S$ is a state in the state space.
- $a \in A$ is an action in the action space.
- $P(s'|s, a)$ is the probability of transitioning from state s to state s' on taking action a. MDPs assume that the environment is Markovian, so the transition probability $P(s'|s, a)$ depends only on the current state-action pair. That is, $P(S_{t+1} = s'|s_t, a_t) = P(S_{t+1} = s'|s_t, a_t, s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, ..., s_0, a_0)$. All the sequence of observations, actions and information on environment obtained by agent during its journey is called **history**.
- $r : S \times A \to \mathbb{R}$ is the reward function which maps the current state-action pair (s, a) to the immediate reward obtained from the set of real numbers.

- $\gamma \in [0, 1)$ is the discount factor. With $\gamma = 0$, it considers only immediate rewards. γ discounts the value of future rewards.

Policy Let $\mathbb{P}(A)$ denote the space of probability distributions over actions; then a policy $\pi : S \to \mathbb{P}(A)$ assigns some probability to each action conditioned on a given state. Policy describes an agent's behavior, that is, tells it what action to take in a given situation. Generally an RL agent starts with a random policy and continues to improve its policy by learning from the outcomes of its actions in the environment. Value-iteration methods help agent estimate the expected reward (or value) it can get by taking an action in its current state. However, various methods also exist in policy space where we directly optimize over policy functions.

2.1.2 Exploration-Exploitation Trade-off

One of the core challenges of RL is the *exploration-exploitation trade-off*. To continue to improve its policy and perform better, the agent needs to keep gathering more and more information about the environment which could mean acting sub-optimally at times (explore). If the agent only manages to visit a portion of the environment (that is only a subset of states), its knowledge about the environment remains low and might perform poorly when it comes to new observations. However, if it focuses too much on the parts of the environment that it has not explored enough, it loses the chance to get immediate rewards that it knows about. This trade-off depends on a number of factors, such as environment dynamics, presence of immediate or long-term rewards, number of states and their properties, set of actions, etc.

Another one of the core challenges in RL is the *temporal credit assignment*. The agent must accurately assign the credit to past actions that helped it in achieving long-term return and in an efficient manner. However identifying such actions is not a trivial problem.

We only focus on the problem of exploration-exploitation trade-off in this work. In fact, the majority of the current work is only focused on the exploration phase.

2.1.3 Exploration

Exploration plays a central role in RL and it has been studied quite extensively in the field. Exploration methods have evolved from simple ideas such as pure randomization, to increasingly effective algorithms that come with theoretical guarantees in various domains and have shown impressive performance in complex problems. In finite state-action spaces, they perform quite well and are well-understood but when it comes to large-scale environments, or sparse reward settings such as in sequential problems, these algorithms are of limited practical use.

We are primarily concerned with exploration in sequential settings where exploration has even greater impact. In sequential settings, exploration controls not only the immediate information but also the potential information an agent could get in the future. There's been a ton of applications of exploration algorithms in bandit settings that perform really well and come with theoretical guarantees (e.g. PAC bounds) [Kearns and Singh, 2002, Brafman and Tennenholtz, 2002, Azar et al., 2017, Wang et al., 2020]. We expand on existing exploration methods in sequential settings in section 3.1 in detail since it's the primary concern of our work.

Based on the information exploration algorithms tend to utilize, they are generally categorized generally into undirected and directed exploration [Thrun, 1992]. In **undirected** exploration methods, an agent selects exploratory actions at random, without using any exploration-specific knowledge. Random walk is the simplest method in this category. Another example is the Boltzmann distribution (based on inverse-temperature parameter for balancing the exploration and exploitation).

In **directed** exploration methods, agents leverage the obtained information to pursue the exploration of states that haven't been visited recently or more generally, the agent thinks would be more informative. How does an agent formalize the notion of *information*? When it comes to pure exploration, what information are they seeking or what information could be useful? For example, in a city or a maze, exploration is necessary for acquiring spatial information. Exploration

can be viewed as a search for undiscovered rewards or looking for alternate ways to get to known rewards or avoid preys. There are various ways to think about this: in terms of "curiosity-seeking", reducing prediction "uncertainty", "sensation seeking", "novelty seeking", energy constraints, etc [Hughes, 1997]. We look at definitions of some of these and the exploration algorithms based on them in section 3.1.

2.2 Temporal Abstraction in RL

It has been argued that in order to scale to large problems, RL agents should be able to reason at multiple temporal scales [Dayan and Hinton, 1992, Sutton et al., 1999, Precup, 2000, Kaelbling, 1993, Dabney et al., 2020]. Learning, planning, and representing knowledge at multiple levels of temporal abstraction are key challenges for AI. To tackle this, Sutton et al. [1999] introduced for the first time, temporally-extended courses of actions - **options**. Options are closed-loop policies to take actions over a period of time. It could be the same action repeated a number of times or a *composite* action. Depending on how options are designed to evolve, we may have **markov options** as well as **semi-markov options**. We do not, however, expand on the closed-loop nature of options and how to learn and improve options in this thesis.

Examples of options include getting on a bike, switching on the TV, going to lunch, and traveling to a distant city, and primitive actions such as muscle twitches or moving left can be regarded as options as well (Figure 2.1). Options are an extension of the general notion of an action - so options may be used interchangeably with primitive actions in existing planning and learning methods. Precup et al. [1998] and Sutton et al. [1999] show that options enable temporally abstract knowledge and action to be included in the RL framework in a natural and general way. By including the options to existing notion of primitive actions, the framework allows an agent to work simultaneously with high-level and low-level temporal representations.

Figure 2.1: A cooking activity involves taking actions at multiple time scales. At high level: choose a recipe, make a grocery list. At medium: get a pot, collect ingredients. At lower: wrist and arm movement, stirring, and such. Example from Precup [2000].

2.2.1 Options and Semi-MDPs

Temporally-extended actions are represented by a policy (behavior) together with a termination condition. An action represented in this way is called an option. In an action-repeat kind of option (e.g. 'go-forward $k = 6$ times'), the termination condition could merely be when the number of times the action was to be repeated becomes zero ($k = 0$).

The current definition of options is designed to make them similar to actions as much as possible while adding the possibility of them being temporally-extended. Options can be easily incorporated in MDPs, allowing an agent to use existing algorithms and heuristics for selecting actions or courses of action. There's a lot of ongoing work on coming up with a different action representation [Sharma et al., 2017] or being able to learn and discover options, either depending on the environment [Kulkarni et al., 2016] or in a task-agnostic fashion [Amin et al., 2021a, Riemer et al., 2018, Harb et al., 2018, Vezhnevets et al., 2017, Fox et al., 2017]. Particularly, the study by Kulkarni et al. [2016] designs intrinsic rewards to aid longer sequences of actions.

Options Options consist of three components: a policy $\pi : S \times A \rightarrow [0, 1]$, a termination condition $\beta : S^+ \rightarrow [0, 1]$, and initiation set $I \in S$ [Sutton et al., 1999]. Note that $[0, 1]$ denotes all the real values from 0 (inclusive) to 1 (inclusive) to denote the probability range. Initiation set consists of states where a certain option can be initiated for execution. An option $< I, \pi, \beta >$ is available in state s_t if and only if $s_t \in I$. That is, given a set of options O, the available options O_s for each state s is implicitly defined by how each option is initialized. When an option is executed, the actions are chosen according to the policy π until the option is terminated according to the probability β.

Markov and Semi-Markov Options Sometimes it is useful to have a timeout on an option to terminate before it has reached a particular intended state [Sutton et al., 1999]. With Markov options, the decision to terminate solely depends on the current state. Semi-Markov options are defined to overcome this problem and extend the framework to even more cases of possible interest. With semi-Markov options, policies and termination conditions can make their choice depending on the events that have occurred since the option was initiated.

In their most general formulation, an option is initiated at some time t, determines the actions selected for some number of steps k, and then terminates in state s_{t+k}. At each intermediate time τ, $t \leq \tau \leq t + k$, the decisions of a Markov option may depend only on s_τ, whereas the decisions of a semi-Markov option may depend on the entire preceding sequence $s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}, \ldots, r_\tau, s_\tau$ but not on events prior to s_t or after s_τ. This sequence is called the *history* from t to τ, denoted by $h_{t\tau}$. The set of all *histories* is usually denoted by Ω. In case of *semi-Markov* options, the policy and termination condition are functions of possible histories, that is, $\pi : \Omega \times A \rightarrow [0, 1]$ and $\beta : \Omega \rightarrow [0, 1]$.

A set of options defined over an MDP formally constitutes a **semi-MDP** or **SMDP**. We state the theorem 1 from Sutton et al. [1999] below and with it, we wrap our review of options and semi-MDPs theory (Figure 2.2).

> (MDP + Options = SMDP). For any MDP, and any set of options defined on that MDP, the decision process that selects only among those options, executing each to termination, is an SMDP.

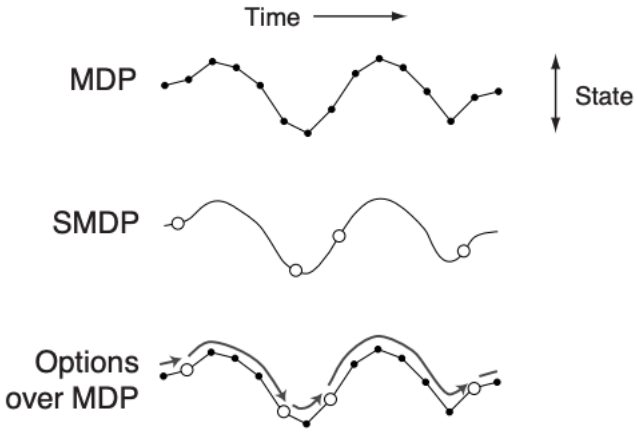

Figure 2.2: Actions in MDP vs SMDP. Figure from Precup [2000].

Options can also select other options, such as in hierarchical structures, giving rise to higher-level options that are also *semi-Markov* (even if all the lower level options are *Markov*). Overall, semi-Markov options include a very general and wide range of possibilities [Sutton et al., 1999].

Frame Skipping in DQN Frame skipping proved to be an important element in the success of Deep Q-Network (DQN) in tackling various Atari games [Mnih et al., 2015]. Skipping few states allowed DQN to be more efficient in learning the policy by reducing the need for observations since the difference between successive observations tends to be small. The agent skips over a few states (frames) and continues to repeatedly playing the same action before deciding to terminate the action-repeat sequence and selecting a new action or option to proceed. Taking it further, various studies [Braylan et al., 2015, Khan et al., 2019] have shown additional improvement in the performance by tuning the skip-size. One line of work tries to learn policies for a fixed skip-size set of action sequences [Metelli et al., 2020, Lee et al., 2020] in parallel to the behavior policy. Along similar lines, Dabney et al. [2020] showed a drastic improvement in the performance as well as generalizability of the principle across domains by using a certain skip-size distribution with motivation from behavioral ecology literature.

2.2.2 Hierarchical Reinforcement Learning

Hierarchical reinforcement learning (HRL) is a type of reinforcement learning that uses multiple levels of temporal abstraction to solve complex problems [Dayan and Hinton, 1992]. HRL divides a complex task into simpler subtasks and assigns each subtask to a different agent. Each agent is trained to optimize its own subtask, while the overall task is optimized by coordinating the agents together. By using multiple levels of temporal abstraction, HRL is able to solve complex problems that may be too difficult for traditional reinforcement learning algorithms.

Chapter 3

Related Work

This chapter attempts to review the advances in the study of exploration and more generally, behavior, in reinforcement learning and ecology literature. In section 3.1, we briefly review the key exploration algorithms and define various notions of intrinsic motivation that have been proposed for sparse sequential settings over the last two decades. In section 3.2, we review some of the key search models from movement ecology that have been used to describe various animal species' movement behavior and the principles behind them.

3.1 Exploration Methods in Sequential Tasks

Exploration in sequential settings is an interesting problem, and even more so when it comes to sparse reward settings. If the agent does not get any immediate reward or does not get any reward at all, it's still helpful for the agent to keep exploring in the hope it will get some in the future. The challenge here is to model such an agent that is always looking for *something* despite any external motivation for any general situation.

As noted in subsection 2.1.3 on exploration in RL, recent studies have come up with definitions of various notions of intrinsic motivation and have utilized them to drive exploration in various sequential tasks such as ATARI games. There are also techniques for exploration that act completely blind, that is, the agent selects actions in the absence of any information obtained from environment or does not

have any metric to measure or track its own performance. We list below a few notable techniques taken from a recent survey on the topic by Amin et al. [2021b] that are relevant for the current work.

