University of Nottingham
UK | CHINA | MALAYSIA

Energy Dependent Reinforcement Learning based on the Neuronal Mechanisms of the Olfactory Processing in Mushroom Body

Jiamu Jiang

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

09 August 2024

This too shall pass.

Abstract

The metabolic energy is crucial for neural processing related to learning, which modulates computational capabilities, neuronal quantities, synaptic connections (Chittka & Niven 2009, Striedter 2006), and long-term memory formation (Suzuki et al. 2011, Trannoy et al. 2011, Plaçais & Preat 2013). From the evolutionary perspective, these neural processes have shaped organisms' adaptive responses to environmental stimuli and enhanced survival chances. Studies have highlighted that associative conditioning extends the lifespan across various species (Morand-Ferron 2017, Wright 2011). Also, insects modulate memory types based on ecological determinants (Smid & Vet 2016).

This thesis concentrates on the olfactory nervous system in Drosophila's Mushroom Body (MB), as a structure paralleling the mammalian brain hippocampus, presenting as a model organism for unraveling memory formation intricacies (Wolff & Strausfeld 2015, Davis 2004, 2005), given its genetic accessibility and well studied olfactory processing (Aso et al. 2014*b*).

This research posits that energy constraints might represent evolutionary adaptations promoting survival and learning efficiency. By dissecting the fruit fly's learning processes—specifically regarding metabolic energy—this study aims to assess potential lifespan extensions via energy modulation during learning and to gauge the efficacy of learning under energy constraints.

We identified three adaptive reinforcement learning variations, each influenced by the energy dynamics observed in fruit flies. The first variation underscores the capability of energy-driven memory pathway regulation to augment the fruit fly's lifespan, particularly when synergized with dopamine regulation. The subsequent variation reveals that the strategy of depressing synapses linked to undesired actions demonstrates high efficiency in synaptic adjustments across both aversive and appetitive conditioning contexts. The final variation applies the energy-adaptive methods to the conventional multi-armed bandit algorithms, such as the Upper Confidence Bound (UCB) and Bayesian-based Thompson Sampling (TS), and emphasizes the capacity of energy-adaptive methods to prolong the agents' lifespan without significant sacrifice in regret.

In summary, the thesis delineates the integral role of energy dynamics in shaping and optimizing learning processes and behaviors, drawing inspiration from the olfactory learning in the MB. These findings contribute to our understanding of the nature of energy, learning, and survival.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisor, Mark van Rossum. His enthusiasm has sparked my passion for research, and pointed me to an interesting research topic. His trust and encouragement in allowing me to explore various research ideas and pursue diverse experiences in my PhD journey. Supervised by Mark has been an immensely rewarding experience, and I am eternally appreciative of his guidance.

In Nottingham, I am grateful for my colleagues in our research group: Silviu, Aaron, Hazem, Carolina, and Holin. Our thought-provoking discussions on both research and life enriched my time here. My sincere thanks to Steve and Rachel for their candid feedback, which often provided the nudge needed to maintain momentum. I'm also thankful to Adam Moss for introducing me to intriguing applications in computer vision and allowing me to delve into TPU. And to George, assisting him in teaching has been an educational experience in itself, equipping me with useful coding tricks in machine learning. His course remains one of my favorites.

At Imperial College, my gratitude extends to Claudia, who first acquainted me with computational neuroscience, sparking my interest. I'm also thankful to Petar, who not only oversaw my master's research but also imparted knowledge on Reinforcement Learning and consistently provided encouragement during my PhD.

Many thanks to my supervisor from the Turing enrichment project, Huy. His guidance on transformers, feature extraction, and clustering algorithms has been enlightening. I'd also like to thank the other PhD student in the Turing project, Andrea, his talent and productivity always inspire me.

A heartfelt thank you to my friends, Yuqi, Shuangyu, Ling, Josh, Dan, and JingJing, for their trust and friendship. My co-workers during my internship, Diana and Laurence, played pivotal roles in improving my coding and communication skills. They also bolstered my confidence, helping me overcome imposter syndrome and improving my belief in my capabilities.

To my parents, Lihong and Long. Their love and unconditional support, even during the family's most difficult period, have been my pillars of strength, giving me the courage to face challenges head-on.

Last but not least, I would like to thank my partner, Adam. His love, encouragement, and shared knowledge on topics ranging from statistics to data science have kept the spark of curiosity alive. His companionship on this PhD journey ensured there was never a dull moment.

List of Publications

Research included in this thesis:

1. Jiang, J., Foyard, E. & van Rossum, M. C. (Under Review), 'Reinforcement learning when your life depends on it: towards a neuroeconomic theory of learning', *PLOS Computational Biology*. (Related to Chapter 2).

2. Jiang, J. & van Rossum, M. C. (2022), 'Energy efficient reinforcement learning as a matter of life and death', *Cosyne*. (Related to Chapter 3).

3. Girard, M., Jiang, J. & van Rossum, M. C. (2023), 'Estimating the energy requirements for long term memory formation', *arxiv* p. 2301.09565. (Related to Chapter 2).

Excluded research:

1. Jiang, J., Smith, P. & van Rossum, M. C. (2020), 'Electro-physiology Models of Cells with Spherical Geometry with Non-conducting Center', *Bulletin of Mathematical Biology*.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Associative
ARM	Anesthesia-Resistant Memory
CS	Conditioned Stimulus
DA	Dopamine
DAN	Dopamine Neurons
DGRP	Drosophila melanogaster Genetic Reference Panel
EA-ϵ-Greedy	Energy adaptive ϵ-Greedy
EA-TS	Energy-Adaptive Thompson Sampling
EA-UCB	Energy-Adaptive Upper Confidence Bound
FT	Fixed Threshold
GABA	Gamma-Aminobutyric Acid
KC	Kenyon Cells
LTD	Long-Term Depression
LTM	Long Term Memory
LTP	Long-Term Potentiation
MAB	Multi-Armed Bandit
MB	Mushroom Body
MBON	Mushroom Body Output Neurons
MTM	Middle-Term Memory
MT	Moving Threshold
NA	Non-Associative

PI	Performance Index
PAM	Protocerebral Anterior Medial
PDF	Probability Density Function
PPL1	Protocerebral Posterior Lateral
RL	Reinforcement Learning
SEM	Standard Error of the Mean
STM	Short-Term Memory
TS	Thompson Sampling
UCB	Upper Confidence Bound
US	Unconditioned Stimulus

Chapter 1

Introduction

The brain, a significant energy consumer, utilizes glucose at a rate more than twice greater than the heart (Holliday et al. 1967). Compared to muscles, the brain's energy usage per gram is ten times more (Balasubramanian 2021). This high energy consumption of the brain mainly (50-80%) goes into signal processing and handling neurotransmitters (Attwell & Laughlin 2001, Bélanger et al. 2011). Metabolic energy plays an important role in various aspects of learning, such as computation ability (Chittka & Niven 2009), number of neurons (Azevedo et al. 2009, Yu et al. 2014), brain size (Fonseca-Azevedo & Herculano-Houzel 2012), length of synaptic connections (Striedter 2006), and long-term memory formation (Potter et al. 2010, Suzuki et al. 2011, Plaçais & Preat 2013).

Since energy is required in various learning-related processes, the learning capacity is believed to be related to the constraints of metabolic energy. For example, in Drosophila, when there is energy deficiency, the formation of energy-intensive memories stops (Plaçais & Preat 2013). Similarly, in the mammalian brain, under conditions of low energy, an enzyme, serving as an energy sensor, reduces the activity of the protein translation pathway that affects the development of Long Term Potentiation (LTP) (Potter et al. 2010). In humans, studies indicate that the visual cortex allocates less energy to neurons that process information not in the focus of attention (Bruckmaier et al. 2020). Therefore, it is posited that the brain might adopt learning strategies to accommodate the energy cost under different conditions, particularly in scenarios of low energy supply.

Learning and memory are essential for survival. Through learning, both animals

and plants can respond to their environment in a way that improves their survival opportunity (Staddon 2016, Gagliano et al. 2016). Studies in wild populations, including species like squirrels, lizards, fish, etc., show that associative conditioning—either aversive or appetitive—enhances survival (Wright 2011, Morand-Ferron 2017). From the metabolic perspective, it is suggested that energy-regulated learning and memory play a role in survival. For instance, in invertebrates like honeybees and fruit flies, starvation before learning enhances appetitive Long Term Memory (LTM) formation. Conversely, feeding prior to conditioning weakens both learning and memory, regardless of training intensity (Müller 2013). Also, Drosophila suspends the formation of LTM in order to conserve energy during starvation, which is suggested as a survival strategy (Plaçais & Preat 2013).

Guided by these insights into energy conservation and adaptive learning, this study focuses on the olfactory nervous system in Drosophila's Mushroom Body (MB). This system was chosen as our biological point of reference for compelling reasons: its genetic clarity and its structural and functional parallels with the mammalian brain hippocampus (Wolff & Strausfeld 2015). Further enriching our choice is the fact that both insects and mammals share significant similarities in their olfactory systems (Davis 2004, 2005). With Drosophila's genetic accessibility and the comprehensively understood MB olfactory processing (Aso et al. 2014*b*), it presents a reliable reference to delve deeper into the intricacies of memory formation.

Given this background, this thesis hypothesizes that energy constraints in learning, and this constriction might not be mere limitations, but evolutionary adaptations to enhance survival. To unravel this, our investigation set on the fruit fly's learning processes, with a particular focus on the role of metabolic energy. The research objectives are:

1. To assess the potential for lifespan extension via energy-adaptive learning.

2. To evaluate learning efficiency and efficacy under metabolic limitations.

This thesis unveils three adaptive variations of reinforcement learning, specifically based on the multi-armed bandit problems, each inspired by the energy dynamics exhibited by the fruit fly. Across these variations, we explore diverse themes—from the role of metabolic energy in modulating memory pathways, to its potential in

fine-tuning exploration-exploitation balances, as well as its integration as a contextual determinant shaping decision-making processes.

This introductory chapter provides the research background, including an in-depth overview of olfactory learning within the MB, and explicates neuronal structures, including the energy's influence on MB's olfactory learning. Furthermore, it introduces the theoretical framework of the Multi-Armed Bandit (MAB) problem. At the end of this chapter, the primary contributions of the thesis are delineated.

The computational experiments in this research were performed using MATLAB, which facilitated the implementation of algorithms and provided tools for result visualization.

1.1 Olfactory Learning in Mushroom Body

In both insects and mammals, the olfactory nervous system shares many structural and functional similarities (Davis 2004, 2005). Intriguingly, genes influencing memory in fruit flies exhibit analogous roles in mammals, suggesting potential conservation in the mechanisms of olfactory learning (Davis 2005). The Mushroom Bodies (MBs) in insects, particularly in fruit flies, are pivotal in olfactory associative memory formation, which is considered to be functionally similar to the mammalian hippocampus (Wolff & Strausfeld 2015). Also, the Drosophila melanogaster Genetic Reference Panel (DGRP) thoroughly documents the correlation between molecular genetic and phenotypic variations (Mackay et al. 2012). The genetic accessibility of Drosophila coupled with its well-understood MB olfactory processing (Aso et al. 2014*b*) makes it a model organism for exploring the genetic underpinnings of memory formation.

Olfactory classical conditioning in Drosophila requires flies to associate an odor, used as a conditioned stimulus (CS), with a negative (electric shock) or positive (sucrose reward) stimulus, used as an unconditioned stimulus (US). Classically, this process is performed using groups of flies alternatively submitted to two different odors, a CS+, and a CS–, with only one of them (CS+) presented at the same time or slightly before the US (Quinn et al. 1974, Tully & Quinn 1985, Busto et al. 2010).

There are various methods to test the memory of this conditioning, where they are generally inspired by two classical setups. One is to test if the flies specifically have different behavior when presenting with a tube with the reinforcement-associated odor, CS+ (Quinn et al. 1974). Another is to put the flies in a T-maze with both two odors, CS+ and CS–, and see if the files have different behavior after learning (Tully & Quinn 1985).