3.1.1 Blind exploration

ϵ-greedy remains one of the widely used methods for exploration still. In the ϵ-greedy approach, the parameter $\epsilon \in [0, 1]$ controls the balance between exploration and exploitation. The action at every step t is chosen such that,

$$a_t = \begin{cases} a_t^* & \text{with probability } 1 - \epsilon \\ random & \text{with probability } \epsilon \end{cases} \tag{3.1}$$

where a_t^* is the greedy action taken at time t. For settings with no external rewards, that is no greedy choice exists, it reduces to simply taking a **random** action at all times. Needless to say, when it comes to being decisive about following a certain direction just to see what's out there, ϵ-greedy is very shortsighted and the probability of moving consistently in a direction decays exponentially with the number of steps [Dabney et al., 2020].

3.1.2 Intrinsically-Motivated Exploration

Novelty

Shyam et al. [2019] proposed an exploration algorithm, Model-based Active eXploration (MAX), which utilizes **novelty** of transitions as a learning signal. They define novelty as the Jenson-Shannon divergence between the predicted space of distributions and the resulting distribution, and choose the action which maximizes this novelty measure at each step. In environments that are inherently static and do not change by agent's interactions, this algorithm quickly reduces to a random walk once the environment transitions are known.

Information Gain

Information Gain has been used in exploration strategies as a measure of intrinsic motivation by many studies. One such study by Little and Sommer [2013], proposes maximizing predicted **information gain** (PIG). Agents keep an internal model of the environment. Upon taking an action, it calculates the KL-divergence of its updated internal model from the one it had predicted before taking the action. The KL-divergence can be used as a measure of inaccuracy or missing information I_m in agent's model. Agents then take actions that decrease this missing information I_m the most, since that would lead to explore relatively more uncertain areas of the environment known to agent, as indicated by the larger I_m.

Entropy

Another measure that has been proposed by Hazan et al. [2019] is to maximize **entropy** of the distribution over visited states. The idea behind it is to efficiently learn policies in an MDP which optimize task-agnostic reward functions, for example, optimizing an objective that is only a function of the state-visitation frequencies. They propose one such objective could be the cross-entropy reward function: cross-entropy between a uniform distribution and the policy-induced distribution over states, and maximize this function. In particular, it generates and optimizes a sequence of intrinsic reward signals. The cross-entropy encourages the "most uniform" random walk over the MDP. The optimal policy obtained is called MaxEnt exploration policy.

Curiosity

A particularly interesting work on the topic is the Intrinsic Curiosity Module (ICM) by Pathak et al. [2017] which has gained popularity in the field. ICM define **curiosity** as "the error in an agent's ability to predict the consequence of its own actions", or simply put, **state prediction error** conditioned on action taken. The authors primarily evaluate it on continuous action and state paradigms such as 3D game of VizDoom or classic game of Super Mario Bros, so the input to their algorithm is pixel frames. It focuses on representing part of the environment that

either affects the agent or is affected by the agent i.e. influential feature space. If there is a source of variation that is inconsequential for the agent, then the agent has no incentive to know about it. To reiterate an example given by the authors to help understand the intuition: if the agent is observing movement of leaves in a breeze, it is hard to predict movement of leaf pixels but the vanilla state prediction error remains high and it actually brings no real utility to the agent.

The Intrinsic Curiosity Module consists of two models - a forward and an inverse, as described below:

- The inverse model helps learn a feature space that encodes information relevant for predicting the agent's actions only. It is trained to predict the action taken using the difference in features of the previous state $\phi(s_t)$ and the state that resulted in after taking that action $\phi(s_{t+1})$. The predicted action is never used, it just ensures the model is only learning the space that is affected by the agent or affects the agent. Since the network is only required to predict the action, it starts getting rid of the state feature space that is not relevant for the action.

- The forward model makes predictions in this feature space and this prediction error is used as a positive curiosity reward to drive the exploration.

In short, it learns to explore the parts of the space (like the video screen) that the agent is not able to predict what would happen if it "goes" there very well. It is also subtly different from relying on the vanilla state prediction error (the famous TV problem).

Space coverage

One algorithm, PolyRL by Amin et al. [2021a], takes a slightly different approach to the problem. The authors intuit that if the reinforcement signal is very scarce, the agent should rely on some form of short-term memory to be able to cover its environment efficiently. They introduce a **measure of spread** in the state space as the metric to optimize avoid getting stuck in a small region and be

able to generate persistent trajectories in a certain direction. They call this feature of the trajectories as locally self-avoiding random walks (LSA-RWs). PolyRL is built upon concepts from statistical physics used to explain behavior of free-rotating chains in polymer physics. It is primarily motivated by the continuous action and state space problems.

A key point about this method is that the PolyRL is able to reproduce the consistent movement behavior without the need for action-repeats. Previous work done on similar topics by Dabney et al. [2020] or Sharma et al. [2017] emphasized the need for either action-repeats, or learning composite actions or learning a different action-representation. This algorithm remains to be tested in our task and it would be interesting to evaluate it with respect to animal behavior in future. As highlighted in section 3.2 later, animal ecologists have identified a form of directional persistence and short-term memory effects in various species, and PolyRL bears a resemblance to the two effects.

A huge argument in favor of PolyRL is none of the modern exploration algorithms address short-term memory or generalizability over multiple environments. An effective exploration algorithm has to be generally applicable. We expand on the notion of temporal abstraction further in section 2.2.

Table 3.1: Examples of some reward-free exploration approaches.

Algorithm	**Intrinsic Motivation**
Shyam et al. 2019	Novelty
Little and Sommer 2013	Information Gain
Hazan et al. 2019	Entropy
Pathak et al. 2017	Curiosity
Amin et al. 2021a	Space Coverage
Dearden et al. 1998	Uncertainty

The above reward-free exploration methods and the corresponding notion of intrinsic motivation they use are summarized in Table 3.1. Various other notions of intrinsic motivation also exist, such as boredom [Schmidhuber, 1991b], adaptive curiosity [Schmidhuber, 1991a], or surprise [Modirshanechi et al., 2021]. For reward-free settings and static environments, these perform relatively similar to those described above, so we omit their discussion here.

3.2 Animal Foraging

The physics of foraging studies and builds mechanistic explanations of animal movement behavior. Animal foraging behavior has been studied in environments varying in the density of food, patchiness or terrains. Although they do not characterize any learning with experience, such limiting models may help to quantify important features of the exploration dynamics and help separate the problem of distinguishing the learning component of behavior from that of the random walking[Viswanathan et al., 2011]. Below we review some of the important concepts from movement diffusion literature and their evidence as studied in various biological entities.

3.2.1 Anomalous Diffusion

Diffusion is the net movement of anything (for example, atoms, ions, molecules, energy) generally from a region of higher concentration to a region of lower concentration. The concept of diffusion is widely used in many fields, including physics, chemistry, biology, sociology, economics, and finance. Diffusion in ecological models is generally characterized as anomalous diffusion and is described by a non-linear relationship between the mean squared displacement (MSD), and time t:

$$MSD \propto t^{\alpha} \tag{3.2}$$

Depending on α, we observe different kinds of behavior:

1. Diffusive Behavior ($\alpha = 1$)
2. Superdiffusive Behavior ($\alpha > 1$)
3. Subdiffusive Behavior ($\alpha < 1$)

In this work, we are primarily concerned with diffusive and superdiffusive behavior.

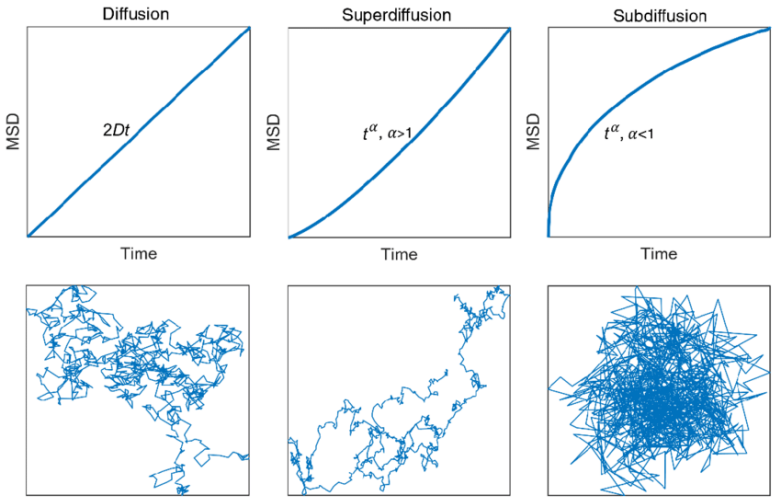

Figure 3.1: The (top) MSD with time for diffusive, superdiffusive and subdiffusive motion and (bottom) examples of the corresponding trajectories. Figure from Wadkin et al. [2021].

3.2.2 Random Walk

Uncorrelated random walkers or random walkers (RW) choose a decision uniformly from the choices at each time step depicting a Brownian motion kind of behavior [Bartumeus et al., 2005]. Standard methods in spatial ecology traditionally have assumed Brownian motion as a basic properties of animal movement. It exhibits normal diffusive behavior and tends to lead to the problem of oversampling. Random Walk does not account for directional persistence and they are relatively short-sighted when it comes to covering large distances. While real organisms have a tendency to continue moving in the same direction and they rarely make 180 degree turns.

To overcome this lack of directional persistence, we have correlated random walks (CRWs) where short-term correlations are introduced in the random walk (Figure 3.2). An example of such a correlation would be constraining the turning angle distribution (ρ) between the next step vector and the current step vector to reduce sharp frequent turns.

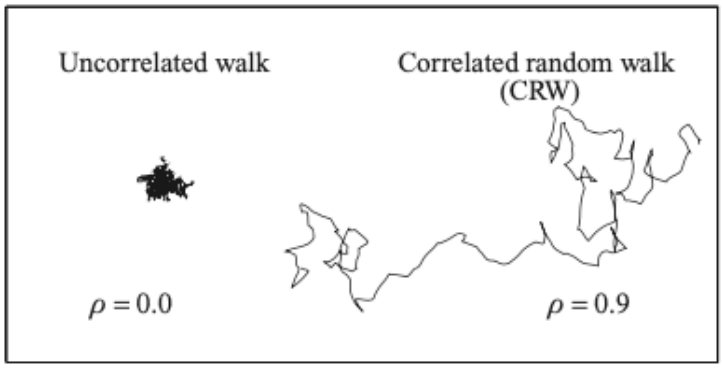

Figure 3.2: Correlated random walks (CRWs) are different from uncorrelated random walks due to directional persistence. Figure by M. L. Felisberto.

3.2.3 Lévy Walk

Lévy Walks are characterized by long periods of wandering in a small area and occasionally longer steps in a random direction. Lévy Walks add further directional persistence to correlated random walks in a scale-free manner. It tends to exhibit superdiffusive behavior and have been known to optimize search efficiencies in a variety of animal species, such as birds and fish [Viswanathan et al., 2002, Hills et al., 2013, Viswanathan et al., 1999, Sims et al., 2019].

To formalize LW(μ):

- Step-lengths are sampled from a heavy-tailed distribution (parameterized by μ).

$$p(l) \sim l^{-\mu}, l > l_0 \tag{3.3}$$

  where $l > l_0$ is a lower cutoff where the power law tail begins.

  Variation of the parameter μ allows superdiffusive Lévy searches as well as Brownian searches involving normal diffusion. By varying μ, we can determine the search efficiency in a given environment and estimate how much advantage can be gained by exploiting diffusivity and randomness. The general range for μ is considered to be between 1 and 3, where $\mu = 1$ denotes long ballistic motion and $\mu = 3$ corresponds to normal diffusion. Lévy walks and flights correspond to intermediate values of μ.

- Step-direction is sampled uniformly over 360 degrees.