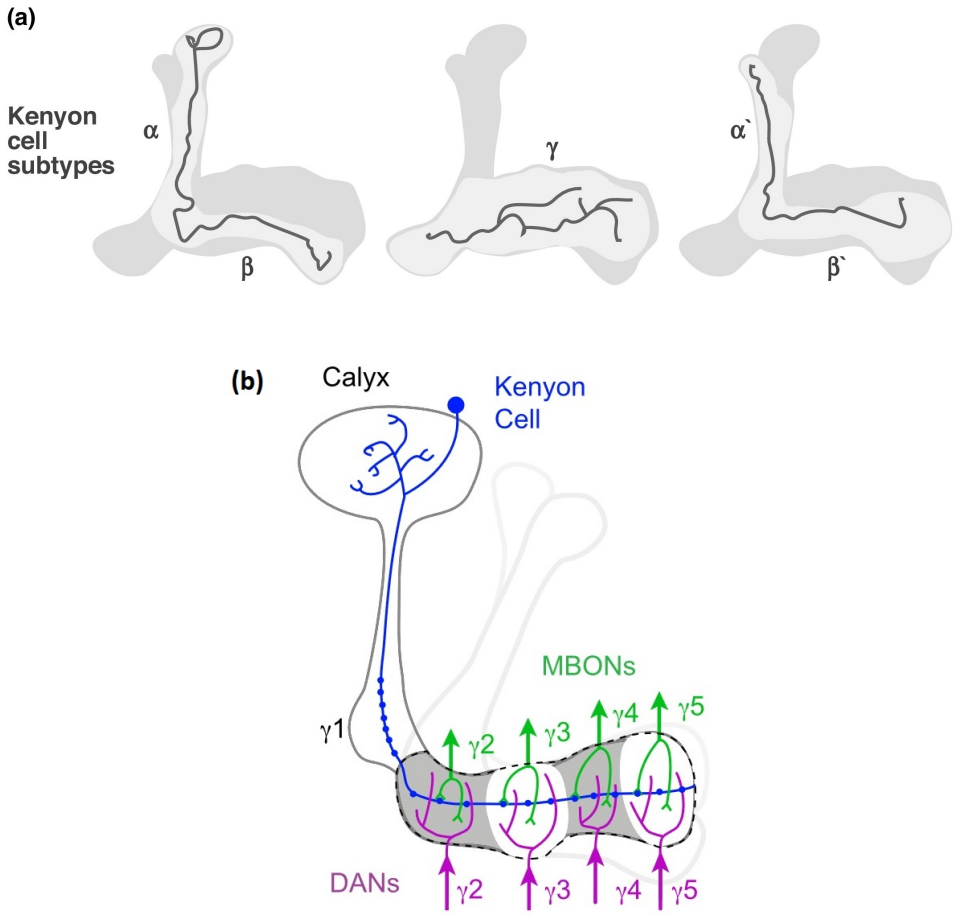

Figure 1.1: Layout of the Mushroom Body lobe. (a) Within the mushroom body neuropil, KCs are arranged in three distinct subtypes (individual KC examples represented in dark grey). Adapted from Cognigni et al. (2018). (b) Compartmentalized structure of the Mushroom Body lobe γ, adapted from Cohn et al. (2015).

Within the MB of Drosophila, approximately 2000 Kenyon cells (KCs) capture signals from a diverse set of olfactory glomeruli in the MB calyx, resulting in a unique sparse code (Turner et al. 2008, Campbell et al. 2013). These KCs transmit odor information through MB lobes, relaying signals to the mushroom body output neurons (MBONs), which subsequently influence odor-driven behavior (Cohn et al. 2015, Hige et al. 2015, Cognigni et al. 2018). The MB houses 34 MBONs, subdivided

into 21 unique cell types, the majority of the MBONs can induce an approach or avoid behavior. Typically, MBONs promoting approach behavior use acetylcholine or GABA, whereas those inducing avoidance behavior use glutamate (Owald et al. 2015, Perisse et al. 2016, Cognigni et al. 2018).

MBONs are distributed across MB output lobes, which are divided by three subregions, $\alpha\beta$–lobe, $\alpha'\beta'$–lobe and γ–lobe, as illustrated Figure 1.1, plot (a). Where each subregion has 5 anatomical compartments, shown in Figure 1.1, plot (b). Each MBON type primarily interfaces with one or two of these compartments (Aso et al. 2014*b*).

Dopaminergic neurons (DANs) play an instrumental role in reinforcing learning by modulating the KC-MBON synapses in response to reward or punishment signals, consequently altering the behavioral response to odors (Waddell 2013). These neurons, comprising 20 distinct types, exhibit compartmental specificity, with each type projecting predominantly to one or two compartments (Cohn et al. 2015). As illustrated in Figure 1.2, plot (b), for DAN neurons, there are two primary categories have been identified: the protocerebral anterior medial (PAM) DANs, generally associated with reward signaling (Burke et al. 2012, Yamagata et al. 2015, Cognigni et al. 2018), and the protocerebral posterior lateral (PPL1) DANs, typically linked to punishment signals (Mao & Davis 2009, Aso et al. 2010, 2012, Cognigni et al. 2018).

As depicted in Figure 1.2, the DANs typically form connections with MBONs of opposing valence within the same compartment (Cohn et al. 2015, Cognigni et al. 2018). A notable pattern in the MB lobe is the pairing of DANs and MBONs according to opposing valence—reward-associated DANs pair with avoidance MBONs, and vice versa. The details are provided in Figure 1.3. This arrangement ensures that when DANs receive reinforcement signals, they can locally depress the KC-MBON synapses within their designated compartment, leading to modified behaviors (Cohn et al. 2015, Hige et al. 2015, Owald et al. 2015). While the majority of these DAN/MBON compartments modulate approach or avoidance behaviors, exceptions exist. For instance, certain compartments drive 'alerting' behaviors in response to novel odors, or influence flight durations (Amin & Lin 2019).

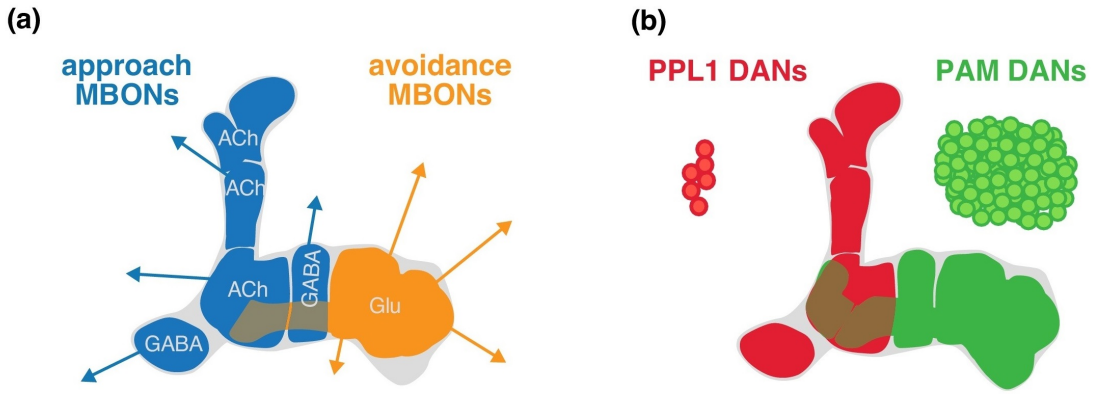

Figure 1.2: Classification of MBONs and DANs within the Mushroom Body, sourced from Cognigni et al. (2018). The dendritic structures of individual MBONs align with the compartmentalized regions of the DANs. Aversive-reinforcement DANs from the paired posterior lateral 1 (PPL1) cluster (in red) intersect with approach-favoring MBONs (in blue). Conversely, the protocerebral anterior medial (PAM) cluster's DANs, mainly relates to appetitive reinforcement (in green), intersect with avoidance-favoring MBONs (in orange). Notations indicate the neurotransmitters primarily utilized by each neuron category: ACh for acetylcholine; DA for dopamine; GABA for γ-aminobutyric acid; and Glu for glutamate.

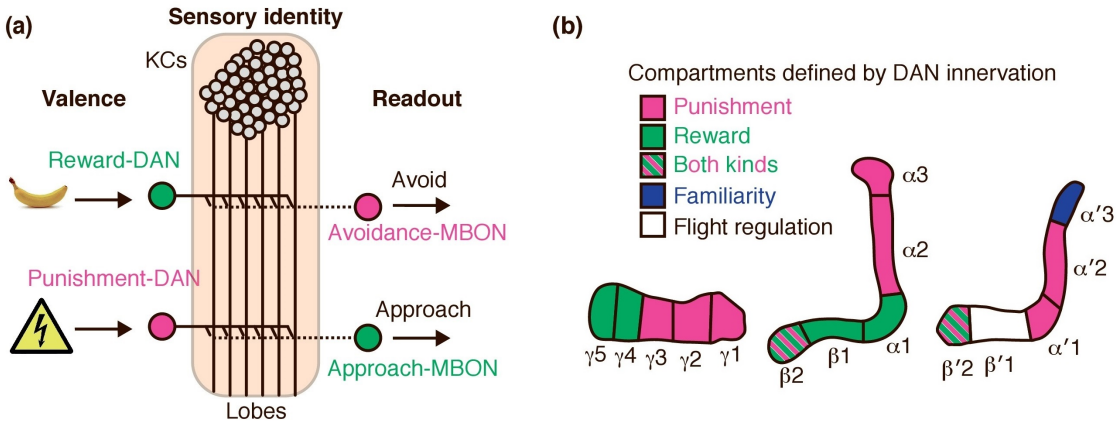

Figure 1.3: Valence of DANs and MBONs in MB sections. (a)The compartmentalized structure of the MB. KCs, depicted in gray, relay sensory identity data to MB lobes, establishing localized synapses with corresponding pairs of DANs and MBONs. When an odor correlates with a reward, reward-driven DANs reduce the KCs' excitatory influence on avoidance-oriented MBONs, which inclines the fly's behavior towards the presented odor. Also, the punishment-sensitive DANs diminish the excitatory impact of KCs on approach-focused MBONs, resulting in the fly's subsequent aversion to the odor. (b) The type of DANs in MB lobes signaling punishment (colored in red), reward (in green), recognition (in blue), or flight control (in white). Adapted from Amin & Lin (2019).

In olfactory classical conditioning, animals learn to associate an odor with a reinforcer, such as an electric shock in aversive conditioning or a food reward in appetitive conditioning, within a set time frame. This learning involves a brief training

phase, followed by stages of memory decay, consolidation, and retrieval. Testing and disruptive methods can be applied at any stage of this process (Busto et al. 2010). In Drosophila, this conditioning reveals various distinct memory phases, including Short-Term Memory (STM), Middle-Term Memory (MTM), Anesthesia-Resistant Memory (ARM), and Long-Term Memory (LTM), each characterized by different retention durations (shown in Figure 1.4).

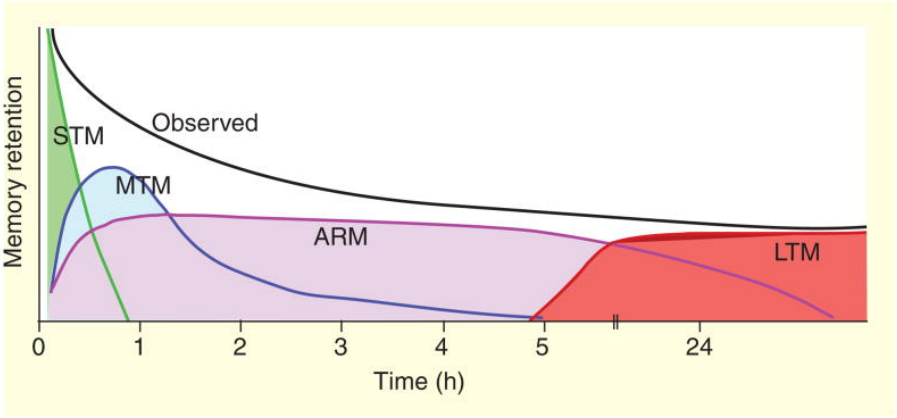

Figure 1.4: The Memory Phases of Drosophila, adapted from Margulies et al. (2005). Memory decay appears as a smooth transition (represented in black). Experiments have delineated separately at least four stages with different memory retention within Drosophila: short-term memory (STM, represented in green), middle-term memory (MTM, represented in blue), anesthesia-resistant memory (ARM, represented in purple), and long-term memory (LTM, represented in red).

The formation of these memory types is initially influenced by the training method and further modulated by the ecological context, such as environmental cues and resource distribution (Smid & Vet 2016). Here, we introduce three training methods: massed training, spaced training, and one-cycle training. Massed training delivers multiple learning trials in rapid succession without significant breaks. Spaced training distributes learning trials over a longer period, with intervals between sessions. One-cycle training involves a single learning trial. STM and MTM are labile memories, which are sensitive to anesthesia, such as CO_2 exposure or low temperatures, these two memories are observed in one training cycle of aversive conditioning (Tully et al. 1994). ARM is more stable and appears resistant to anesthesia, which appears in both single and massed training cycles, with enhanced retention in the latter (Tully et al. 1994). Spaced training, which incorporates intervals between training cycles, triggers the development of LTM, while concurrently suppressing ARM formation (Isabel et al. 2004). Notably, the LTM formation involves protein

synthesis, when inhibiting the neural activity related to the protein synthesis, the memory consolidation will be disrupted (Tully et al. 1994, Krashes & Waddell 2008, Colomb et al. 2009). Interestingly, although LTM and ARM activate neurons in the same MB compartments, these two memory pathways are mutually exclusive (Isabel et al. 2004). The activation of these pathways is influenced by factors including hunger levels (Plaçais & Preat 2013) and the energy fluxes associated with glucose consumption (Placais et al. 2017), the details of this energy-related regulation are introduced in the Section 1.2.

1.2 The Energy Impact on Olfactory Conditioning in MB

In the animal brain, the way networks of neurons are designed reflects the need for energy efficiency (Niven & Laughlin 2008). The energy-saving mechanism is seen in many sensory systems, for example, sparse coding in the sensory system is believed as a method where the brain uses fewer resources to represent information (Olshausen & Field 2004). Also, researchers suggest that memory-related brain function is modulated by the intake of energy resources such as glucose (McNay & Gold 2002).