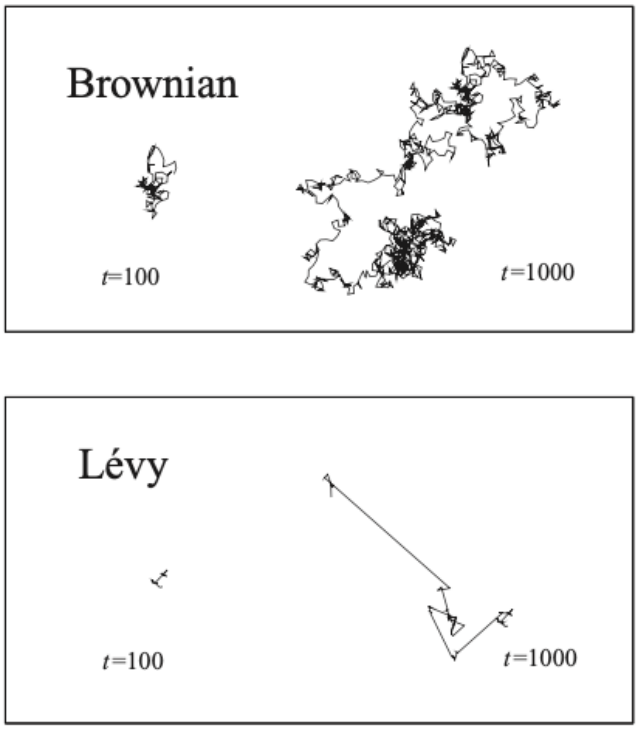

Figure 3.3: An example of 1000 steps of a Lévy flight in two dimensions in comparison to 1000-step Brownian motion. Figure from Viswanathan et al. [2011].

Levy flight foraging hypothesis Since Lévy flights have been shown to optimize random searches, the Lévy flight foraging (LFF) hypothesis states that the biological organisms must have evolved to exploit Lévy flights [Viswanathan et al., 1999]. But Viswanathan et al. [2011] attempts to generalize this hypothesis to describe species and circumstances where a more mixture set of search processes have been observed. The exact reformulated LFF hypothesis is stated below.

> Superdiffusive motion governed by fat-tailed propagators optimizes encounter rates under specific (but common) circumstances; hence some species must have evolved mechanisms that exploit these properties of Lévy walks.

3.2.4 Intermittent Search

It is shown alternating between walking and intensive search tends to explain the behavior of animals behavior. Intermittent searches capture the behavior at two scales of movement: walking followed by a period of intensive search in a small area [Bénichou et al., 2006]. The movement reflects a ballistic relocation to a far away area and generally the search is turned off during this relocation.

To formalize IS:

- Move to a neighboring site with probability p
- Relocate ballistically off in random direction, with $1 - p$

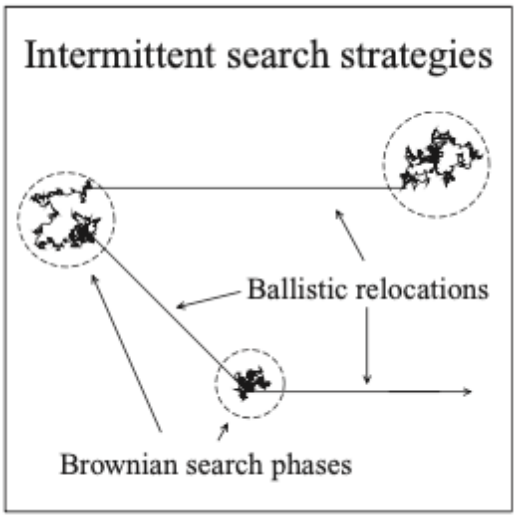

Figure 3.4: Intermittent searches consist of two phases: a search phase alternating with a relocation phase. Figure from Viswanathan et al. [2011].

Utilizing the known efficiency of Lévy walks for movement, studies have shown the lévy-modulated intermittent searches (LWIS) are more advantageous compared to ballistic relocations [Lomholt et al., 2008]. The resulting Lévy walks reduce oversampling and further optimize the search strategy in the situation of very sparse rewards.

To formalize LWIS(μ):

- Lévy Walk to move throughout the maze

- Intensive Search when it hits the boundary

It is important to note here that the above described models are better treated more alike than different even though they differ in the statistical properties of the movements shown. Studies have shown one form of movement behavior can be understood within the context of another, or there have been hybrid models of CRWs and Lévy walks, and such. Nature is full of sub-optimal mechanisms and the evolutionary basis of their origins are continuously being studied. Lévy walks and Brownian motion are merely idealizations, it is not wise to expect to observe perfect Lévy walks or perfect random walks in real organisms. There is no theoretical argument to expect Lévy walks in all biological organisms. The structure of the environment also plays a huge role in the efficiency and applicability of the above stochastic search models [Viswanathan et al., 2011].

Chapter 4

Mouse Maze Dataset

In this chapter, we describe an experiment conducted by Rosenberg et al. [2021] which resulted in a rich dataset on animal exploration in a sufficiently complex maze. We describe certain physical features of the maze that are important to keep it noise-free as much as possible and give us a controlled setting to analyze while keeping any interference to the naturalistic motion or drive to navigate as low as possible. We then state some of the findings from original analysis of the dataset that are relevant to our study.

4.1 Experiment

The maze is an enclosed complex labyrinth, as shown in Figure 4.1. A short tunnel offers free access to a maze, a wide network of corridors. The home cage has bedding and food. The animal's movement in the maze is recorded from below.

The logical structure of the maze is a binary tree, with 6 levels of branches, from home to 64 endpoints (end nodes). The levels are numbered 0, 1, ..., 6 where level 0 is the central point of the maze and at level 6 are the leaf nodes (Figure 4.2). At leaf nodes, the only action animals can take is to return back to its parent node at level 5. One of the 64 endpoints of the maze is fit with a water port (sometimes referred to as reward port or reward node). After activation by a brief poke by an animal, the port delivers a small amount of water, followed by a 90s time-out period.

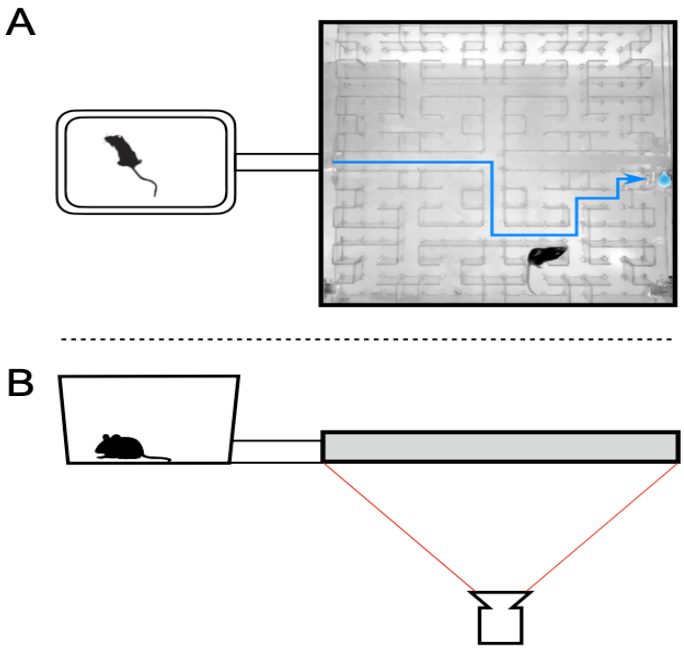

Figure 4.1: The maze environment. Top (A) and side (B) views of a home cage, connected via a tunnel to the labyrinth. Figure from Rosenberg et al. [2021].

All observations were made in darkness during the animal's subjective night. There were two groups of animals: First, 10 rewarded animals who were mildly water-deprived before the experiment, had access to food in the home cage but water only through the water port. Second, 10 unrewarded mice who had free access to food and water in the cage, and received no water from the port in the maze. One unrewarded animal did not pass between home and the maze enough and was discarded from further analysis.

Each animal was recorded continuously for 7 hours and there was no human interference whatsoever. The animal was free to move between the cage and the maze as it wished.

We retain the definitions of bout (each foray into the maze from entrance to exit through home), step (transition from one node to another), trajectory (sequence of nodes visited one after another), node (a node in the binary tree - a junction or an end node), node sequence, as per the original study. Three sample trajectories of an animal in the maze is shown in Figure 4.3.

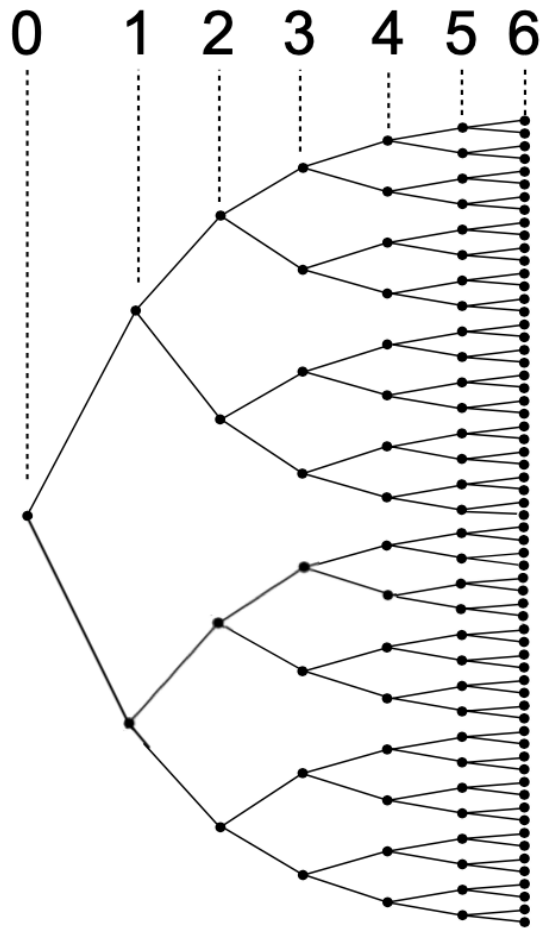

Figure 4.2: The maze is structured as a complete binary tree with 63 branch points (in levels numbered 0,. . . ,5) and 64 end nodes. Figure from Rosenberg et al. [2021].

Rotation A rotation experiment on a subset of the animals to help gain some insights about animals' learning in the maze and if the physical or sensory cues such as glue odor, urine trail have anything to do. Once the animals have been exposed to the maze for several hours, the experimenters rotated the maze by 180 degrees. If the animals did follow the odor to reach the water port, they would travel to the original water port node which lies opposite to the current water port location now, as shown in Figure 4.4. On the other hand, if they learned the sequence of turns, they would reach the right water port node which is at the same location as before.

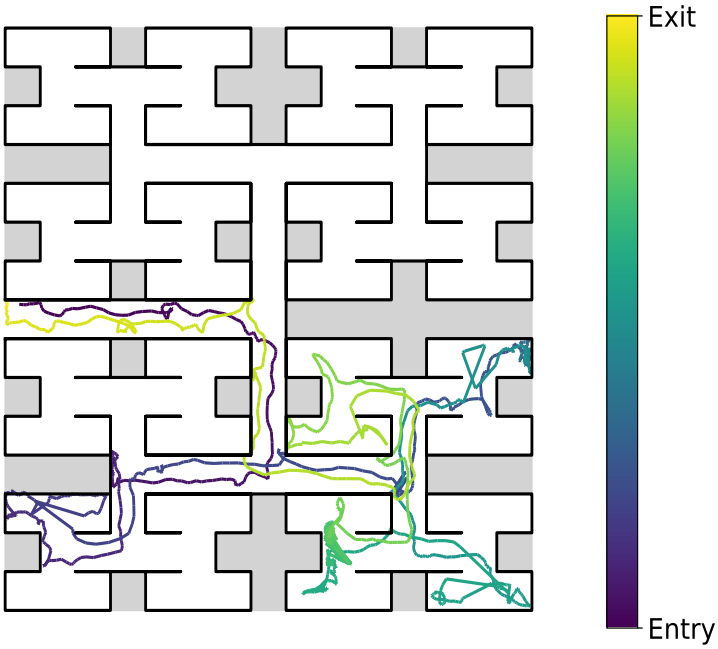

Figure 4.3: The maze environment with a sample trajectory of an animal from entry to exit plotted on it.

4.2 Maze Construction Features

1. **Possibility of using external light as cues**: The maze is built to limit the passage of external light through walls, floor and ceiling as much as possible. The bottom and walls of the maze were constructed of black plastic that is transparent in the infrared. The room is kept dark except for infrared illumination for recordings and the walls and floor are opaque for animals. Even if the animal finds enough light, the goals (water port or the home cage) are invisible within the maze except from the immediately adjacent corridor. A lot of previous studies have concluded the use of light, odor and other sensory cues by rodents to help navigate a place. The current design of this maze tries to restrict those options [Rosenberg et al., 2021, Munn, 1950].