In Drosophila, the energy-regulated learning mechanism is studied in aspects from the behavior level to the molecular and cellular level. Behaviorally, metabolic energy can influence exploratory actions during learning. This energy dynamic plays a pivotal role in neurobehavioral decision-making and reinforcement learning, especially regarding the dilemma of exploration versus exploitation. This dilemma requires the learner to weigh the benefits of utilizing known resources against the potential gains from discovering new ones (Cohen et al. 2007). Specifically, in the context of foraging and associative learning with neuronal rewards, the energy implications of such decisions are accentuated, food deprivation in Drosophila foraging results in extended durations of local searches, indicating an intensity to remain near the food source (Bell et al. 1985). This is evident in the balance struck between the urgency to locate nutrient-abundant sustenance and the energy/effort expended in exploration (Krebs et al. 1978, McNamara & Houston 1985, Bell 2012, Katz & Naug

2015).

From the aspect of memory formation, compared to the memory that attenuates after learning, the formation of the LTM is energy-costly. For instance, during the early phase of forming LTM, fruit flies doubled their sucrose consumption, and the energy consumption in the MB is also elevated (Placais et al. 2017). Also, when the fruit flies have no nutritious intake, the energy cost of forming LTM will significantly reduce their lifetime (Mery & Kawecki 2005). When artificially activated DANs modulate the formation of LTM, a higher energy consumption rate in MB neurons will be triggered (Placais et al. 2017).

The metabolic energy also regulates the retention of the memory. In aversive conditioning, Drosophila stops forming the energy-costly LTM, and only forms the short-term ARM when they are starving (Plaçais & Preat 2013). In appetite conditioning, when the unconditioned stimulus is non-digestible sugar, the flies only form short-termed lived memories, where the sweet stimulus that can be metabolized leads to LTM (Yamagata et al. 2015). When the reward is delayed, the long-term appetite memory can only be constructed after taking the nutritional reward (Trannoy et al. 2011). Notably, for both aversive and appetitive conditioning, the activity of a type of PPL (punishment-related) DANs, called MP1, are involved with this long-term memory formation (Trannoy et al. 2011, Placais et al. 2017).

The activities of MP1 DANs are influenced by Drosophila *neuropeptide F* (dNPF), a neuropeptide that represents the hunger state in flies (Krashes et al. 2007, 2009), which is a homolog of mammalian *neuropeptide Y* (Brown et al. 1999). Stimulating NPF-expressing neurons can replicate the hunger state in fed flies. When the NPF receptor does not function well, the starved flies will not display the hunger-dependent memory and behave as if they were full (Krashes et al. 2009).

As summarized in Figure 1.5, the MP1 DANs play multifaceted roles in the MB, beyond participating in the energy-costly process of long-term memory formation, they also interface with hunger signals during olfactory conditioning. Within the structured architecture of the MB lobes, MP1 DANs form connections with MVP2 MBONs, promoting approach behaviors (Perisse et al. 2016). During appetitive memory retrieval, the energy state of the fly modulates odor-driven behavior by channeling the hunger-dependent dNPF signal through MP1 DANs. In hungry flies,

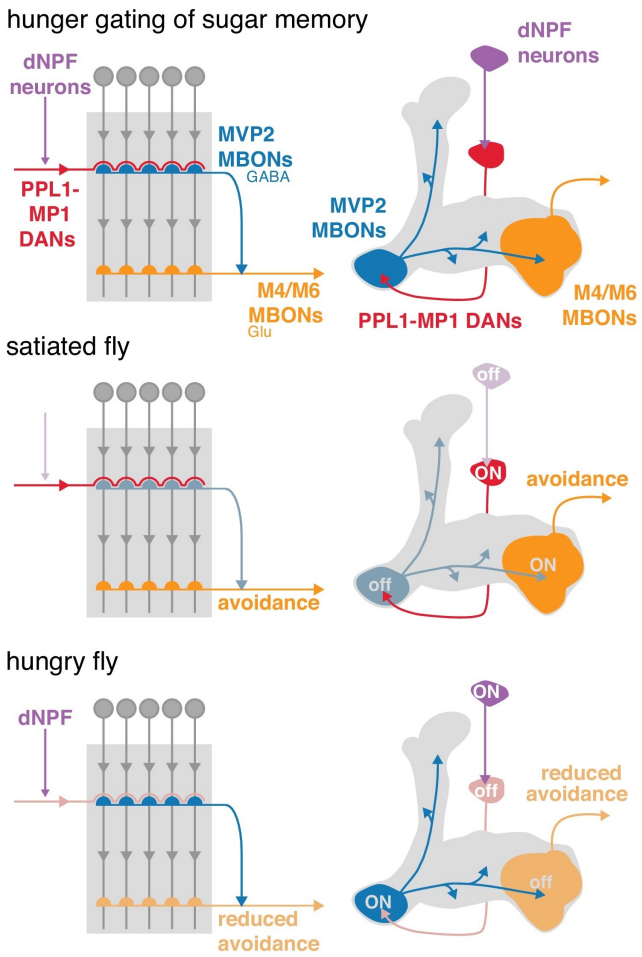

Figure 1.5: Hunger-Modulated food memory retrieval, adapted from Cognigni et al. (2018). Feed-forward inhibition governs the hunger-driven expression of food memory retrieval. The hunger state modulates odor-responsive behaviors by channeling hunger-dependent dNPF signaling (highlighted in purple) via the PPL1-MP1 DANs (depicted in red) and the GABAergic MVP2 neuron (in blue). This MVP2 neuron delivers feed-forward inhibition to the M4/M6 cluster of glutamatergic MBONs (portrayed in orange). When the fly is satiated, approach behaviors are counterbalanced by the avoidance-driven M4/M6 MBONs, which remain uninhibited by MVP2. Conversely, in a famished state, the MVP2 neuron becomes active and curtails the M4/M6 MBONs' activity, diminishing avoidance tendencies and promoting approach behaviors.

the positive-valence MVP2 MBON is activated under the influence of MP1 DANs. Conversely, in satiated flies, MVP2 MBON activation is suppressed, and the activities of its counteracting negative-valence counterparts, the M4/6 MBONs, are triggered, leading to avoidance behaviors (Cognigni et al. 2018). This intricate neural circuitry ensures that flies act on their food-associated memories in a contextually appropriate manner, primarily when they are hungry. Essentially, the fly's neural framework possesses an intrinsic mechanism that adjusts behavior to its physiolog-

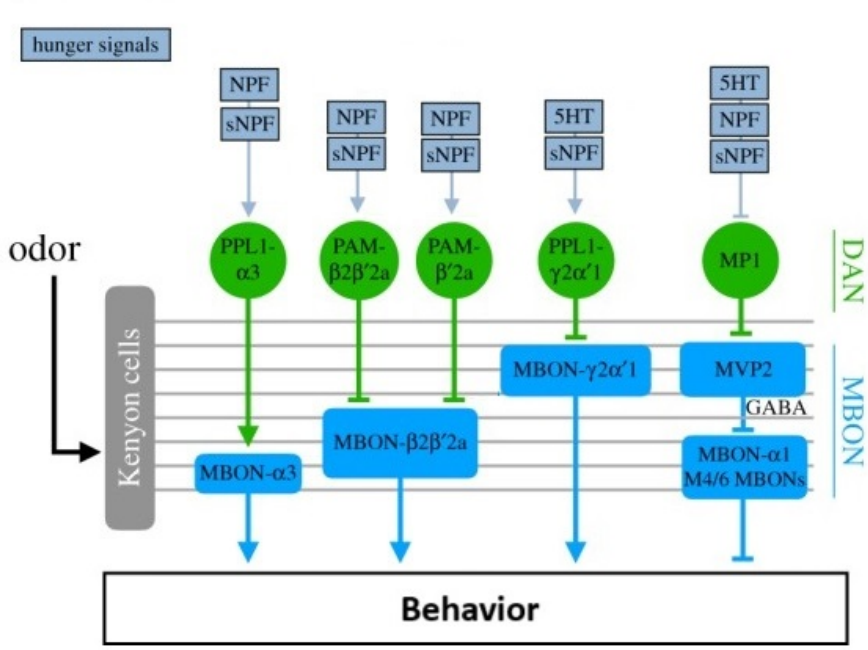

Figure 1.6: Hunger-Regulated odor responses, modified from Lin et al. (2019). Five MBON pathways (in blue) which target varying compartments of the KC axons facilitate odor-induced food-seeking behaviors. The activities of these MBON pathways are modulated by their associated dopaminergic neurons (DANs; shown in green). Each of these DANs processes a unique blend of hunger and satiety signals. Notably, the MP1-MVP2-M4/6 MBON route plays a key role in mediating the hunger influence on learned foraging behaviors.

ical needs.

While the MP1 neurons serve as pivotal regulators in energy-dependent behaviors, several DANs and MBONs are subject to the modulation of various hunger signals. Apart from dNPF, neuropeptides like sNPF (short Neuropeptide F) and neurotransmitters such as serotonin (5HT) play integral roles, interacting with DANs and MBONs across diverse MB output compartments (Lin et al. 2019). The details of the hunger-dependent odor responses are provided in Figure 1.6.

1.3 Multi-Armed Bandit Problem

In the 1930s, Thompson (1933) considered the choice of optimal medical treatments between two options, he argued against the practice of rejecting a treatment based on one unfavorable clinical trial, and introduced a method to estimate the likelihood of one treatment being superior to another, using a Bernoulli probability to repre-

sent the success rate of each alternative. This approach led to the formulation of the Multi-Armed Bandit (MAB) Problem, a framework for decision-making under uncertainty that balances exploration and exploitation.

Nowadays, the MAB problem is treated as a simplified setting of the reinforcement learning problem with a single state (Sutton & Barto 2018), which involves a decision-maker (the agent) choosing from multiple options (or 'arms'), each with unknown rewards. The agent's goal is to maximize cumulative rewards or minimize regret (the gap between optimal and actual rewards) over several trials, without knowing the exact reward probabilities. In a K-armed Bandit scenario, the agent faces K alternatives. For each trial t, the agent must:

1. Select an arm i.
2. Receive reward $r_{t,i}$ from that arm.
3. Update their reward estimation based on the outcome.

In this process, the reward estimation and arm selection can be guided by various algorithms. When an algorithm frequently chooses the most promising option based on current knowledge, it is considered more 'greedy.' Such a greedy approach allows the agent to capitalize on the arm believed to offer the highest reward. However, this comes with a drawback: it limits exploration of other potential options, which might be more beneficial but less apparent, specifically in the early stage. Thus, a critical challenge in MAB algorithms is striking an optimal balance between exploration (investigating various alternatives) and exploitation (maximizing returns from known options).

In human and animal decision-making studies, MAB tasks are essential for understanding brain mechanisms and cognitive processes. For example, Daw et al. (2006) used MAB tasks in human decision-making research and employing fMRI to the subjects. They discovered activity in two prefrontal cortex (PFC) regions during these tasks - the ventromedial and frontopolar PFC. The ventromedial PFC is known for processing reinforcement signals across various tasks, while the frontopolar PFC is more active during exploratory choices. In avian studies, birds facing a two-armed bandit problem with feeding posts exhibited decision patterns aligned with the Gittins index predictions (Krebs et al. 1978), a MAB algorithm that estimates expected future rewards considering reward discounting (Gittins et al. 2011). Additionally, research on Drosophila's olfactory learning using a two-armed bandit framework has

helped in estimating neuron dynamics during decision-making processes (Loewenstein 2008, Bennett et al. 2021).

This thesis also develops computational models based on MAB frameworks, Chapter 2 and Chapter 3 use a two-armed model to explore the neural mechanisms of approach and avoidance behavior. In Chapter 4, the framework of three prevalent MAB algorithms is introduced, together with their effectiveness, especially in the context of metabolic energy impact.

1.4 Main Contributions

The primary contributions of this thesis are threefold, each detailed in subsequent chapters.

Chapter 2 delves into the energy-dependent regulation of memory pathways, producing two memory types with distinct retentions. It investigates the impact of such energy-adaptive regulation on survival. Within a basic aversive conditioning framework, we integrate the energy expenditure associated with memory formation. This framework examines the approach and avoidance responses to an odor coupled with a stimulus. The aim is to unveil strategies that modulate memory formation contingent on available energy reserves. Concurrently, we calculate the prospective lifespan of fruit flies, factoring in the dangers from both energy deficiencies and aversive stimuli. The result of the simulation reveals that the energy-regulated memory pathways can enhance the longevity of fruit flies. This longevity augmentation is amplified when dopamine regulation participates with the energy-adaptive memory pathway adjustments.

Chapter 3 unfolds around the efficiency of synaptic plasticity, focusing on the olfactory learning processes in fruit flies. It establishes a theoretical framework that incorporates energy as a form of contextual information, which influences decision-making processes. This chapter underscores the pivotal role of dopaminergic neurons, which serve as interpreters of energy signals during decision-making. We investigate various synaptic weight adjustment tactics. The evaluation criterion is the weight update efficiency evaluated by the enhancement in the rate of exhibiting

correct behavior given a uniform synaptic weight change. The model results indicate that suppressing the weight associated with the undesired behavior, which is the weight update strategy aligns with MB olfactory learning mechanics, exhibits a higher efficiency compared with other learning strategies. This trend persists across both aversive and appetitive conditioning, and remains consistent when considering energy levels as contextual information that biases decision-making.

Chapter 4 shifts the focus to a more complex foraging context. Here, we explore energy-regulated decision-making across multiple odors, each with a distinct amount of food resources. In a departure from the previous chapters, a more algorithmic approach is applied. In this chapter, energy becomes a regulatory factor for balancing exploration and exploitation, based on three well-established Multi-Armed Bandit approaches. By weighing the energy cost of food-seeking activities against the energy intake from the appetitive reward, we can deduce an agent's lifespan. Intriguingly, our investigations highlight the advantage of energy-adaptive versions of the "optimistic" Upper Confidence Bound and Bayesian-based Thompson Sampling methods. These techniques not only extend lifespans but also achieve this with small increase in regret compared to their conventional counterparts.

Chapter 2

Energy-Adaptive Aversive Conditioning

2.1 Synopsis

Synaptic plasticity enables animals to adapt to their environment, but making reliable changes in synaptic strength consumes a substantial amount of metabolic energy, potentially impairing survival. Hence, the brain must regulate learning wisely. Indeed, during starvation, Drosophila suppress the formation of energy-intensive aversive memories. Here, we include energy considerations in a two-armed bandit framework, and investigate strategies that regulate memory formation depending on the animal's energy reserve. Simulated flies learned to avoid noxious stimuli through synaptic plasticity in either the long-term memory (LTM) or the anesthesia-resistant memory (ARM) pathway, each with distinct energy demands and decay rates. We propose two energy-adaptive learning approaches: one with a fixed energy threshold and another with a dopamine-dependent threshold. Consistent with experimental results, we show that regulating LTM and ARM pathways based on energy reserve prolongs lifespan, also under stochastic conditions, highlighting the significance of energy-regulated memory pathways and dopaminergic control in adaptive learning and survival.

2.2 Introduction

In the natural environment, a primary function of animal learning is to adapt to their surroundings, evade dangers, and enhance survival prospects. However, learning itself is not risk-free as it demands considerable energy. For instance, experiments have shown that fruit flies that learn a classical conditioning task, perish 20% faster when subsequently starved (Mery & Kawecki 2005). Yet, despite such experimental evidence as well as its potential relevance for computational hardware (Han et al. 2016), the energy requirements of learning have thus far been mostly overlooked in the computational community.

In Drosophila memory is expressed in two distinct and mutually exclusive pathways, each with different energetic demands (Isabel et al. 2004, Mery & Kawecki 2005, Plaçais & Preat 2013). The LTM pathway is characterized by high energy demands and persistent memory. Conversely, the ARM pathway requires negligible amounts of energy, but typically dissipates within four days (Tully et al. 1994, Isabel et al. 2004). A circuit of dopamine neurons in the Mushroom Body (MB) of the fly signals the availability of energy (Musso et al. 2015, Placais et al. 2017). Notably, flies halt energy-demanding LTM formation under conditions of starvation (Plaçais & Preat 2013).

Inspired by these observations, this study is centered around two primary objectives: 1) examine the energy cost and benefit of learning on survival, 2) identify an optimal learning strategy contingent upon reward magnitude and plasticity costs. To address the first objective, we introduced a hazard framework to examine the trade-off between the energy expenditure incurred during learning and the consequent reward/punishment. Learning to evade aversive stimuli decreases the stimulus hazard, while concurrently, the energy expenditure associated with learning increases the starvation hazard. The objective for the flies in this experiment is to maximize their survival, seeking a robust strategy that prolongs their lifetime. To address the second objective, our model alternates between the LTM and ARM memory pathways. Switching between them is contingent on an energy threshold. The model uses the LTM pathway when energy reserves exceed this threshold, and shifts to the more energy-conservative ARM pathway when energy falls below this value.
We devised two strategies to regulate this energy threshold: a fixed-threshold ap-

proach and a dopamine-regulated moving threshold approach. Both strategies were found to prolong the lifespan of fruit flies and exhibited robust performance in stochastic environments. The dopamine-regulated energy adaptive learning demonstrated a more pronounced effect on lifespan extension. This observation suggests the incorporation of dopaminergic signals in the regulation of energy-adaptive learning.

2.3 Learning Network

In pursuit of an optimal memory regulation strategy to maximize the lifespan, we developed a feedforward decision network reflecting the Drosophila brain's anatomical structure, shown in Figure 2.1, and a complementary feedback network associated with reinforcement, shown in Figure 2.2.

2.3.1 Decision-Making Network

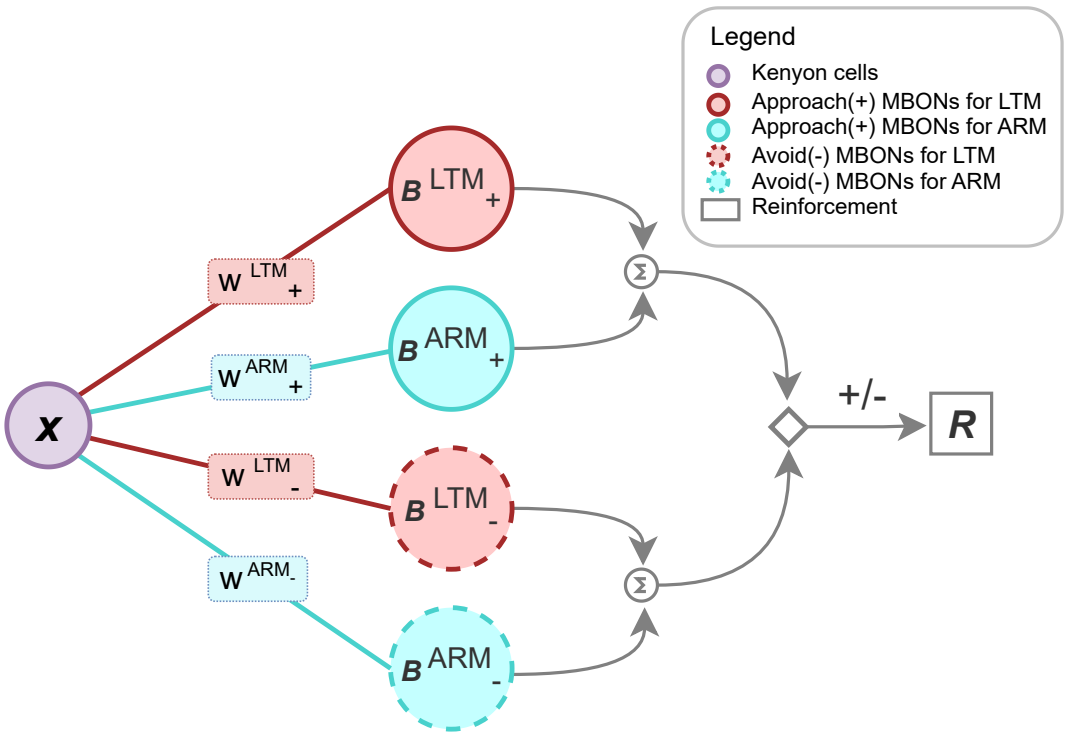

Figure 2.1: A feedforward Decision-making network based on the anatomical structure of the Drosophila brain.

In Drosophila aversive conditioning experiments, an odor (Conditioned Stimulus,

CS) is paired with a shock (Unconditioned Stimulus, US). By repeating exposure to the CS-US pairs a few times, the flies learn to avoid the odor, as can be subsequently tested in a T-maze. The underlying circuitry, involving sensory encoding Kenyon Cells (KCs) and action-driving Mushroom Body Output Neurons (MBONs), is relatively well-understood (Tempel et al. 1983, Tully et al. 1994, Plaçais & Preat 2013, Aso et al. 2014*a*).

Our decision-making network comprises sensory KCs that respond to the given odor, which subsequently interfaces with MBONs that drive behavior, Figure 2.1. We assume that neural processing and decision-making are inherently noisy. Without this noise, even a minor imbalance could determine the outcome. Given independent Poisson spike-time variability, the variance of input to decision-making neurons matches the mean input. At sufficiently high firing rates, this input can be approximated by a normal distribution with variance equal to the mean. Consequently, we model the firing rate of KCs in response to the odor x as a noisy normal distribution characterized by mean μ and variance σ^2, where $\mu = \sigma^2$.

The activities of the MBONs are modeled as linear neurons $B_{\pm}^{LTM} = w_{\pm}^{LTM}x$, and $B_{\pm}^{ARM} = w_{\pm}^{ARM}x$, where $\pm$ indicates approach (+) and avoidance (-) behaviors, and the parameters $w_{\pm}^{LTM}$ and $w_{\pm}^{ARM}$ denote the synaptic strengths from the KCs to the MBONs. Given the association of the odor with aversive stimuli, the reward for approach behavior R_+ equals the strength of the aversive stimulus, set within a range of -1 to 0, $R_+ \in [-1, 0]$. The reward for avoiding the odor, R_-, is set to 0.

Given the additive nature of MBON signals (Aso et al. 2014*a*), we posit that total neuronal activity driving the approach and avoidance behaviors result from the sum of the ARM and LTM components. Hence

$$B_{\pm} = (w_{\pm}^{ARM} + w_{\pm}^{LTM})x \tag{2.1}$$

The total weight for approach and avoidance behaviors is $w_{\pm} = w_{\pm}^{ARM} + w_{\pm}^{LTM}$. To ensure model stability and biological plausibility, we constrain the total weight $w_{\pm}$ to be non-negative and saturate at 1. Winner-Take-All competition among MBON neuron populations determines the fly's action. Although the competition process is not explicitly modeled, it can potentially be captured using lateral inhibition and

attractor dynamics. Because of the neural noise in x, the decision is stochastic. The probability is a sigmoidal function of the difference between B_+ and B_-.

The probability of choosing an avoidance action based on the corresponding weight values can be found as $P(B_- > B_+)$. We introduce a new random variable $Z = B_- - B_+$, then the mean (μ_Z) and variance (σ_Z^2) of Z are given by:

$$\mu_Z = (w_- - w_+)\mu$$

$$\sigma_Z^2 = (w_-^2 + w_+^2)\sigma^2$$

The probability $P(Z > 0)$. Since Z follows a standard normal distribution, we can express this probability in terms of the error function (erf):

$$P(Z > 0) = 1 - P(Z \leq 0)$$

$$= \frac{1}{2}\left(1 - \text{erf}(\mu_Z/(\sigma_Z\sqrt{2}))\right)$$

Substituting the values for μ_Z and σ_Z, we get:

$$P(Z > 0) = \frac{1}{2}\left(1 - \text{erf}\left(\frac{(w_- - w_+)\mu}{\sqrt{2\sigma^2(w_-^2 + w_+^2)}}\right)\right)$$

Considering the properties of a Poisson distribution, we let $\mu = \sigma^2$ in our study, then the probability of taking the avoidance action becomes:

$$P_-(w_-, w_+, \mu, t) = \frac{1}{2}\left(1 - \text{erf}\left(\sqrt{\mu}\frac{(w_- - w_+)}{\sqrt{2(w_-^2 + w_+^2)}}\right)\right) \tag{2.2}$$

The value of μ has been estimated to be 10, based on experimental data. A detailed explanation of this calculation is presented in Section 2.4.2.

2.3.2 Energy-Adaptive Learning Driven by Reinforcement Signals

Next, we describe the feedback circuit and learning, as displayed in Figure 2.2. In the MB of Drosophila, reinforcement-related signals are encoded by DANs, and these DAN signals modulate the plasticity of the synapse connecting KCs to MBONs

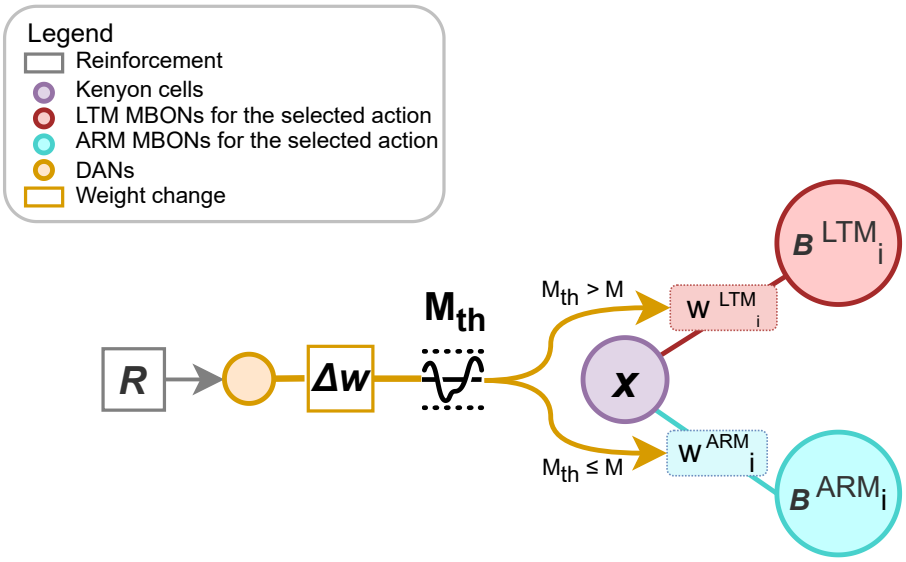

Figure 2.2: A Feedback Energy-Adaptive learning driven by reinforcement signals. Where M_{th} is the energy threshold index that gates the memory pathways, M is the energy index indicating the remaining energy. Δw is the weight change modulated by the reinforcement.