2. **Symmetry**: The maze is constructed with maximal symmetry around the center point. If the maze wasn't symmetrical, the learning of reward or home

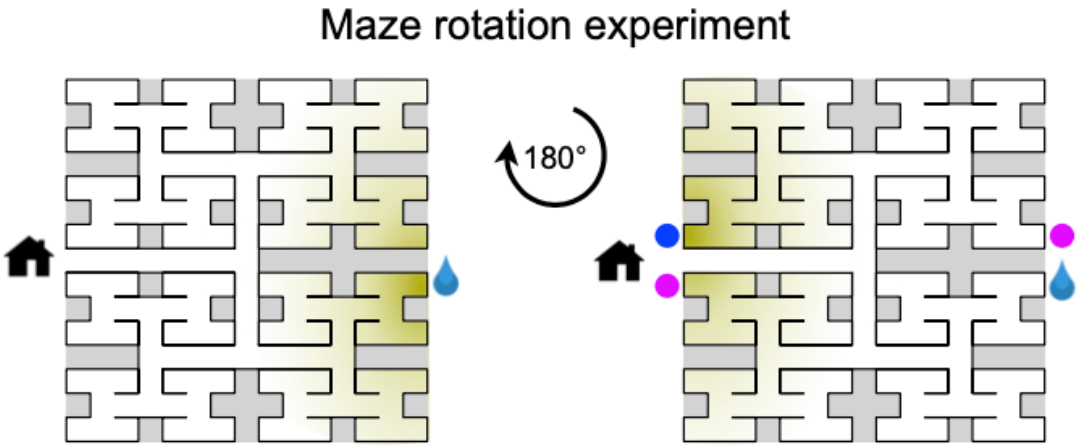

Figure 4.4: The rotation experiment. Figure from Rosenberg et al. [2021].

path could be partially attributed to sensory cues (even if no light, animals can possibly use the kinesthetic intelligence and the asymmetry in surroundings to get a sense of their location in the maze space). All the junctions at a level are visually and geometrically identical and the two branches out of a parent node are identical locally. The two children nodes of a parent can be distinguished in terms of direction if the physical maze is viewed from the top, that is, at global level but not locally.

4.3 Behavioral Insights

Here we present some of the general behavioral insights as analyzed by the authors of the original study. The two groups of animals have the exploration phase in common while the animals exposed to the water port exhibit certain goal-oriented learning and structure in their navigation behavior.

Insights common to all animals

1. **Animals initially hesitate to enter the maze**. Initially animals would not cross the main corridor from home to the maze or do frequent hesitant entries until level 1 before returning to the cage. After a few such entries,

the animals seem to feel more comfortable and start going further into the maze and do longer bouts into most or all of the end nodes.

2. **Exploration occupies a huge fraction of the animal's time in the maze**. Animals could have chosen to stay at the home cage or next to the water port, rather they seem to keep exploring throughout the night (Figure 4.5). Exploration is defined as all periods in which the animal is in the maze but not a direct path to water or to the home cage. We retain this definition of exploration in our analysis as it is.

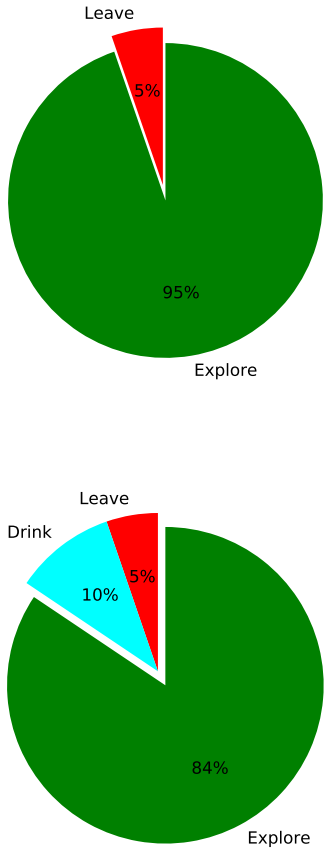

Figure 4.5: Exploration constitutes a significant proportion of the animal's activity. Pie charts show time spent in each mode averaged over animals and duration of the experiment. Figure from Rosenberg et al. [2021].

3. **Animals are efficient at exploring**. As seen before, animals tend to explore the space throughout the night that could be for a number of reasons: look for hidden reward [Berlyne, 1955] or satisfy their curiosity of being in a new space [Berlyne and Slater, 1957] or perhaps they are simply acting upon their hyperactive tendencies [Jones et al., 2017]. We measure how quickly the animals are able to cover the area of maze by being efficient about not visiting areas they just visited and avoid getting stuck. We again use the definition of efficiency of exploration as defined in the experimental study: number of visits N_{32} required to survey half the **end nodes**, that is:

$$E = \frac{32}{N_{32}} \tag{4.1}$$

An optimal agent would visit the end nodes systematically and without repeats, thus covering all the end nodes at just 64 visits giving rise to an efficiency of 1.0. Behavior like this would require perfect memory to track every node. On the other hand, a random agent makes 3 decisions at every junction without any notion of recency leading to highly inefficient behavior (E = 0.23 calculated by simulating random agents). The mice show an efficiency of about 0.39 with very little variability among them and interestingly this lies in the middle of the efficiency of a perfect agent and that of a random agent (Figure 4.6). We believe this observation holds the key to possibly uncover a certain fundamental principle of animal navigation.

4. **Animals exhibit strong biases about where to go further on arriving at a junction**. They have a strong bias to keep moving **forward** ($P_{SF} = 0.8$, $P_{BF} = 0.8$) when they face a branch (B) or a stem (S), that is, a fixed probability of going back at any point in the maze. They also have a tendency to **alternate** ($P_{SA} = 0.8$), that is, if they took a left turn at the previous junction, they seem to prefer taking right turn at the next junction. Lastly, they also have a mild preference ($P_{BS} = 0.65$) for **going into the branch** instead of going straight in a corridor. The 4 biases are depicted in Figure 4.7. As with efficiency of exploration, the study found a remarkable degree of

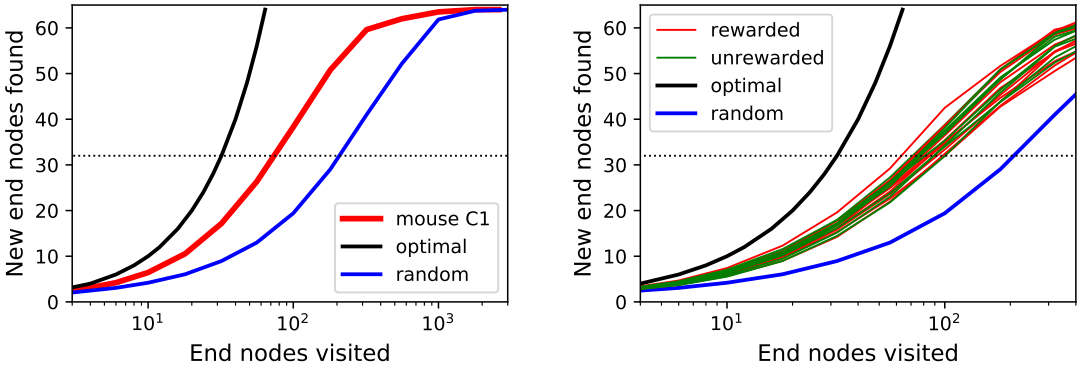

Figure 4.6: Animals are quite efficient at exploration w.r.t. space coverage. Exploration efficiency as defined in text plotted for one animal (Left) and all 19 animals (Right). Figure from Rosenberg et al. [2021].

consistency in these local rules of behavior as well. We will exploit some of these biases later in this thesis and also try to interpret them slightly differently to help correlate some of these with the ecology literature and develop a model to explain navigation in the maze. An important point to note here is that some of these biases (P_{SA}) are not realizable when the maze is viewed as a symmetric binary tree with no difference in the two children at any node but only carry meaning when the maze is viewed as an expanded physical space as it actually is.

5. **Animals visit certain end nodes a lot more frequently than others**. While the animals vary quite a bit in showing this preference, the effect remains consistent and significant across all animals of the two groups. Specifically, animals tend to prefer nodes that lie on the outer edge of the maze than those on the inner, by a factor of 2.2 (ranges from 1.8x to 4x across animals), depicted in Figure 4.8. This preference again is an example of a behavior variable that is only meaningful in the physical notion of the maze and not the abstract binary tree one.

   To note, the presence of water port on the outer node of the maze does not affect this bias as evidenced by the similar preference for peripheral nodes in unrewarded animals that do not have any water port.

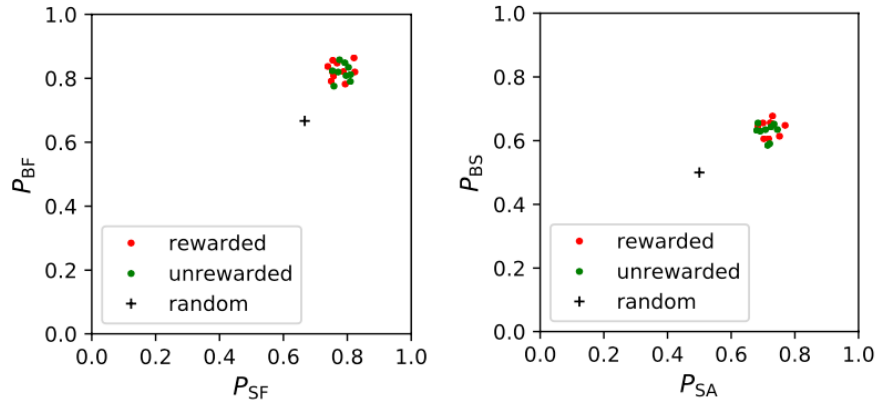

Figure 4.7: Scatter graph of the biases P_{BF} and P_{SF} (left) and P_{BS} and P_{SA} (right). Each dot represents a mouse. Cross: a random agent. Figure from Rosenberg et al. [2021].

6. **Animals go from end nodes to home by taking a direct path**. "Home runs" are direct paths without reversals that take the animal to home. Home runs have a special significance in ethology literature because factors such as fear or uncertainty make animals remember a route to escape, which in this case is the route to the maze entrance (home) [Rosenberg et al., 2021, Tchernichovski et al., 1998, Fonio et al., 2009]. As authors of the original study find after day analysis, animals do not seem to practice the home path explicitly by taking incremental steps and gradually building it, neither they retrace their path that they took while entering the maze (Figure 4.9). We will later try to argue how these home runs do not necessarily denote learning of the six decisions.

   We will see similar behavior in section 4.3 on insights from rewarded animals towards navigation to the water port as well. Animals learn to take direct paths from home or even from deeper end nodes to water port very soon in the experiment. How do the animals navigate when they perform direct paths to the water port or to the exit? The original study doesn't touch upon it but it does point us towards certain construction features of the maze (covered in section 4.2 earlier) and the observed behavior that can help us constrain some answers to the question and discard others.

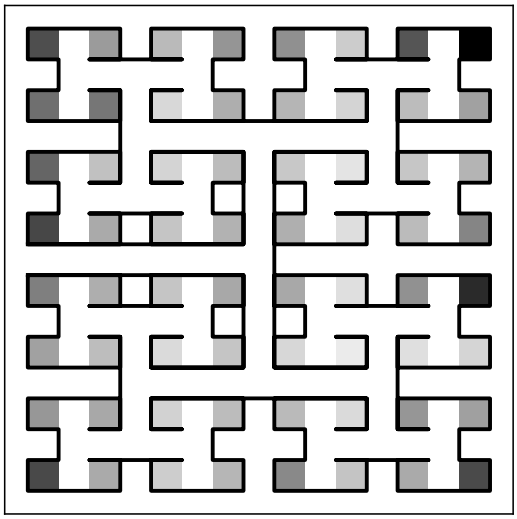

Figure 4.8: Preference for outer end nodes during exploration. The number of visits to different end nodes encoded by a gray scale, for unrewarded animals. Darker nodes mean higher preference. Figure from Rosenberg et al. [2021].

Insights specific to rewarded animals

1. **Animals start visiting water port at a higher rate suddenly**. Animals exhibit a sudden change in behavior with respect to rate of visits to water port at a certain point of life, long after discovering the water port for the first time (Figure 4.10). It is also followed by a higher rate of longer direct paths from anywhere in the maze to the water port ("sudden insight"). The authors observe this behavior in about 5 out of the 10 rewarded animals and a gradual but similar performance change for remaining 5 animals.

2. **Animals use physical cues to learn the goal path, at least in the beginning**. As part of the rotation experiment, the maze is rotated by 180 degrees after exposing the animals to it for a few hours and the original water port node get shifted to its mirror image location. Only 1 out of 4 animals confuses the two locations for the first trip but quickly gets back to the correct reward path to the new water port node and the rest 3 animals went straight to the correct water port before ever visiting the image location. However, for the first hour following it, the visit frequency to the image

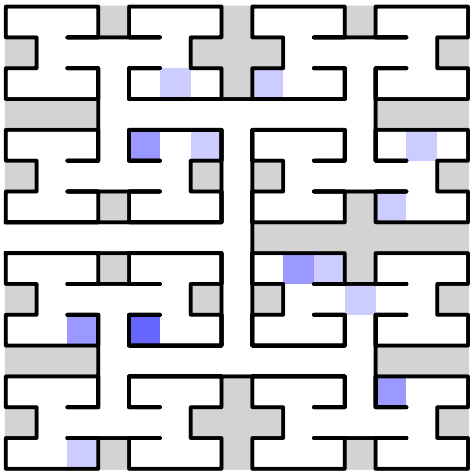

Figure 4.9: Learning of the home path. Locations in the maze where the 19 animals started their first home run. Darker color indicates locations used by 2 or 3 animals. Figure from Rosenberg et al. [2021].

location increased by 1.8x compared to before the rotation and the reward node visits and reward rate both declined. This led the authors to conclude that navigation does not strictly depend on the physical cues but animals do notice a change in these cues which is reflected by the decreased reward rate and increased visits to image location.

4.4 Significance

Most animal behavior studies employ far-simpler tasks than the current experiment that involve learning between left or right, or similar. These behaviors have a complexity of 1 bit or less, and often animals can learn these associations after a single trial. The tasks a mouse performs in the maze are far more complex and are in a close to naturalistic setting. For example, the path from the maze entrance to the water port involves 6 junctions, with 3 options at each. The current binary maze has 64 branches of equal length with only one leading to the water port. The probability of animal making the choice that helps it move forward towards the water port or the exit becomes increasingly low.

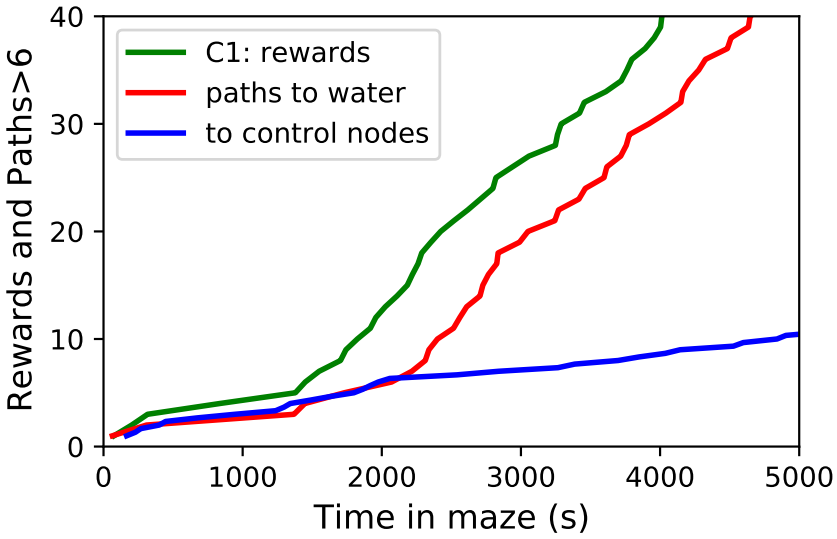

Figure 4.10: The bump in reward rate at 1350s depict sudden changes in behavior. For C1, plotted are the cumulative number of rewards; of long paths to water (red); and of similar paths to 3 control nodes (blue, divided by 3). Figure from Rosenberg et al. [2021].

The absence of any human interference and trial structure is another big aspect of the current experimental design significant for future animal behavioral studies. The current dataset is vastly rich in terms of behavior syllables and complexity. The rich nature of this dataset is also prone to a lot of vague and ambiguous hypotheses about the structure of navigation in the maze because of the human tendency to try to generalize from only a few observations in day-to-day life. There is also very limited amount of work on experiments in such highly-constrained physical space in the field of ethology or human behavior or artificial intelligence. This raises the importance of systematically studying the behavior even more and only then making any conclusions to help come up with valid hypothesis. Further, this also signifies how much we can learn about animal behavior by conducting behavior experiments in complex environments with many choice points. As pointed out by the experimenters, with the rising advances in computer vision and user-friendly tools for single and multi-animal tracking, we are already seeing some [Alonso et al., 2020, Grobéty and Schenk, 1992, Nagy et al., 2020,

Wood et al., 2018] and should expect more experiments of similar complexity in the near future.

Chapter 5

Models of Exploration in Mouse Maze

This chapter describes the set of methods used in the project for modeling and analysis. We first describe an existing model of animal exploratory behavior observed in the Mouse Maze dataset in section 5.1 and later introduce our modeling setup in section 5.2. We then present the results of our model in section 5.3 and compare the two using a set of evaluation metrics, and then provide an interpretation of our results in section 5.4.

5.1 Biased Walk

Rosenberg et al. [2021] proposed a simple yet an effective model of mice exploratory behavior using decision turning biases shown by animals as highlighted in their data analysis in chapter 4. Biased Walk is a variation of the Correlated Random Walk where the current decision is related to the previous turn in the sense that the previous turn determines the probabilities of three possible choices at the current turn. This model was not presented as to be the best model of mice behavior but it nonetheless serves as a good candidate model to compare our results against.

The correlations in this biased walk are built upon the below 4 turning biases (hence, we label it as BiasedWalk4 model) calculated using the number of actions

taken at a particular type of junction:

1. A strong preference to go forward when the animal faces a junction. The two types of junctions are distinguished as stem and branch based on how the animal is entering the junction (Figure 5.1). The two biases are labeled P_{SF} (forward through stem) and P_{BF} (forward through branch).

2. A strong preference (P_{SA}) to take alternating turns left and right rather than repeating the same direction turn at junctions. Note that the left and right turns are defined globally and not with respect to where the animal is coming from.

3. A slight preference to take a branch that leads the animal out of the maze, that is, from a deeper level to a shallower level in terms of navigating on the binary tree (P_{BS}: from branch-to-stem).

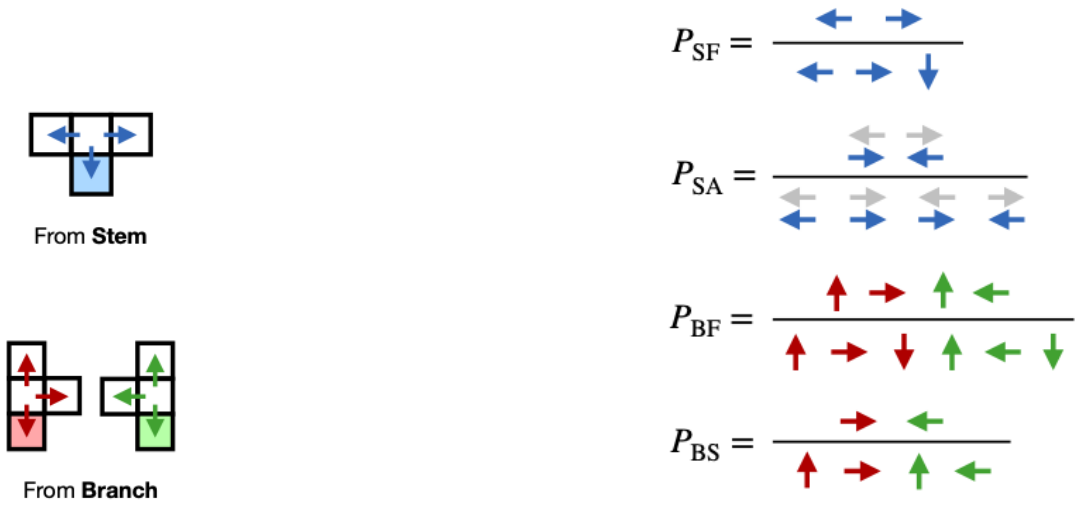

Figure 5.1: Definition of four turning biases at a T-junction based on the ratios of actions taken. For details, refer Rosenberg et al. [2021].

The four biases are depicted in Figure 5.1. The intuition behind the model is as follows: The forward biases keep the animal from re-entering the territory it has already covered, and the branch-to-stem makes the animal go to different parts of

the maze. This is different from a purely random walker which will keep getting stuck in the tips of the tree and oversample an area of the maze.

As we will see in section 5.3 on results, BiasedWalk4 performs quite well in capturing the exploration efficiency and some of the preferences of animals. This was also an indication that the underlying search algorithm of animals possibly does not need a global memory of places visited and can be explained by purely local turning rules to a large extent, as inferred by the authors of the original study. Although the model replicates a good amount of the pattern of preferences and efficient to a certain degree, the model in its current form is tough to interpret and the 4 parameters seem highly specific to this maze. The model does not provide a full description of the animal's search strategy.

5.2 Temporally-Extended ϵ-Greedy

After a thorough review of ecology literature on foraging, we found Lévy walks are known to optimize search efficiency in particularly sparse-reward environments [Viswanathan et al., 1999]. Lévy walks in animal foraging are random walks built with directional persistence to enable longer steps. Further, an intermittent strategy which alternates an intense search phase with Lévy walks (section 3.2) is claimed to have higher search efficiency than just the Lévy walks [Lomholt et al., 2008, Oshanin et al., 2009] in various biological species. After an iterative process of building a handcrafted model of the navigation behavior and closely analyzing the behavior data, we hypothesized one component of the behavior is the movement around the corridors and another is the intensive search in a sub-quadrant when they hit an end (Figure 5.2). Note that this does not necessarily imply they "switch off" the search component when they are Lévy-walking but that they tend to search more intensively in the corresponding sub-quadrant area when hit they an end node.

On the other side of the spectrum, in deep reinforcement learning, Dabney et al. [2020] recently proposed an exploration framework, temporally-extended ϵ-greedy, built upon the properties of ϵ-greedy with an addition of temporal persistence to it.

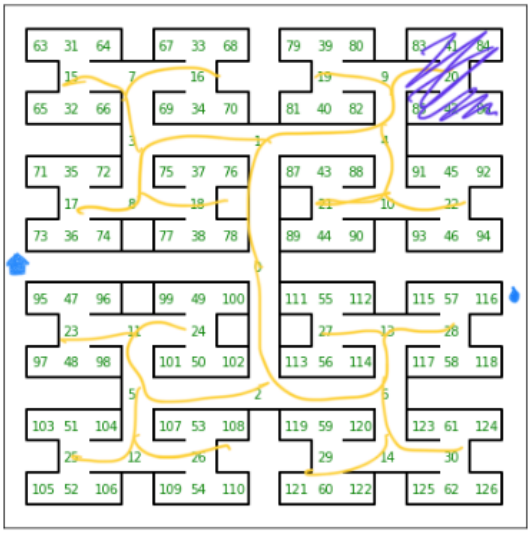

Figure 5.2: The hypothesized two components of behavior: efficient movement and intensive search.

By building upon ϵ-greedy, temporally-extended ϵ-greedy retains the convergence properties of vanilla ϵ-greedy without the need for rigorous proofs or to modify the greedy policy.

As we can see, the principle of temporal persistence bears a direct correspondence with the directional persistence that motivated the correlated random walk and Lévy walk models in ecology literature (section 3.2). Because of this strong resemblance, temporally-extended ϵ-greedy seems to be a promising approach to study the exploration behavior, if we can adapt it to our maze environment.

Temporally-Extended ϵ-greedy Temporally-extended ϵ-greedy replaces actions with temporally extended sequence of actions, or options, in standard ϵ-greedy. Options provide a way to temporally abstract away the decisions at each time step and instead help encode the long-term intention. Using options helps us learn behavior at multiple time scales and by appropriately defining a set of options, we can "align" the exploratory behavior of an agent with a given environment and control the nature of exploration. However, learning a set of options for an environment automatically remains a challenge in the field. We overcome it by defining them instead by hand and leave the work on the discovery of options in

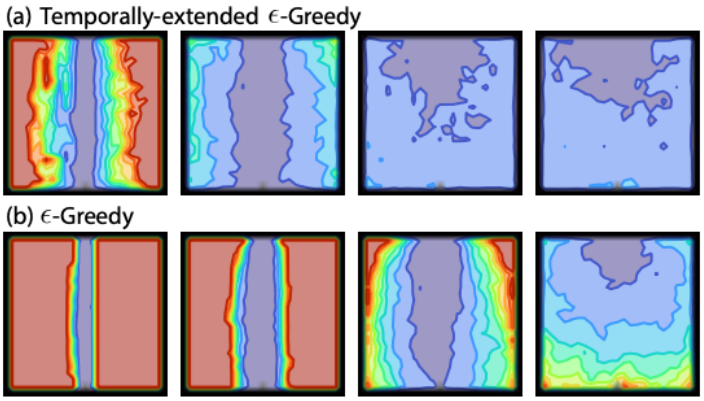

Figure 5.3: Average first-visit times comparing ϵ-greedy approach and its temporally-extended version. Blue represents fewer steps to first-visit and red states rarely or never seen. Figure from Dabney et al. [2020].

this maze and similar environments to future.