(Cohn et al. 2015, Bennett et al. 2021). The synaptic strength associated with the selected behavior is updated based on the difference between the reward from the current trial $R_i(t)$ and the moving average of the reward from previous trials $\overline{R}(t-1)$, compounded by the odor-related input x from KCs. Here, the reward $R_i(t)$ is either R_+ (for approach) or R_- (for avoid), depending on the selected action. The corresponding synaptic weight modification is

$$\Delta w_i = \eta \left(R_i(t) - \overline{R}(t-1) \right) x \tag{2.3}$$

Where η denotes the learning rate, which has been calibrated based on experimental results (Tully et al. 1994) to a value of 0.6, the details of this calibration are presented in Section 2.4.2. $\bar{R}$ denotes the running average of the reward computed with a time constant τ_R:

$$\bar{R}(t) = \bar{R}(t-1) + \frac{(R(t) - \bar{R}(t-1))}{\tau_R}$$

For our study, we set the parameter τ_R to 10. In accordance with this model, the weight inducing approach behavior will diminish over time, thereby progressively reducing the probability of the organism approaching the aversive stimulus.

Building on experimental evidence which indicating mutual exclusivity between ARM and LTM memory formation (Isabel et al. 2004), we postulate that weight

changes contribute exclusively to either LTM or ARM. To facilitate this, we define a metabolic energy value, denoted as M, which is defined within a range from 0 to 1, representing the energy reservation of the fruit flies. The specific definition and calculation of M are discussed in Section 2.4. The selection of the LTM and ARM pathway is controlled by a threshold of this energy index, denoted as energy threshold (M_{th}). If the energy index exceeds this threshold ($M > M_{th}$), the LTM weight ($w_{\pm}^{LTM}$) is updated, and an amount of energy proportional to the weight change is expended, the estimation of this energy expenditure is introduced in Section 2.4.1. Conversely, if the energy index falls below this threshold, the weight in the ARM pathway ($w_{\pm}^{ARM}$) is updated. This process consumes a negligible amount of energy but leads to weight decay over time, then we have

$$w_i^{ARM}(t) = \gamma_{ARM}(w_i^{ARM}(t-1) + \Delta w_i(t-1)) \tag{2.4}$$

Here, γ_{ARM} is the ARM retention factor, and time is denoted in days. Using experimental data, we estimate γ_{ARM} to be 0.34. Detailed calculations can be found in Section 2.4.2).

2.4 Analysis of Metabolic Energy and Lifetime

To quantify the trade-off between metabolic energy usage and the learning required to avoid aversive stimuli, we utilize the concept of a hazard function (Clark et al. 2003, Gerstner et al. 2014). In this context, the hazard represents the probability of mortality within a specific time interval (measured in days). The overarching goal for the Drosophila in our model is to optimize their lifespan.

We categorize two types of hazards that contribute to a reduction in lifespan. The first is the hazard associated with starvation, denoted as h_M. Where M is the energy level for monitoring the metabolic energy reservation, which varies between 0 and 1. In conditions of food abundance, the energy level reaches its maximum, at which point the value of M equals 1. We assume that under these conditions, the Drosophila would live a natural lifespan (l_N). Based on experimental results (Mair et al. 2005, Min et al. 2007, Fanson et al. 2009, Krittika & Yadav 2019), we suppose the excessive calorie intake does not prolong lifespan in this study.

Within our model, the starvation hazard escalates exponentially as the energy reserve diminishes. Therefore, the hazard from starvation can be summarized as follows:

$$h_M(t) = \exp(-c_m M(t)) \tag{2.5}$$

Here, c_m is a factor calibrated such that flies have their natural lifespan when energy is plentiful (i.e., M=1), see details in Section 2.4.2.

In the real world, metabolic energy is influenced by many factors, such as metabolic processes as well as random events such as food availability. However, we focus on the changes in the value of the energy index M as a result of the weight change being incorporated into the LTM pathway. The underlying reasons for the metabolic cost learning are not known, see e.g. (Girard et al. 2023). We assume that the metabolic energy cost of LTM formation decreases the energy reserve by an amount proportional to the weight change (Li & Van Rossum 2020)

$$M(t) = M(t-1) - c_{LTM}\left(|\Delta w_+^{LTM}| + |\Delta w_-^{LTM}|\right) \tag{2.6}$$

The parameter c_{LTM} denotes the energy cost of LTM, for calibration, see Section 2.4.1. Hence, LTM learning increases the starvation hazard. ARM learning is assumed to not decrease the energy reserve (Mery & Kawecki 2005).

The second source of hazard, denoted h_s, is from the aversive stimulus. Although laboratory experiments generally involve non-lethal shock stimuli, in a natural environment, such shocks could potentially forebode a perilous event. As elucidated in the prior section, we propose to equate the perceived hazard arising from the stimulus h_s to the reinforcement received when the fly approaches the odor R_+. When the fly avoids the odor, this hazard is avoided. Consequently, we assume that Drosophila has a probability of dying when it fails to evade the US.

Given that hazards represent probabilities, the total hazard, $h_\sum$, at time t can be calculated as follows:

$$h_\Sigma(t) = 1 - (1 - h_s(t))\,(1 - h_M(t)) \tag{2.7}$$

The survival function, denoted as $S(t)$, provides the probability that Drosophila will

survive up to a specific time t. When the time is continuous, the survival function is

$$S(t) = \exp\left(-\int_0^t h_\Sigma(t')dt'\right) \tag{2.8}$$

At the onset of the experiment, the survival function is initialized at 1, and gradually decreases to 0 as t approaches infinity. In our simulations, we employ discrete time intervals, measured in days. Consequently, the relationship between the hazard function and the survival function is

$$S(t) = \exp\left(-\sum_{t'=0}^{t} h_\Sigma(t')\right) \tag{2.9}$$

The expected lifetime of a fruit fly population who have hazard over time $h_\sum(t')$ can be approximated by

$$l = \int_0^\infty S(t)dt = \int_0^\infty \exp\left(-\int_0^t h_\sum dt'\right) dt \tag{2.10}$$

When the time is discrete, the expected lifetime can be found by

$$l = \sum_{t=0}^{\infty} S(t) = \sum_{t=0}^{\infty} \exp\left(-\sum_{t'=0}^{t} h(t')_\sum\right) \tag{2.11}$$

Throughout the modeling process, we determine the expected lifetime subsequent to the learning. This value serves as an indicator of the efficacy of the strategy for memory pathway selection. Hence, the optimal memory pathway regulations are those that effectively maximize the lifespan.

2.4.1 The Energy Cost of LTM

The precise metabolic cost of plasticity is not known, nor is its origin – protein synthesis, receptor transport, or replay processes are among the many candidates (Laughlin et al. 1998, Herculano-Houzel 2011, Wang et al. 2016, Karbowski 2019). We assume that metabolic energy is proportional to the absolute value of the weight change (Li & Van Rossum 2020). We quantify the energy consumption in each trial by ΔM. According to experimental results (Mery & Kawecki 2005), the formation of ARM does not lead to a considerable reduction in lifespan compared to control flies. Hence, we assume that the cost of ARM is negligible, and assume the metabolic

energy cost of LTM is proportionate to the weight change. Thus, the energy cost for learning is

$$\Delta M = c_{LTM}\left(|\Delta w_{+}^{LTM}| + |\Delta w_{-}^{LTM}|\right) \tag{2.12}$$

Where c_{LTM} is the rate of energy consumption when using LTM, we estimate this value based on the experimental research from (Mery & Kawecki 2005), which investigated the lifetime of the fruit flies directly afterward the conditioning when depriving food and water, showing that compare with the flies subjected to non-associative conditioning, flies conditioned in the associative conditioning died on average 4 hours earlier. This means this 4-hour difference is caused by the formation of LTM. In this model, we assume that the flies achieve flawless after the conditioning learning and are thus able to consistently avoid the odor. When the initial value of LTM weights is 0.5, this implies that the total weight change is equal to one, i.e., $|\Delta w_{+}^{LTM}| + |\Delta w_{-}^{LTM}| = |-0.5| + |0.5| = 1$. Then we have $\Delta M_{LTM} = c_{LTM}$ in Equation (2.12).

To model the post-conditioning scenario, we posit that the conditioning concludes at time zero, and the energy state at this moment denoting as M_0, represents the residual energy reserves immediately after conditioning. Given the absence of food supplementation post-training, and this holds for both Associative (AC) and Non-Associative (NA) protocols, we hypothesize a linear decline in energy reserves, governed by a constant basal energy consumption rate (β). Consequently, the energy state in the post-conditioning scenario (M_{pc}) can be mathematically represented as:

$$M_{pc}(t) = M_0 - \beta t \tag{2.13}$$

Since the flies are not pre-starved in this experiment, we set the initial energy M_0 for non-associative conditioning as its maxima $M_0^{NA} = 1$. For associative conditioning, the energy consumption of LTM is consumed during the training process, hence we suppose $M_0^{AC} = 1 - \Delta M_{LTM}$. Substitute Equation (2.13) into Equation (2.5), the post-conditioning hazard can be found as

$$h_{\mathrm{pc}}(t) = \exp(-c_m(M_0 - \beta t))$$

Where c_m stands for the steepness of the starvation hazard, the estimation of its value is introduced in Section 2.4.2. Given the survival function Equation (2.8), we

get the survival function under this protocol:

$$S_{\text{pc}}(t) = \exp\left(-\frac{\exp(-c_m M_0)(-1+\exp(c_m \beta t))}{c_m \beta}\right)$$

Then the lifetime can be found as:

$$l_{\text{pc}} = -\frac{\exp\left(\frac{\exp(-c_m M_0)}{c_m \beta}\right) \text{Ei}\left(-\frac{\exp(-c_m M_0)}{c_m \beta}\right)}{c_m \beta} \tag{2.14}$$

The variable Ei represents the exponential integral function. Drawing on the experimental study (Mery & Kawecki 2005), we note that subjecting fruit flies to AC spaced conditioning results in a reduction of the lifespan by approximately four hours when compared to those subjected to NA spaced conditioning. This suggests that the formation of LTM effectively reduces the lifetime by about four hours. The lifespan of male and female flies post-NA conditioning stands at roughly 20 and 25 hours respectively.

In our model, we have made the simplifying assumption that energy expenditure during the learning protocol can be neglected, and that the fruit flies subjected to non-associative conditioning possess maximal energy immediately after the learning phase. Subsequently, we can estimate the energy cost in two steps as follows:

Step 1: Determination of the energy depletion rate, β. By assuming the energy right after the NA condition is at the maxima, i.e. $M_0^{NA}1$, we can substitute this value into the equation to derive the energy depletion rate given the lifetime, l. This is achieved through a numerical solution of Equation (2.14), yielding values of β for male and female flies as 1.59 and 1.20 respectively.

Step 2: Estimation of initial energy. To ascertain the initial energy at the onset of starvation, we can substitute the derived value of β into Equation (2.14). As mentioned earlier, $M_0^{AC} = 1 - \Delta M_{LTM}$, where $\Delta M_{LTM} = c_{LTM}$, we have $M_0^{AC} = 1 - c_{LTM}$. Considering that the lifetimes for males and females post-AC training are 4 hours shorter than those post-NA conditioning, we use 16 and 21 hours respectively as the initial energy values for males and females in our computations. With the corresponding values of β calculated in step 1, we obtain c_{LTM} values of 0.27 and 0.21 for males and females respectively.

In our simulation, we adopt a value of $c_{LTM} = 0.21$. It's worth mentioning that this might result in an underestimation of the energy savings accrued from employing ARM in our simulation. Consequently, the lifetime difference when alternating memory pathways could exhibit greater significance in real-life scenarios.

2.4.2 Model Parameters Calibration

In this study, we utilized empirical data (Tully et al. 1994) to calibrate parameters for ARM retention (γ_{ARM}), the mean (μ) and variance (σ^2) of the sensory input distribution, and the learning rate η. The experiment involved ten cycles of massed aversive conditioning training in Drosophila. As introduced in Section 1.1, massed training involves rapid, consecutive learning trials without breaks and the fruit flies generated ARM exclusively within this training regime. In each training cycle, flies were exposed to an odor (CS+) paired with electric shock (US) and another odor (CS-) without shock.

To indicate learning performance, the Performance Index (PI) is used as a metric ranging from 0 to 100. The PI is defined by the relationship with P_-, representing the proportion of correct responses or desired outcomes. The calculation of PI is given by the equation:

$$PI = 100(2P_- - 1) \tag{2.15}$$

This formula allows us to translate the value of P_- into a scaled index that reflects the effectiveness of learning. After training, an immediate PI between 80 and 90 was observed, which diminished to approximately 5 after four days. Leveraging these findings, we formulated several assumptions to aid in the estimation of ARM retention:

1. The input sensory signal follows a normal distribution with a constant mean (μ) and variance (σ^2). As stated in the previous section, we establish $\mu = \sigma^2$ within our simulation framework.

2. The weights for LTM ($w_{\pm}^{LTM}$) and ARM ($w_{\pm}^{ARM}$) are initially set to 0.5 and 0, respectively. Since the weight varies between 0 and 1, initializing LTM at 0.5 enables both potentiation and depression during learning, providing a 50% probability that flies will approach the odor in the absence of learning.

Since ARM decays over time, it is assumed that no ARM is present before the learning process begins.

3. Following 10 cycles of massed training, the weight inducing avoidance behavior ($w_- = w_-^{ARM} + w_-^{LTM}$) reaches its maximum value. As the massed training protocol updates only ARM and not LTM, the change of the avoidance ARM weight is 0.5 after the massed training.

4. In the scenario of one-cycle training, the weight does not reach saturation, allowing us to estimate the change in avoidance ARM ($\Delta w^{ARM}-$) using the equation $\Delta w_-^{ARM} = \eta \Delta R \mu$. Here, η denotes the learning rate, ΔR represents the mean reinforcement obtained from the one-cycle training, and μ is the average value of the sensory input signal.

The Sensory Input Distribution

Building upon **Assumption 1**, we can determine the probability of odor avoidance (P_-) using Equation (2.2). Following **Assumption 2** and **3**, the weight values immediately following the massed training can be found by $w_+^{ARM} = 0$, $w_+^{LTM} = 0.5$, $w_-^{ARM} = 0.5$ and $w_-^{LTM} = 0.5$. This results in $w_+ = 0.5$ and $w_- = 1$. Based on empirical observations, we assume a PI of 85. Given that the relationship between PI and P_- is defined by Equation (2.15), we can infer that $P_- = 0.925$. Consequently, the mean and variance μ for the sensory input distribution $X \sim \mathcal{N}(\mu, \mu)$ can be calculated through Equation (2.2), yielding a value of $\mu = 10.36$. For our simulation, we round this figure to $\mu = 10$.

The ARM retention rate

Utilizing the calculated $\mu = 10.36$, we can subsequently determine the retention rate of ARM (γ_{ARM}). Given that the observed PI is 5 at four days after learning (Tully et al. 1994), we deduce that the corresponding P_- equals 0.525. This results in a w_- value of 0.514. By invoking **Assumption 2**, we have $\Delta w = 0.014$ four days after learning, as extrapolated from Equation (2.2). We then derive γ_{ARM} from $\Delta w^{1/4}$, yielding a value of 0.34.

The Learning Rate

The learning rate for this study was calibrated using the learning performance from one-cycle training. As this performance index post one-cycle training is not influenced by protein synthesis inhibition (Tully et al. 1994), it is plausible to assume the learning rates for LTM and ARM are identical. When the PI ranges from 60-80 within the first hour post one-cycle training. Considering a performance index of 70, we have $P_- = 0.85$. From Equation (2.2), the post-learning value of w_- is determined to be 0.8. Then the weight change of $\Delta w_-^{ARM} = 0.8 - 0.5 = 0.3$. Following **Assumption 4**, we know that $\Delta w_-^{ARM} = \eta \Delta R \mu$ after the one-cycle training, and considering that the probability of approaching the odor without learning is 0.5, we posit the expected reward before learning is 0, then the reward difference of this one cycle learning is $\Delta R = 0.5R$-0, then we have $\Delta w_-^{ARM} = 0.5R\eta\mu$. In this model, the value of the reinforcement aligns with the hazard of the stimuli (h_s). As previously discussed, this value ranges from 0 to 1. Given our estimated μ value of around 10, we reasoned that the hazard induced by the stimuli is not lethal and estimated $R\mu = 1$. Therefore, $\Delta w_-^{ARM} = 0.5\eta$, allowing us to calculate the learning rate for this scenario as approximately 0.6.

The Starvation Hazard Steepness

As previously mentioned, we assume that Drosophila will experience its natural lifespan (l_N) when the energy level remains at its maximum throughout its life. Therefore, by letting $M = 1$ in Equation (2.5), we derive $h_M(t) = \exp(-c_m)$. Substituting this into Equation (2.10) yields $c_m = \log(l_N)$. For the purpose of this study, we postulate that the natural lifespan (l_N) of the fruit flies is approximately 50 days, as supported by the experiment from Linford et al. (2013). Consequently, we compute the starvation hazard steepness c_m to be approximately 3.96.

List of Parameters

The model parameters used in this study can be summarized in Table 2.1:

Table 2.1: List of parameters

Name	Symbol	Value
Odor input	$x \sim \mathcal{N}(\mu, \sigma^2)$	$\mu = 10$, $\sigma^2 = 10$
Initial weight value (LTM)	$w_{\pm}^{LTM}(t=0)$	0.5
Initial weight value (ARM)	$w_{\pm}^{ARM}(t=0)$	0
Time constant of the average reward estimate	τ_R	10
ARM retention	τ_{ARM}	0.34
Learning rate	η	0.6
Starvation hazard steepness	c_m	3.96
Energy cost of LTM	c_{LTM}	0.21

2.5 Energy-Adaptive Learning

In the context of this research, we introduce a concept termed Energy Adaptive learning. This concept refers to the selection of a memory consolidation pathway based on metabolic energy levels. The regulation of this process is centered on the energy threshold (M_{th}), which can be applied in two distinct ways.

2.5.1 Fixed Threshold Model

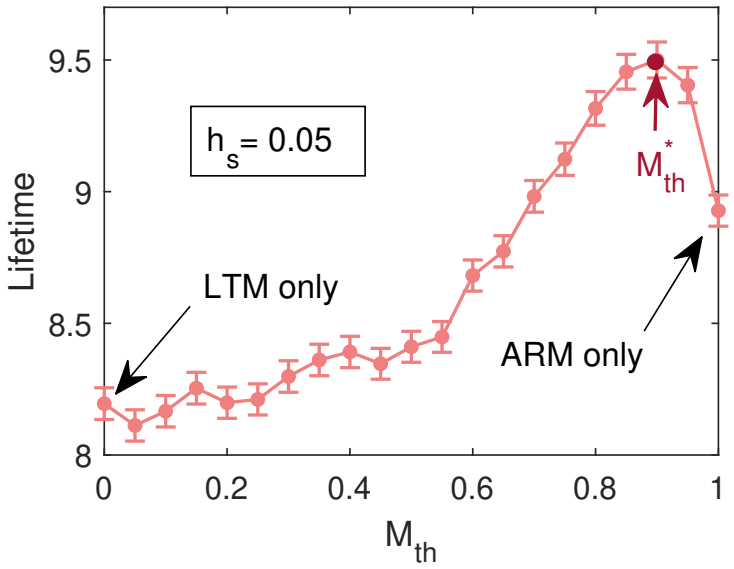

Figure 2.3: Variation in lifetime with different fixed energy thresholds. This result depicts the simulation of 10000 flies. The energy prior to learning was modeled by a uniform distribution ranging from 0 to 1. The error bars represent the Standard Error of the Mean (SEM) calculated from the data. Where the optimal energy threshold (M_{th}^*) is highlighted in red. The hazard of stimuli in this case is 0.06.

The first model of energy adaptive learning employs a Fixed Threshold (FT) approach, implying a static value for M_{th} that remains unchanged over time. This model was devised to investigate whether a basic energy regulation of the memory pathway exerts a positive influence on the learning and survival of Drosophila. At the same time, this model helps us find the optimal threshold (denoted as M_{th}^{*}) for situations with different hazards from the stimuli, giving us insight into how the brain might adjust memory pathways in response to various scenarios.

To investigate the optimal threshold, we postulate that the energy prior to learning is uniformly distributed from 0 to 1, and the optimal energy threshold is identified by the value that maximizes the expected lifetime of 10,000 flies. An example of lifetime variations with a hazard level (h_s) set at 0.05 corresponding to a threshold ranging from 0 to 1 is illustrated in Figure 2.3. In this instance, the value of the optimal threshold approximates 0.9. Furthermore, since the model employs LTM when the energy level surpasses the energy threshold (M_{th}), at extremely low thresholds, LTM learning will always be employed; at extremely high thresholds, all learning takes place via ARM.

The value of the fixed threshold with different stimulus hazards is shown in Figure 2.4 (a). For the minimal stimulus hazard, LTM is relatively so expensive that the average lifetime is even reduced compared to the no-learning condition. ARM-only learning comes at no energetic cost and so avoiding the stimulus always increases the lifetime. However, adequately tuned the adaptive model outperforms the fixed strategies. With the increase in stimulus hazard, the optimal threshold becomes fairly insensitive to the precise stimulus hazard and has a broad optimum (see Figure 2.4 b).

2.5.2 Moving Threshold Model

In the fixed threshold model, LTM occurs when the learning hardly reduces the hazard exposure, as happens late in learning. To further prevent the formation of LTM with low utility, specifically in late learning, we introduce the Moving Threshold (MT) model. Research suggests that Dopamine (DA) in the hippocampus has been found to influence memory persistence (O'Carroll et al. 2006, Bethus et al. 2010, Lisman et al. 2011). Similarly, DA has a role in shaping the formation of long-term

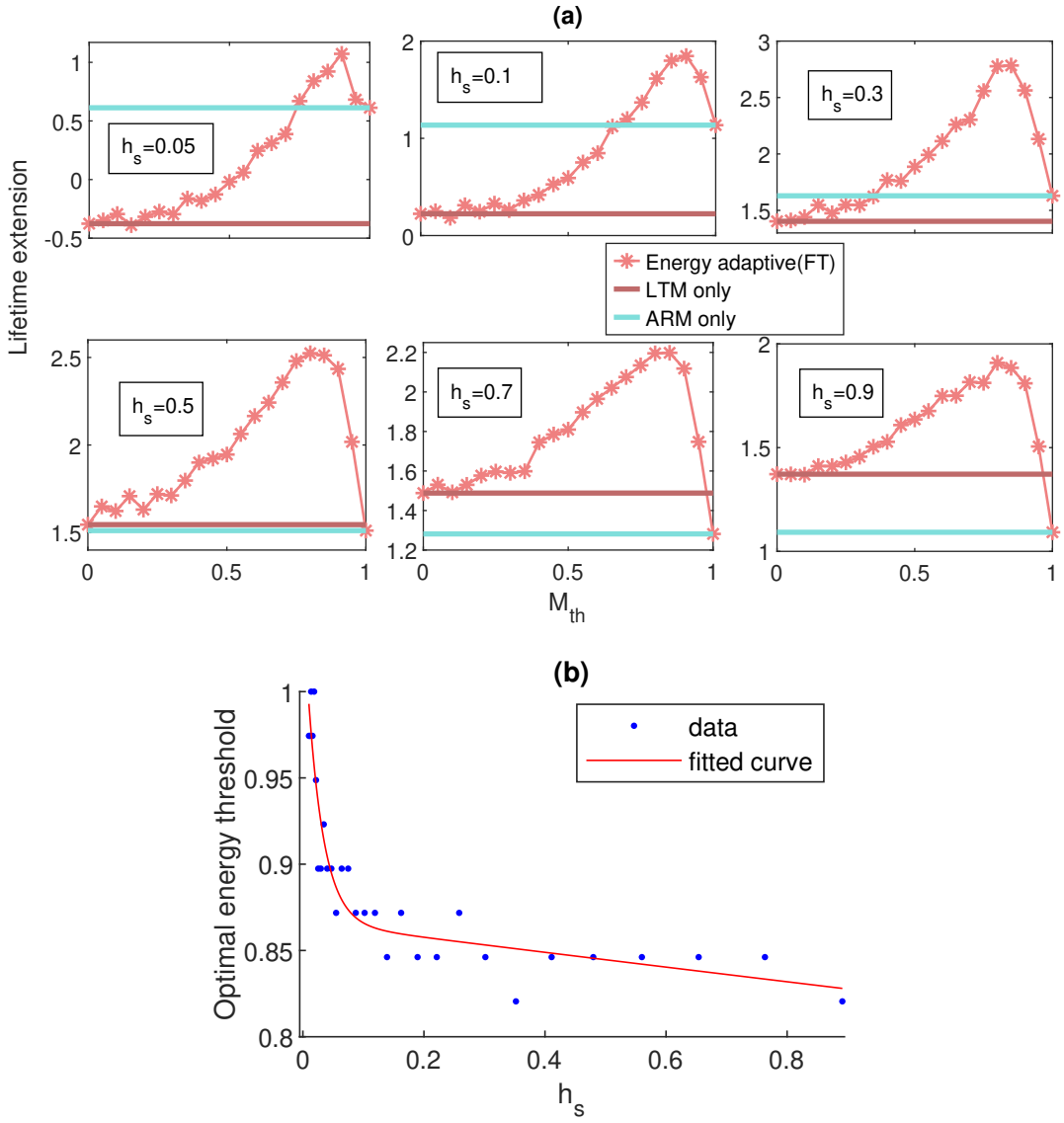

Figure 2.4: Variation in lifetime extension with different fixed energy thresholds. Where these results present the average lifetime of 10,000 flies, with a uniform distribution of initial energy reserves. (a) The lifetime extension (compared with no learning) of the FT model with different energy thresholds (M_{th}). (b) The optimal energy threshold of the FT model against the hazard of stimulus.

memory in Drosophila MB (Huetteroth et al. 2015, Placais et al. 2017). Extrapolating from these findings, we hypothesize the formation of LTM is postulated to be influenced by DANs.