A temporally-extended ϵ-greedy exploration strategy depends on choosing an exploration probability ϵ, a set of options O, and a sampling distribution p with support O. Then, on each step:

- With probability ϵ, sample an option $w \sim p(O)$ and follows it until termination.
- With probability $1 - \epsilon$, follow the current policy π for one step.

For purely exploration settings such as ours, we set ϵ to 1.0 which removes the need for learning a policy π. The maze in fact does not offer any external reward. To adapt temporally-extended ϵ-greedy to our setting, we want to ideally capture the hypothesized Lévy walk intermittent search strategy in our set of options.

The idea is to use options to encode extended sequences of actions, which corresponds to longer paths in a purely spatial setting (Figure 5.4). Lévy walks sample a step of length n and continue moving in the same direction for n steps. Given the structure of our maze and the walls, we first have to define the notion of "directional persistence" for our environment.

We define "moving in the forward direction" as taking alternate turns at subsequent junctions into the maze since an alternate path in our maze corresponds

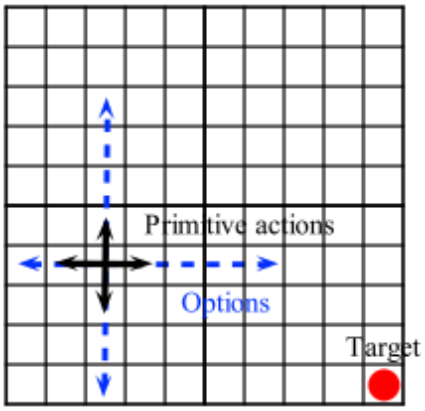

Figure 5.4: Primitive actions and options in a spatial grid world setting. Figure from Fruit et al. [2017].

to the least amount of direction or orientation changes in our setting. The forward direction defined this way corresponds to taking certain branches down the tree with a rough consideration of the forward direction on the physical maze. Similarly, we defined "moving in the backwards direction" as the opposite of moving forward, taking actions as going out of the maze from deeper to shallower levels. This backwards direction is equivalent to going up the tree. Finally, to define temporally-extended random movements, we consider all the possible paths in the maze without any consideration of forward and backwards direction. This corresponds to having all the branches up and down the tree as choices.

We formalize the problem of exploration in the current MDP as following:

- **States**: The set of states constitutes all the 63 turning points, all the 64 end nodes in the maze and 1 home node - a total of 128 states.
- **Primitive actions**: At home, the only action is to go to 0. At end nodes, the only action is to go back. At all the other nodes in the maze, there are 3 actions available: go-further-left, go-further-right or go-back.
- **Transitions**: The transition probabilities in the current MDP are entirely deterministic. Taking left at a state where left action is available, it will always go left.

- **Reward**: There is 0 reward throughout the maze and hence no value propagation. Therefore, we do not also attempt to learn a policy or define a discount factor here.

- **Options**: Using the above definition of longer straight paths, we construct options of lengths 1 to 12 (since the maximum distance between any two states in in the maze is 12) for each state as following:

    1. **For states within the maze**: We hypothesized animals to be using Lévy walks to move around the maze. Lévy Walks are encoded as long sequences of length n in forward and back directions (Figure 5.5).

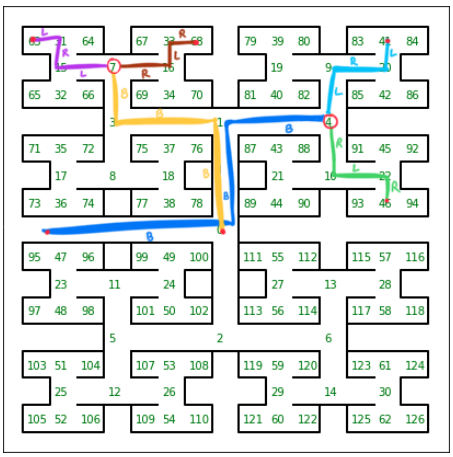

Figure 5.5: The set of Lévy Walk options of length 3 at two states in the maze. The red **o** indicates option initiation and the red • indicates the option termination.

    2. **For states at end nodes**: Animals are hypothesized to search more thoroughly in the "corner" areas. Thus, Intensive Search options at end nodes are encoded as temporally-extended random movements of length n (Figure 5.6).

Lévy Walk (LW) model We use the zeta distribution $z(n) \sim n^{-\mu}$ with $\mu = 2$ to first sample a step-length n. Then we sample one option of all n-length options available at that state in a uniform manner. Zeta distribution is one heavy-tailed distribution widely known to optimize search efficiency in ecology [Viswanathan et al., 1999].

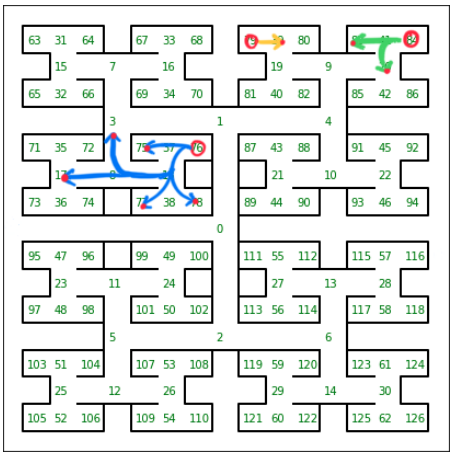

Figure 5.6: The set of Intense Search options of length 1 (yellow), length 2 (green) and length 4 (blue) at three different end nodes in the maze. The red **o** indicates option initiation and the red • indicates the option termination.

A simple interpretation of LW agent is that the agent tends to follow a certain notion of "going further" to navigate around the maze but is also aware of a different spatial scale and starts an intensive back-and-forth search in the sub-quadrant when it hits an end point.

5.3 Results

We simulated a random agent, a BiasedWalk4 agent and a LW agent each for a 25000 step long trajectory. Then we split the trajectory into bouts based on when it went to the home cage, to have our data in the same form as the original study. Below we compare the three models using efficiency and other evaluation metrics as described in chapter 4, and describe how close they come to capturing animal's preferences.

1. Lévy Walk model captures exploration efficiency of animals with respect to end nodes fairly well (Figure 5.7). Exploration efficiency is indication of how efficiently the space is being covered throughout the course of the experiment. Lévy Walks in ecology are known to optimize the efficiency of random searches in sparse-reward environments. However, we see Lévy Walk

model offers only a minor improvement over BiasedWalk4 model [Rosenberg et al., 2021].

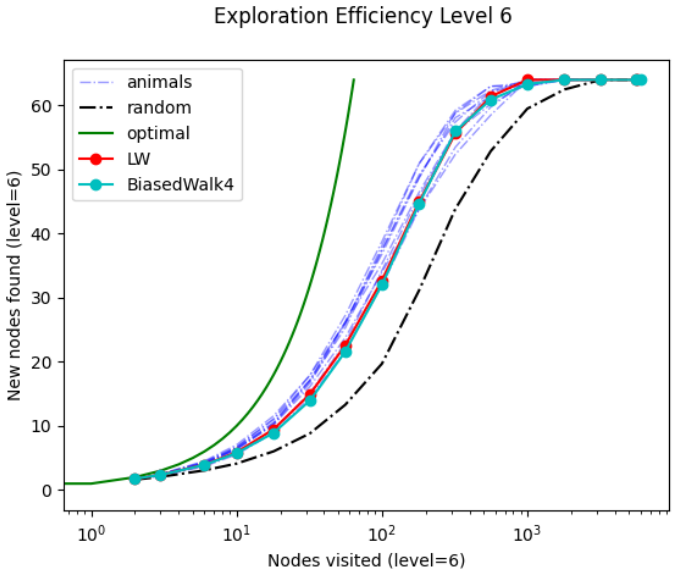

Figure 5.7: Exploration efficiency of animals is well captured by the Lévy Walk model.

2. Lévy Walk model also captures exploration efficiency better than BiasedWalk4 with respect to the inner levels of the maze (level 3, 4 and 5), indicating that movement of animals in the maze and corridors is better captured by a search efficiency-optimized model than only a turn decision optimized model (Figure 5.8). The effect is especially large for efficiency w.r.t. level 3.

3. Lévy Walk model exhibits very similar the decision biases as animals do. Since BiasedWalk4 is built upon these very four bias probabilities, it's not a fair comparison with the Lévy Walk model. Lévy Walk shows slightly higher forward bias (P_{SF}) indicating a higher directional persistence in our model as compared to animals. A more detailed analysis by varying Lévy walk parameter μ can help understanding this gap (Figure 5.9).

4. Lévy Walk model exhibits an outgoing tendency which is in the range of most animals (Figure 5.10). Lévy walk is optimized to take longer alternate straight paths and the structure of the maze is such that alternate straight

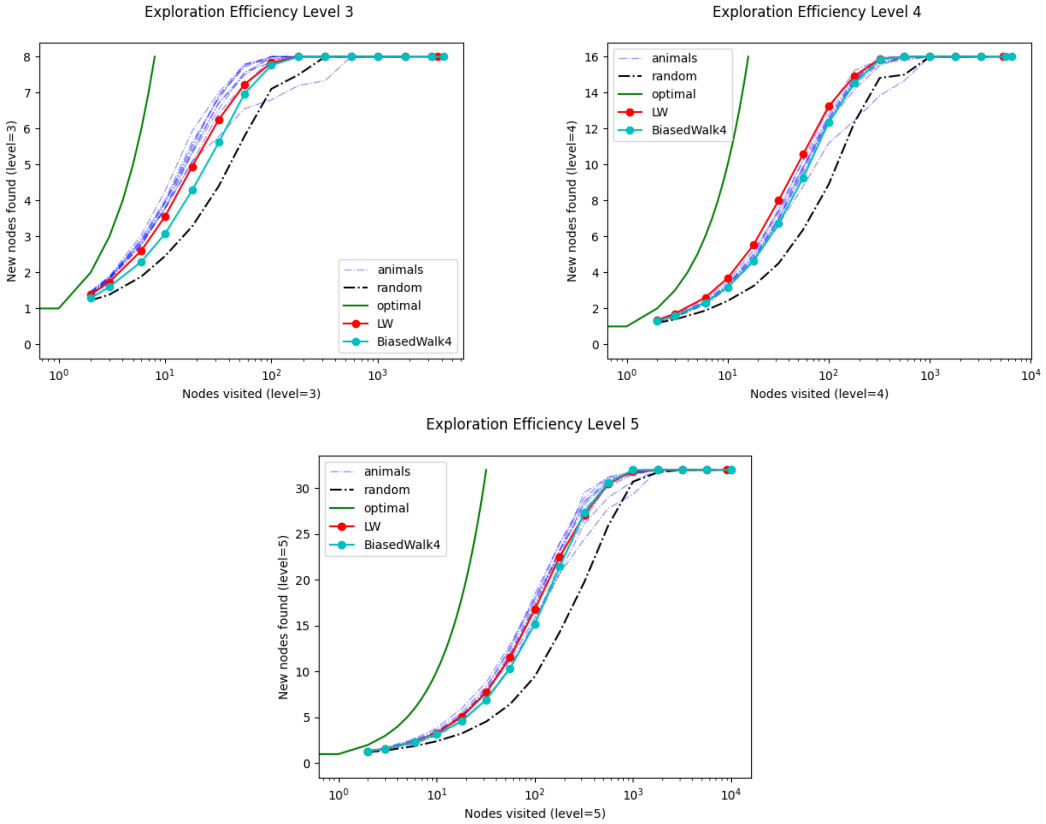

Figure 5.8: Lévy walk model captures the exploration efficiency with respect to nodes at level 3, 4 and 5 in addition to end nodes.

paths increase the probability of reaching outer nodes more. While the mean outside-to-inside ratio for unrewarded animals is 2.2, there is a large variation among them. Lévy walk model exhibits a preference of 2.28 while BiasedWalk4 shows 2.25, both lying in the range of what animals show.