In associative conditioning, the reinforcement signals including reward and punishment are modulated by DANs (Waddell 2013), we posit the reinforcement predict error is proportional to the DANs signal (D). Hence, the MT model proposes that the energy threshold is influenced by the difference between the current reward and a moving average of past rewards, which is encoded by DANs, as expressed by the

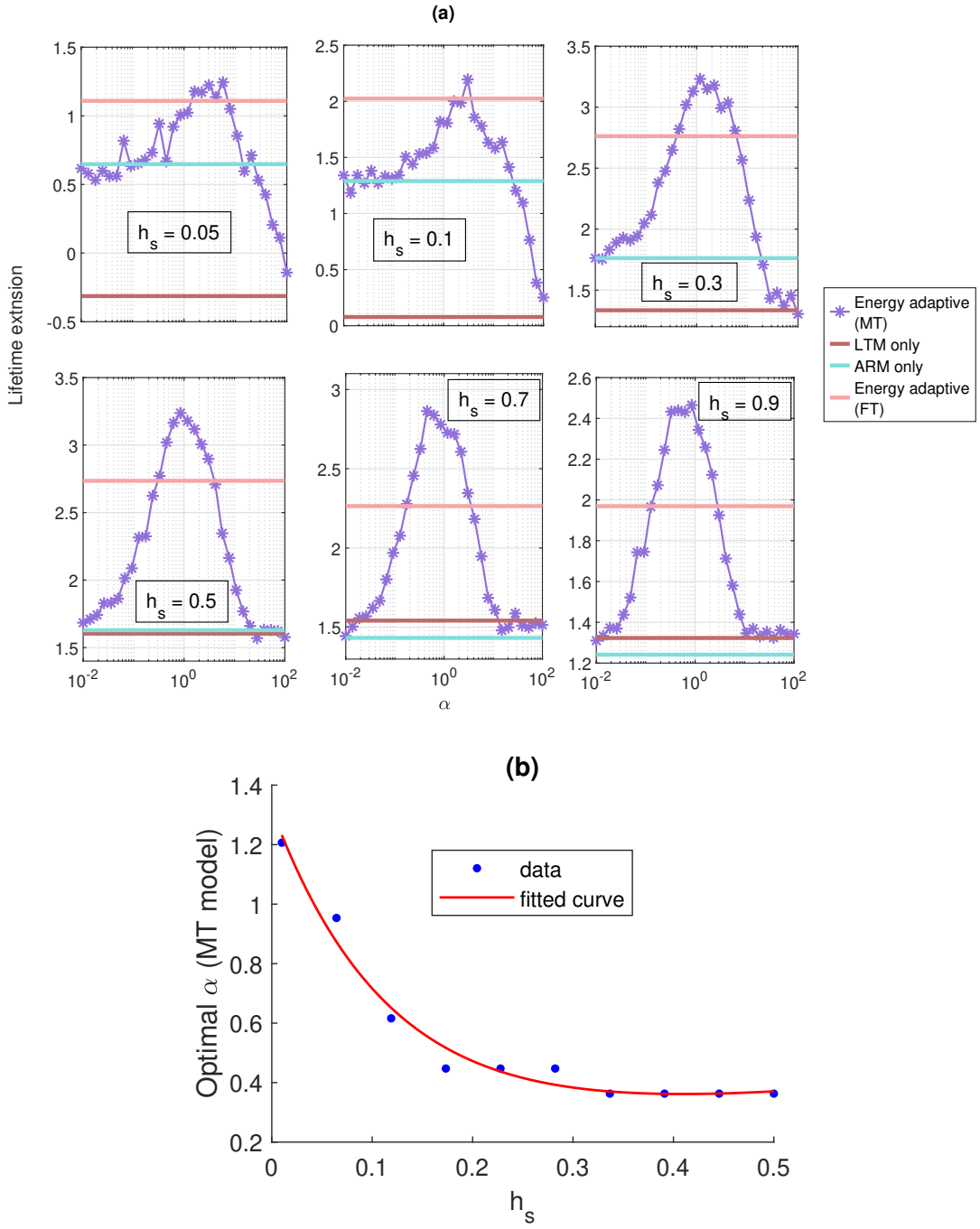

Figure 2.5: Lifetime extension in comparison to the no-learning scenario as a function of α. Compared with the lifetime extension observed when exclusively employing LTM, ARM, and the energy-adaptive MT model. This figure presents the average results obtained from a population of 10,000 fruit flies, wherein the energy levels prior to learning are uniformly distributed between 0 and 1. (a) The average lifetime time extension of the MT model, with the influence factor of DA (α) varying from 0.01 to 100. (b) The optimal α of the MT model with different stimulus hazards.

equation:

$$M_{th}(t) = 1 - \alpha D(t) \tag{2.16}$$

Where D is the DANs signal regulated by the difference between the current reward and its past moving average

$$D(t) = |R_i(t) - \bar{R}(t-1)| \tag{2.17}$$

Here, α is the influence factor of DA, with a larger α signifying a stronger DA influence on memory pathway regulation. When α is small, ARM is used even when the energy reserve is large. As α is increased, LTM is increasingly used when the reward/punishment is different from expected.

The MT model demonstrates the highest lifetime extension when α falls within a certain range, as illustrated in Figure 2.5 (a). Notably, as the stimulus hazards increase, the MT model displays increased robustness to variations in α, Figure 2.5 (b).

Furthermore, it is noteworthy that the MT model demonstrates an energy threshold that grows over time, with a more pronounced variation of M_{th} during the initial stages of learning (see Figure 2.6). This means more LTM is implemented in the early stages of learning. This feature is conducive to enhancing longevity due to the implications of reward discounting, that is, the value of the future reward in reinforcement learning is diminishing over time (Sutton & Barto 2018).

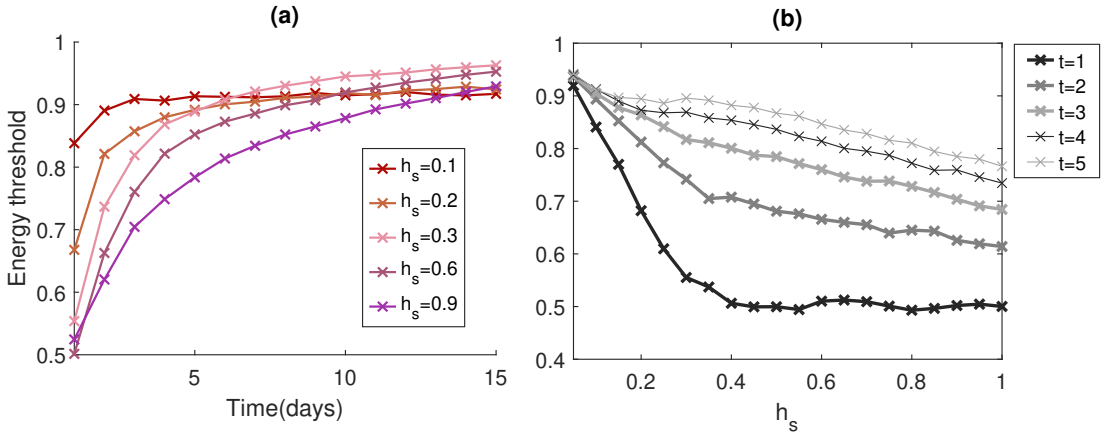

Figure 2.6: The energy-threshold of the MT model grows over time. (a) The threshold change of the MT model with different times after learning. (b) The threshold change of the MT model with different stimulus hazards.

2.6 Model Evaluation

In assessing the model's performance, we utilized the optimal energy threshold for the FT model and the optimal α for the MT model. The evaluation involved contrasting the performance of these energy-adaptive models with the models featuring a singular memory pathway.

2.6.1 Lifetime Extension Resulting from Energy Adaptive Learning

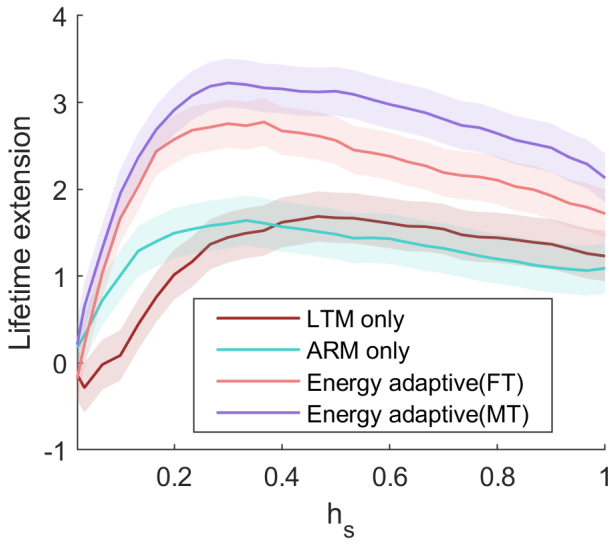

Figure 2.7: The lifetime extension resulting from the energy adaptive learning. Where this is the average lifetime extension of 2000 flies compared with the case of no learning. The initial energy levels prior to learning are uniformly distributed from 0 to 1. Shaded regions indicate the SEM.

In order to investigate the potential influence of energy adaptive learning on the lifespan of Drosophila, we conducted a simulation involving 2,000 fruit flies and implemented the model parameters calibrated by the experimental data, listed in Table 2.1. The initial energy levels prior to learning were assigned using a uniform distribution ranging from 0 to 1. In assessing the model's performance, we utilized the optimal energy threshold for the FT model and the optimal α for the MT model, as shown in Figure 2.4 (b) and Figure 2.5 (b). These lifetimes were subsequently compared with those resulting from the exclusive use of either the ARM or LTM pathway. As the hazard levels varied within the range of 0 to 1, the mean lifetime extension (defined as the additional lifespan compared to the case where no learn-

ing transpires) corresponding to different hazard levels (h_s) is depicted in Figure 2.7.

The results reveal that the lifetime associated with energy-adaptive learning outperforms that of the use of a singular memory pathway. This means the energy-regulated memory pathway optimizes the trade-off between learning and energy expenditure, aligning with experimental results (Plaçais & Preat 2013).

In the context of the testing setup depicted in Figure 2.7, we observed that the employment of ARM resulted in a prolonged lifespan when the reinforcement, or stimuli hazard, was minor. Conversely, utilization of LTM contributed to longer survival when faced with substantial reinforcement. An energy investment in learning to evade hazards is beneficial when those hazards could have severe consequences. This underscores the significance of the memory pathway regulation under varying levels of reinforcement strength.

Meanwhile, in Figure 2.7, the MT model with the optimal α outperforms the FT model with the optimal energy threshold in this regime. To delve deeper into the reasons behind the superior performance of the MT model and the underlying mechanisms of energy adaptive learning, we set the energy prior to learning, M_0, to 1 and traced the hazards from stimuli h_s, energy deficiency h_M, and the total hazard $h_{\sum}$. The results are shown in Figure 2.8 (a-d). These results demonstrate that energy adaptive learning manages to maintain a similar performance in avoiding the hazard as the case of solely using LTM (see Figure 2.8 b), while concurrently exhibiting reduced h_M during learning (see Figure 2.8 c). The energy adaptive method succeeds in conserving energy while preserving robust learning performance. This explains the observed increase in lifetime when employing energy-adaptive learning in comparison to solely using a single memory pathway. Furthermore, in Figure 2.8 (c), the MT method demonstrates superior energy conservation compared to the FT method. This accounts for the more pronounced lifetime extension observed in the MT method relative to the FT approach.

Also, exploring memory pathway utilization during learning reveals notable differences between the FT and MT models (see Figure 2.8 (e-f)). The MT model displays a significantly reduced usage of LTM compared to the FT model, while maintaining comparable hazard avoidance (see Figure 2.8 (b)). There are plausible reasons to

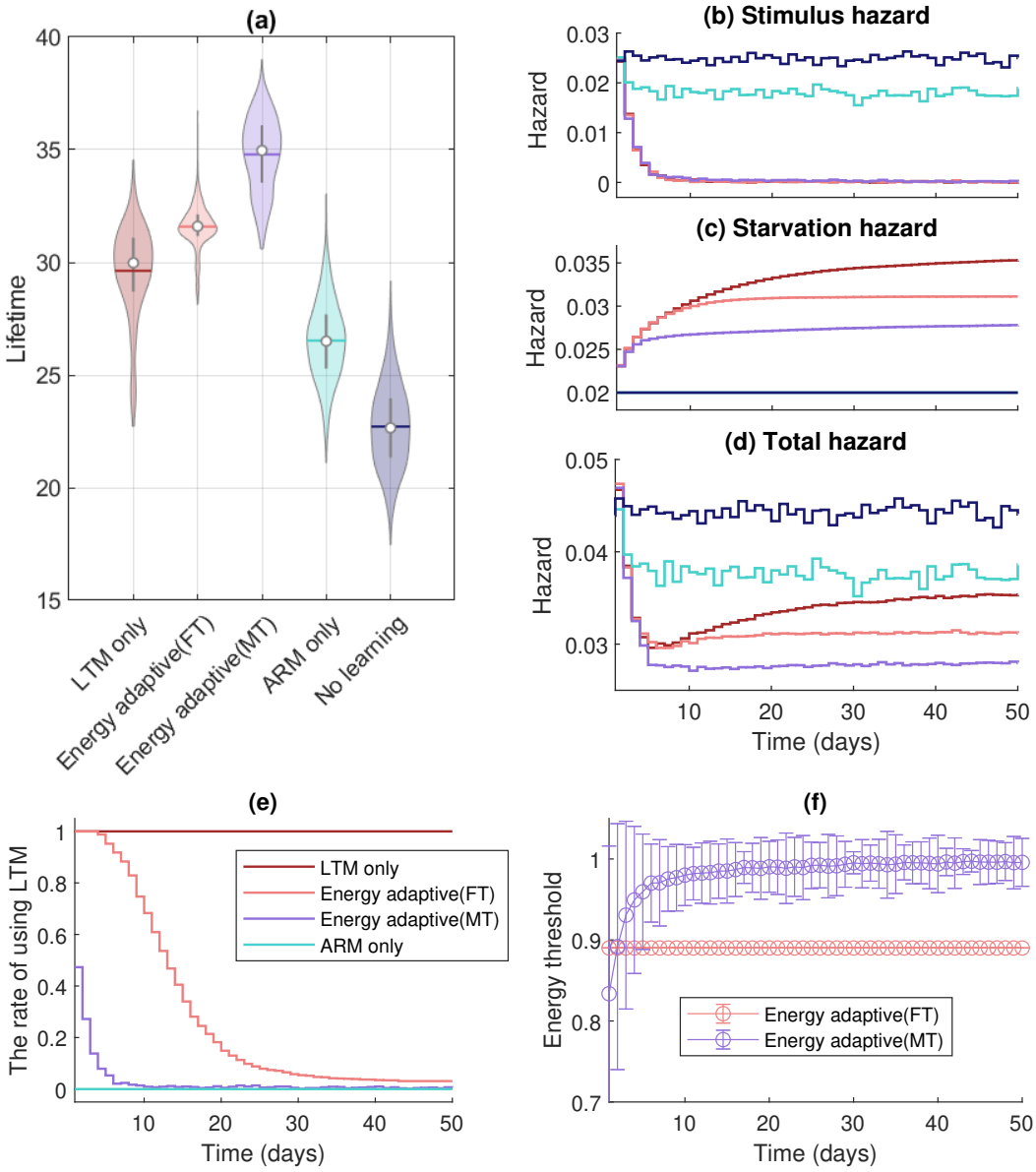

Figure 2.8: The hazard trace, the rate of using LTM and the energy threshold while learning. (a) The lifetime comparison. (b) The hazard from the stimuli. (c) The hazard of energy deficiency. (d) The total hazard. (e) the rate of using LTM. (f) The energy threshold of the energy adaptive models, where the error bar indicates the std error.

explain this: firstly, as illustrated in Figure 2.6, the sliding energy threshold in the MT model reduces the usage of LTM over time. Secondly, the individual-specific threshold in the MT model caters to each fly's distinctive learning regime, which halts superfluous LTM usage. Thus, the MT model harmonizes energy preservation and learning effectiveness in a more precise manner.