5. Lévy Walk comes close to animals in terms of occupancy at different levels of the maze. We estimate the occupancy at a level as normalized number of visits to all the nodes at that level which is a rough measure of time spent at different levels. Notably animals spend a significant fraction of time in smaller scale regions of the maze (sub-quadrants). BiasedWalk4 performed very close to animals which indicates a more random component of the search in sub-quadrants. The gap between the Lévy walk and animals in occupancy of different zones in the maze indicates the, possibly noisy, component of

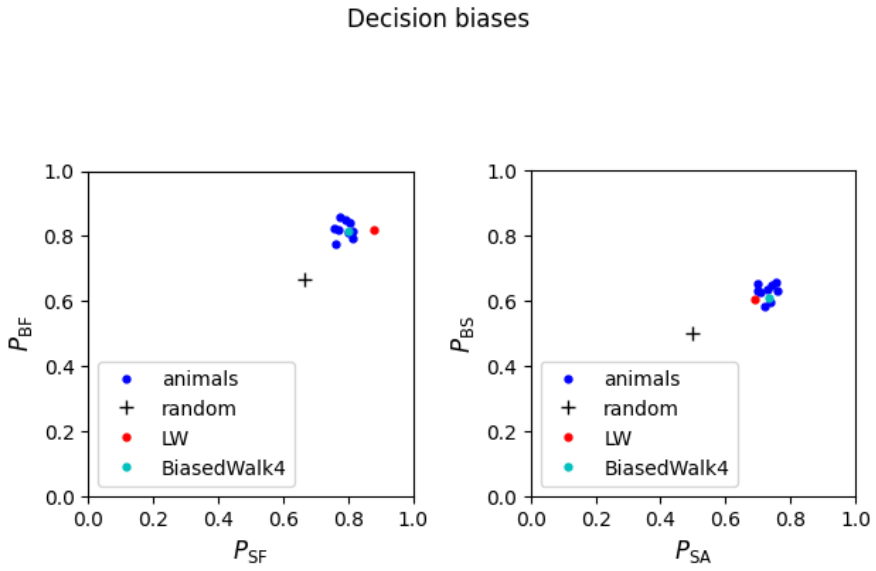

Figure 5.9: Lévy Walk model captures the decision biases quite well. For the definitions of x and y axes, please refer text.

animals' search strategies that still remains to be understood (Figure 5.11).

6. Lévy Walk model shows a variable markov depth of slightly below 4, indicating the next turn can be predicated by the current location, preceding 2 turns (Figure 5.12) and some contribution from further before. It closely resembles the markov depth shown by animals as indicated by Rosenberg et al. [2021] in their analysis. It is an improvement over the BiasedWalk4 model which only captured the contribution of previous 1 location. This finally brings down the cross-entropy (uncertainty of decisions) from 1.59 for a random agent to approximately 1.35 for a Lévy Walk agent. Lévy Walks in ecology are shown to exhibit memory effects due to the embedded directional persistence which helped us improve the cross-entropy in our model here [Viswanathan et al., 2011].

5.4 Summary

An ecologically inspired Lévy Walk Intermittent Search explains the mice exploratory behavior in the labyrinth quite well. Our agent is as efficient as animals

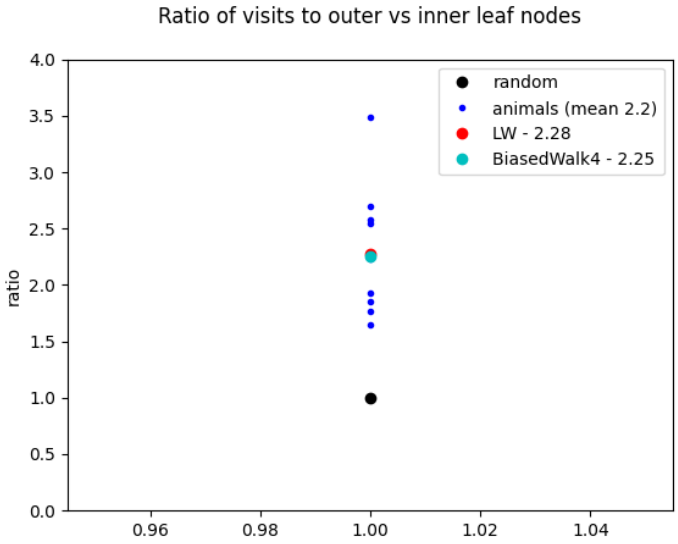

Figure 5.10: Lévy Walk model exhibits a similar outgoing tendency which is in the range of most animals.

are and it shows largely similar biases and preferences as animals do. And the search strategy can be conveniently coded in the framework of temporal abstraction in RL. According to this model and its performance across various metrics, the animals tend to be aware of different spatial scales in the maze and behave accordingly. Taking short paths and occasionally longer paths in the maze helps it cover the space efficiently, and the intensive search helps search it more in the corners. They abstract away their decisions at each junction using temporal persistence to continue to move in one direction in the maze.

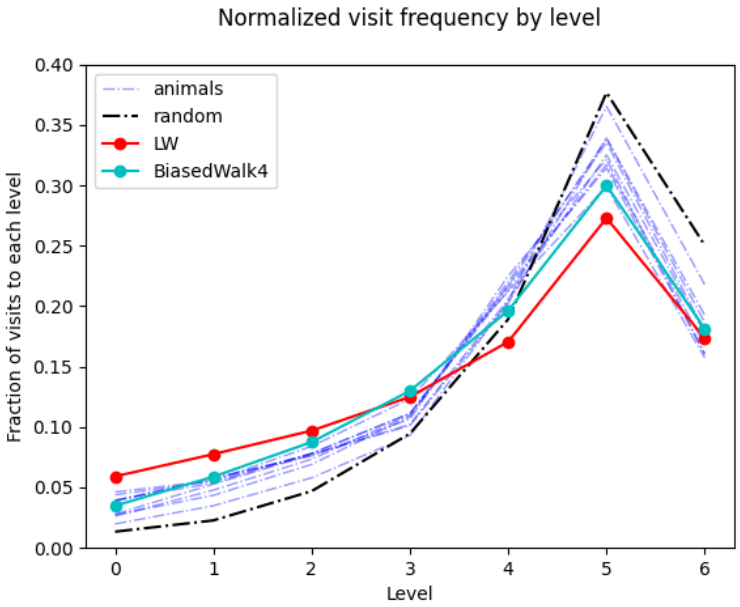

Figure 5.11: Lévy Walk performs very close to animals in terms of occupancy at different levels of the maze.

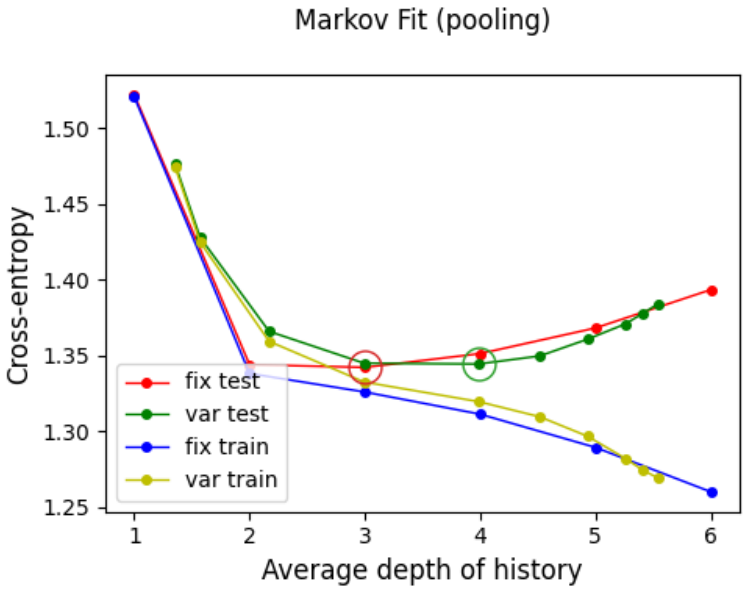

Figure 5.12: Cross-entropy of the Lévy Walk model's prediction. For details, see text and Rosenberg et al. [2021].

Chapter 6

Discussion

This chapter seeks to interpret and extend our results by drawing from the literature. We first bring perspectives of interpretation of our results in section 6.1. We then identify shortcomings and provide clues for possible future work in section 6.2.

6.1 Our Work

Exploration and search are important phenomena in the natural world around us, and yet we don't have a good understanding about what gives rise to the specific structure we observe in real-life settings. Animals and humans are limited memory beings but we navigate quite efficiently - not as good as an information-theoretic optimal agent would do but not as bad as a completely random walker. Understanding the rules that underlie animal search and learning not only can provide neuroscientists with insight into the animals, but can also provide concrete examples of biological algorithms to the ML community. Such work will also help us in the field of AI to create autonomous agents that can behave more closely like living organisms.

The exploration and search algorithms in the field of computer science or reinforcement learning are generally designed to perform optimally which could mean excessive usage of memory or employing systematic search that animals clearly don't possess or exhibit. For example, count-based algorithms try to store how many times a state has been visited [Bellemare et al., 2013]. Ecoffet et al. [2021]

tracks what parts of the environment have been recently or partially explored. Zhang et al. [2021] tracks boundaries where it left the previous search. Exploration by animals in discrete choice tasks has been studied quite extensively but rarely in sequential settings. But at the same time, exploration in sequential settings in biological agents has rarely been explored and no existing MDP algorithm has been able to capture exploratory patterns in animal movement in a complex naturalistic setting. Animals have certain biases, limited memory and physical energy constraints and these all interact in a complex manner to give rise to their behavior in the natural world.

In ecology, it's been reported across species to have efficient random search wired into them to help them hunt and forage. The animals show scale-free dynamics in their movement patterns and adapt their strategy according to the food density in the environment Viswanathan et al. [2011]. The search patterns also exhibit certain fractality and short-term memory effects Ferreira et al. [2012]. Further, the complexity of maze geometry and experience are known to affect the locomotive behavior of animals. For example, Uster et al. [1976] observed in their experiments with a hexagonal maze that the locomotion speed was highest in straight sections, decreased at corners and branchings. The change in locomotion behavior could also reflect a shift of the behavior towards a more detailed inspection of interesting parts of the maze (corners, alleys). We take the search strategies from ecology literature and model them in an RL framework, and show that this behaves much like animals in one rich naturalistic setting of a labyrinth through a ton of evaluation metrics. While we do not characterize the learning and changed in the behavior of mice as they gathered more experience throughout the course of experiment, we found that our model explains the efficiency and captures the preferences of an average mice exploring the labyrinth quite well when there's no reward. The principles of directional persistence in ecology and temporal persistence in RL are equivalent which allowed us to use the framework of temporal abstraction in RL directly. We found that animals exhibit super-diffusive behavior and leverage temporal persistence to navigate the maze rather than making decisions at each intersection. The animals have an efficient movement component based on

Lévy walks, and an intensive search component in the corners. Our study provides a new perspective on Lévy flight foraging and opens new avenues for investigating the interaction between exploration dynamics and the environment.

6.2 Outlook and Shortcomings

We provide a mechanistic view of the exploration behavior that animals seem to be exhibiting by capturing many of the qualitative evaluation metrics. Since the animals keep exploring throughout the night, our model does not provide any indication to what really is the intrinsic reward structure that animals are acting upon. Recent work by Ashwood et al. uses an Inverse Reinforcement Learning (IRL) approach to reverse-engineer the animal's behavior and infer the underlying intrinsic reward functions. Further, since we only focus on the exploratory component of the behavior, the work on characterizing animal learning in the maze remains. Does the exploration structure remain the same when we introduce a reward? How does learning of reward affect their paths and internal states? How to infer and switch between "explore" state, "are lost" state, "towards goal" or "towards home" state? Few recent studies have successfully used the GLM-HMM approach to infer animal's internal states in simpler tasks [Ashwood et al., 2022, Coen et al., 2014] but it remains to see if a similar approach could work in our environment.

Our modeling attempt indicated an ecologically-inspired search efficiency strategy explains the efficiency and preferences of mice exploration in the labyrinth quite well. But to encode it in the general framework of temporal abstraction in reinforcement learning, we handcrafted options for the two components of the exploration - movement and search. In particular, we had to provide a description of what a straight path could look like in the environment and how to search the corners once in a smaller zone of the maze without getting stuck. While the framework and the strategy remains general, the designing of options limits our ability to extend the conclusions to other environments readily. Thus we ask, how can an agent learn notions of direction in a space? We can further ask if this behavior

and a similar search strategy would be seen in human mobility patterns. Garg and Kello [2021] report efficient Lévy walks in humans while they are playing a virtual reality game set in a mountainous region. But the effect in more complex and day-to-day environments remains to be understood.

Finally, is there a way to infer the underlying diffusion process using the trajectory data itself? It's well accepted in the ecology field that the search strategies can be composed of a number of random search processes and it is rather difficult to distinguish them using available statistical methods [Viswanathan et al., 2011, Palyulin et al., 2014]. The discretization granularity of the trajectory data can have an affect on our conclusion of the underlying search process [Edwards et al., 2007]. Here we simulate a Lévy walk model which matches the exploration efficiency of animals very well but this could very well be a result of our specific modeling setup and doesn't rule out alternative step-length distributions. A rigorous analysis by performing goodness-of-fit tests could provide a more accurate description.

Bibliography

Alejandra Alonso, Jacqueline van der Meij, Dorothy Tse, and Lisa Genzel. Naïve to expert: Considering the role of previous knowledge in memory. *Brain and neuroscience advances*, 4:2398212820948686, 2020.

Susan Amin, Maziar Gomrokchi, Hossein Aboutalebi, Harsh Satija, and Doina Precup. Locally Persistent Exploration in Continuous Control Tasks with Sparse Rewards. *arXiv:2012.13658 [cs]*, June 2021a. arXiv: 2012.13658.

Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A Survey of Exploration Methods in Reinforcement Learning. *arXiv:2109.00157 [cs]*, September 2021b. arXiv: 2109.00157.

Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. In *Advances in Neural Information Processing Systems*.

Zoe C. Ashwood, Nicholas A. Roy, Iris R. Stone, The International Brain Laboratory, Anne E. Urai, Anne K. Churchland, Alexandre Pouget, and Jonathan W. Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, February 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-01007-z.

RPD Atkinson, CJ Rhodes, DW Macdonald, and RM Anderson. Scale-free dynamics in the movement patterns of jackals. *Oikos*, 98(1):134–140, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret

bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Frederic Bartumeus, M G E da Luz, Gandhimohan M Viswanathan, and Jordi Catalan. Animal search strategies: a quantitative random-walk analysis. *Ecology*, 86(11):3078–3087, 2005.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

O Bénichou, C Loverdo, M Moreau, and R Voituriez. Two-dimensional intermittent search processes: An alternative to lévy flight strategies. *Physical Review E*, 74 (2):020102, 2006.

Daniel E Berlyne. The arousal and satiation of perceptual curiosity in the rat. *Journal of Comparative and Physiological Psychology*, 48(4):238, 1955.

Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

DE Berlyne and J Slater. Perceptual curiosity, exploratory behavior, and maze learning. *Journal of Comparative and Physiological Psychology*, 50(3):228, 1957.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Alex Braylan, Mark Hollenbeck, Elliot Meyerson, and Risto Miikkulainen. Frame skip is a powerful parameter for learning to play atari. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Philip Coen, Jan Clemens, Andrew J. Weinstein, Diego A. Pacheco, Yi Deng, and Mala Murthy. Dynamic sensory cues shape song structure in Drosophila. *Nature*, 507(7491):233–237, March 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13131.

Vincent D Costa, Andrew R Mitz, and Bruno B Averbeck. Subcortical substrates of explore-exploit decisions in primates. *Neuron*, 103(3):533–545, 2019.

Will Dabney, Georg Ostrovski, and André Barreto. Temporally-Extended {\epsilon}-Greedy Exploration. *arXiv:2006.01782 [cs, stat]*, June 2020. arXiv: 2006.01782.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992.

Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-Learning. page 8, 1998.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, February 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-03157-9.

Andrew M Edwards, Richard A Phillips, Nicholas W Watkins, Mervyn P Freeman, Eugene J Murphy, Vsevolod Afanasyev, Sergey V Buldyrev, Marcos GE da Luz, Ernesto P Raposo, H Eugene Stanley, et al. Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449(7165):1044–1048, 2007.

A.S. Ferreira, E.P. Raposo, G.M. Viswanathan, and M.G.E. da Luz. The influence of the environment on Lévy random search efficiency: Fractality and memory effects. *Physica A: Statistical Mechanics and its Applications*, 391(11):3234–3246, June 2012. ISSN 03784371. doi: 10.1016/j.physa.2012.01.028.

Ehud Fonio, Yoav Benjamini, and Ilan Golani. Freedom of movement and the stability of its unfolding in free exploration of mice. *Proceedings of the National Academy of Sciences*, 106(50):21335–21340, 2009.

Roy Fox, Sanjay Krishnan, Ion Stoica, and Ken Goldberg. Multi-level discovery of deep options. *arXiv preprint arXiv:1703.08294*, 2017.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in mdps with options without prior knowledge. *Advances in Neural Information Processing Systems*, 30, 2017.

Ketika Garg and Christopher T Kello. Efficient Lévy walks in virtual human foraging. *Scientific Reports*, 11(1):5242, December 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-84542-w.

Marie-Claude Grobéty and Françoise Schenk. Spatial learning in a three-dimensional maze. *Animal Behaviour*, 43(6):1011–1020, 1992.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Elad Hazan, Sham M Kakade, and Karan Singh. Provably Efficient Maximum Entropy Exploration. page 16, 2019.

Thomas T. Hills, Christopher Kalff, and Jan M. Wiener. Adaptive Lévy Processes and Area-Restricted Search in Human Foraging. *PLoS ONE*, 8(4):e60488, April 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0060488. URL `https://dx.plos.org/10.1371/journal.pone.0060488`.

Robert N Hughes. Intrinsic exploration in animals: motives and measurement. *Behavioural Processes*, 41(3):213–226, December 1997. ISSN 03766357. doi: 10.1016/S0376-6357(97)00055-7. URL `https://linkinghub.elsevier.com/retrieve/pii/S0376635797000557`.

Samantha Jones, Elizabeth S Paul, Peter Dayan, Emma SJ Robinson, and Michael Mendl. Pavlovian influences on learning differ between rats and mice in a counter-balanced go/nogo judgement bias task. *Behavioural Brain Research*, 331:214–224, 2017.

Leslie Pack Kaelbling. Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the tenth international conference on machine learning*, volume 951, pages 167–173, 1993.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Adil Khan, Jiang Feng, Shaohui Liu, and Muhammad Zubair Asghar. Optimal skipping rates: training agents with fine-grained control using deep reinforcement learning. *Journal of Robotics*, 2019, 2019.

John R Krebs and David W Stephens. *Foraging theory*. Princeton University Press, 2019.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Jongmin Lee, Byung-Jun Lee, and Kee-Eung Kim. Reinforcement learning for control with multiple frequencies. *Advances in Neural Information Processing Systems*, 33:3254–3264, 2020.

Daniel Y Little and Friedrich T Sommer. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7:37, 2013.

Michael A Lomholt, Koren Tal, Ralf Metzler, and Klafter Joseph. Lévy strategies in intermittent search processes are advantageous. *Proceedings of the National Academy of Sciences*, 105(32):11055–11059, 2008.

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. In *International Conference on Machine Learning*, pages 6862–6873. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Alireza Modirshanechi, Johanni Brea, and Wulfram Gerstner. Surprise: a unified theory and experimental predictions. preprint, Neuroscience, November 2021.

NL Munn. The role of sensory processes in maze behavior. *Handbook of psychological research on the rat*, pages 181–225, 1950.

Máté Nagy, Attila Horicsányi, Enikő Kubinyi, Iain D Couzin, Gábor Vásárhelyi, Andrea Flack, and Tamás Vicsek. Synergistic benefits of group search in rats. *Current Biology*, 30(23):4733–4738, 2020.

John O'keefe and Lynn Nadel. Précis of o'keefe & nadel's the hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979.

Bence P Ölveczky. Ordered randomness in fly love songs. *Nature*, 507(7491): 178–178, 2014.

Gleb Oshanin, Katja Lindenberg, Horacio S Wio, and Sergei Burlatsky. Efficient search by optimized intermittent random walks. *Journal of Physics A: Mathematical and Theoretical*, 42(43):434008, 2009.

Vladimir V. Palyulin, Aleksei V. Chechkin, and Ralf Metzler. Lévy flights do not always optimize random blind search for sparse targets. *Proceedings of the National Academy of Sciences*, 111(8):2931–2936, February 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1320424111.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-Driven Exploration by Self-Supervised Prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 488–489, Honolulu, HI, USA, July 2017. IEEE. ISBN 978-1-5386-0733-6. doi: 10.1109/CVPRW.2017.70.

Doina Precup. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst, 2000.

Doina Precup, Richard S Sutton, and Satinder Singh. Theoretical results on reinforcement learning with temporally abstract options. In *European conference on machine learning*, pages 382–393. Springer, 1998.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Matthew Riemer, Miao Liu, and Gerald Tesauro. Learning abstract options. *Advances in neural information processing systems*, 31, 2018.

Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *Elife*, 10:e66175, 2021.

Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991a.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991b.

Sahil Sharma, Aravind Srinivas, and Balaraman Ravindran. Learning to repeat: Fine grained action repetition for deep reinforcement learning. *arXiv preprint arXiv:1702.06054*, 2017.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International conference on machine learning*, pages 5779–5788. PMLR, 2019.

David W Sims, Nicolas E Humphries, Nan Hu, Violeta Medan, and Jimena Berni. Optimal searching behaviour generated intrinsically by the central pattern generator for locomotion. *eLife*, 8:e50316, November 2019. ISSN 2050-084X. doi: 10.7554/eLife.50316.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Ofer Tchernichovski and Yoav Benjamini. Part II. The dynamics of long term exploration in the rat. page 8.

Ofer Tchernichovski, Yoav Benjamini, and Ilan Golani. Part I. A phase-plane analysis of the relationship between location and velocity. page 10.

Ofer Tchernichovski, Yoav Benjamini, and Ilan Golani. The dynamics of long-term exploration in the rat. *Biological cybernetics*, 78(6):423–432, 1998.

Sebastian B. Thrun. Efficient exploration in reinforcement learning. Technical report, USA, 1992.

H. J. Uster, K. Bättig, and H. H. Nägeli. Effects of maze geometry and experience on exploratory behavior in the rat. *Animal Learning & Behavior*, 4(1):84–88, March 1976. ISSN 0090-4996, 1532-5830. doi: 10.3758/BF03211992.

Christina-Anna Vallianatou, Alejandra Alonso, Adrian Zapata Aleman, Lisa Genzel, and Federico Stella. Learning-Induced Shifts in Mice Navigational Strategies Are Unveiled by a Minimal Behavioral Model of Spatial Exploration. *eneuro*, 8(5):ENEURO.0553–20.2021, September 2021. ISSN 2373-2822. doi: 10.1523/ENEURO.0553-20.2021.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. FeUdal networks for hierarchical reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3540–3549. PMLR, 06–11 Aug 2017.

G. M. Viswanathan, Sergey V. Buldyrev, Shlomo Havlin, M. G. E. da Luz, E. P. Raposo, and H. Eugene Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, October 1999. ISSN 0028-0836, 1476-4687. doi: 10.1038/44831.

G M Viswanathan, F Bartumeus, Sergey V Buldyrev, J Catalan, U L Fulco, Shlomo Havlin, M L Lyra, E P Raposo, and H Eugene Stanley. Levy flight random searches in biological phenomena. *Physica A*, page 6, 2002.

Gandhimohan M Viswanathan, Marcos GE Da Luz, Ernesto P Raposo, and H Eugene Stanley. *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, 2011.

Laura E Wadkin, Sirio Orozco-Fuentes, Irina Neganova, Majlinda Lako, Nicholas G Parker, and Anvar Shukurov. An introduction to the mathematical modeling of ipscs. In *Recent Advances in iPSC Technology*, pages 115–156. Elsevier, 2021.

Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. In *International Conference on Learning Representations*, 2020.

Ruth A Wood, Marius Bauza, Julija Krupic, Stephen Burton, Andrea Delekate, Dennis Chan, and John O'Keefe. The honeycomb maze provides a novel test to study hippocampal-dependent spatial navigation. *Nature*, 554(7690):102–105, 2018.

Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuandong Tian. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34:25217–25230, 2021.