Moreover, we observed variations in the consistency of memory pathway regulation

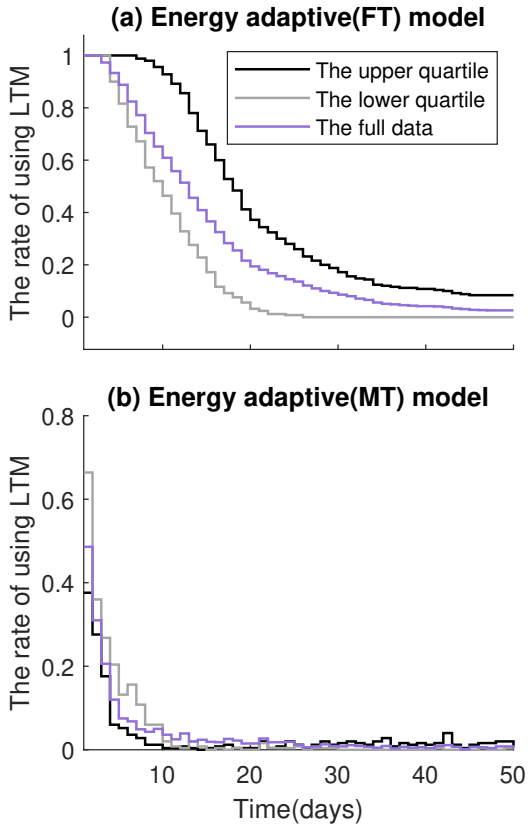

Figure 2.9: The difference of memory pathway selection of different quartiles. (a) Energy adaptive model with a fixed threshold. (b) Energy adaptive model with moving threshold.

between the FT and MT models. This was assessed by examining the rate of LTM utilization for flies within the upper and lower quartiles of lifetime, as illustrated in Figure 2.9. In the FT model, the LTM usage rate exhibits significant variation between flies from the upper and lower quartiles, whereas the difference in LTM usage in the MT model is less pronounced. Hence, the MT model is capable of regulating memory pathways in a more consistent manner.

2.6.2 Model Performance under Stochastic Environment

In a real-world scenario, the consistency of reinforcement is uncertain, in this section, we examined the model performance under stochastic stimulus conditions. This involved adjusting the hazard occurrence probability, P_h, which above was held constant at 1. Assuming the energy prior to learning is uniformly distributed between 0 and 1, we simulated the lifespan of 2000 flies. Energy adaptive learning extends lifespan over single memory pathway models, when reward probability ranges above

0.2, shown in Figure 2.10 (a-d). However, if P_h drops below 0.2, relying solely on ARM proves more beneficial, especially in low-hazard stimuli scenarios Figure 2.10 (f). In this case, the occasional appearance of the stimulus leads to a large dopamine signal, driving LTM. However, the cost for LTM now outweighs the rare stimulus occurrence. This finding is in accordance with prior research on Drosophila learning in stochastic environments, which posits that forgetting can be viewed as an optimally adaptive behavior in changing contexts (Brea et al. 2014).

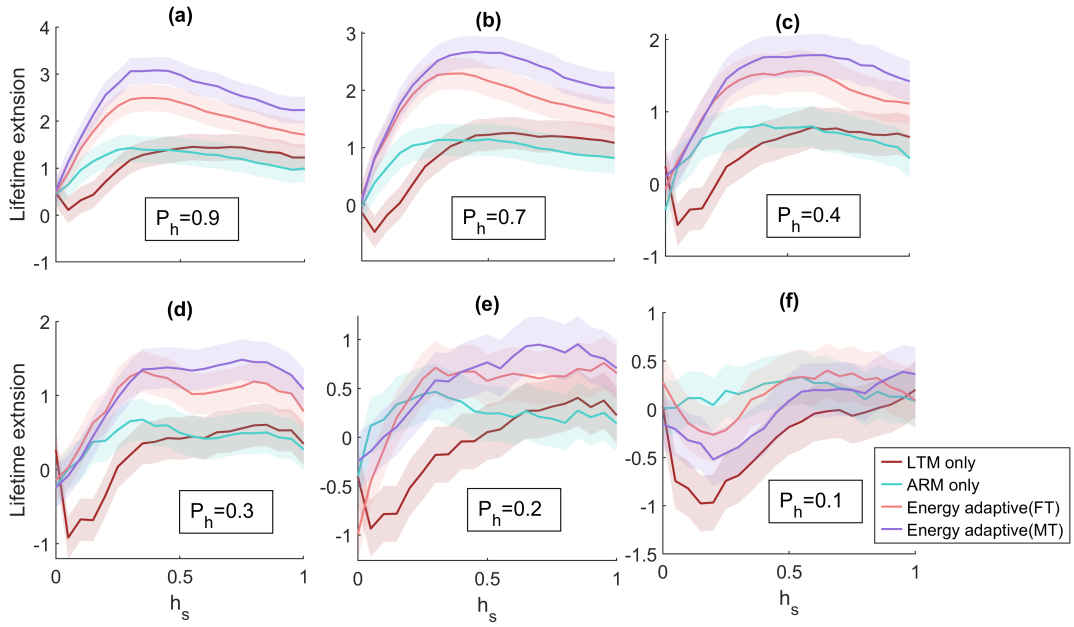

Figure 2.10: The lifetime extension compared to no-learning of 2,000 flies under stochastic conditions. The initial energy is uniformly distributed between 0 and 1. The probability of the occurrence of the hazardous stimulus (P_h) and the strength of the stimuli (h_s) were varied. The shaded region indicates SEM. For the FT model, the optimal threshold value is determined based on the level of h_s. In the MT model, the threshold value is proportional to the influence factor of DA, α, where α is optimized for the given h_s.

To understand the mechanisms underlying this phenomenon and identify the conditions favoring exclusive ARM use, we varied the initial energy (M_0) from 0 to 1 and compared the lifetime extension when solely using LTM and ARM. With the hazard fixed at 0.5, we observed the lifetime extension of the flies under different P_h. As depicted in Figure 2.11, we identified a crossover point in the lifetimes associated with exclusive use of LTM and ARM when P_h ranged from 0.4 to 0.2. Below this crossover point, it is advantageous to use ARM, whereas above it, LTM use is beneficial. As P_h decreases, this crossover point shifts towards higher values. Notably,

at $P_h = 0.1$, the crossover point vanishes, and using ARM consistently results in a longer lifespan across all M_0. This observation elucidates why ARM use is preferable in stochastic environments characterized by extremely low P_h.

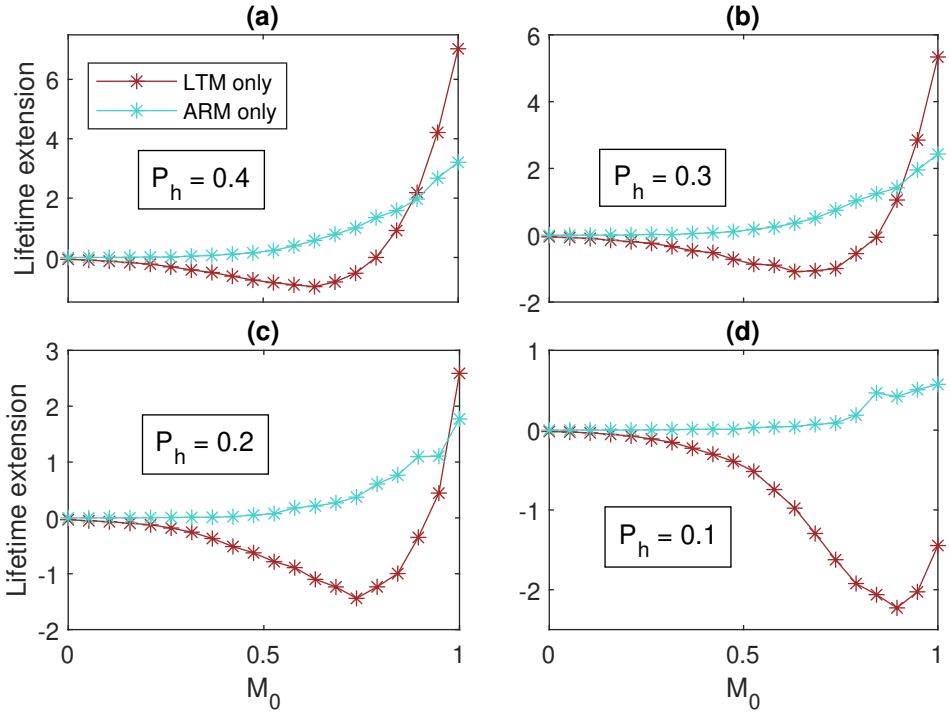

Figure 2.11: The lifetime extension of using LTM and ARM only under stochastic environment, varying the probability of the occurrence of electrical stimuli P_h. The hazard value from the stimuli is fixed at 0.5 in this example.

2.7 Discussion and Conclusion

This chapter introduces a computational model that explicates the energy-based regulation of learning and memory pathways in biological systems. The model, framed within the context of fruit flies, exhibits significant energy efficiency and adaptability to the aversive stimuli, leading to an extended lifespan in contrast to singular memory pathway models. This regulation extends the flies' lifetime in hazardous environments, aligning with empirical data (Plaçais & Preat 2013), thus suggesting potential energy-efficient memory regulation mechanisms in biological systems.

The dopamine-modulated MT model posits an energy-adaptive learning system that is regulated by reward prediction error signals originating from dopaminergic neurons. Within this framework, a more significant stimulus elicits a larger error signal,

thereby promoting the utilization of LTM. The superior performance of this model hints at the potential adaptability of biological memory systems to the intensity of dopaminergic signals, aligning with the existing research on dopamine's pivotal role in memory persistence, also in mammals (Lisman et al. 2011, O'Carroll et al. 2006, Bethus et al. 2010). Furthermore, the MT model further shows a decrease in superfluous LTM usage compared to the FT model, suggesting that the Dopamine-Modulated learning systems might implement strategies to avoid unnecessary engagement of energy-intensive LTM processes.

While the current study concerned insects, given the many parallels with mammals, the results have likely a much wider applicability. First, analogous to ARM and LTM, mammals express both transient early-phase and persistent protein synthesis dependent late-phase plasticity (although their pathways are not segregated like they are in flies). Second, under low energy conditions the persistent form, but not the transient form, is inhibited, suggesting a similar difference in energetics (Potter et al. 2010). Finally, in both insects and mammals the dopamine reward system plays a similar role in signaling reward prediction error and boosting persistent forms of plasticity. Hence, all necessary ingredients are also present in mammals. Yet the effect of metabolic state on animal (including human) reinforcement learning has to our knowledge not been extensively explored; this study suggests that this could be a fruitful research topic.

Our results demonstrate the robustness of energy threshold parameters in both the FT and MT models in response to variations in stimulus hazards, especially at higher hazard levels. This suggests that biological systems might demonstrate resilience to non-trivial stimuli, preserving a steady memory regulation mechanism under hazardous environments.

In both FT and MT models, we observed that the energy-adaptive memory regulation mechanisms oriented toward higher utilization of LTM during the early phases of the learning process. This feature is conducive to enhancing longevity due to the implications of reward discounting, that is, the value of the future reward in reinforcement learning is diminishing over time (Sutton & Barto 2018). This tendency of discounting the future reinforcement is also used to uncover the neurobiological mechanisms of animal risk behavior (Raiff & Yoon 2010) and reflect the urgency of

the reward (Carter & Redish 2016). As such, this research underscores the importance of initial decision-making and early LTM consolidation, thereby contributing significantly to the broader field of adaptive learning mechanisms.

Interestingly, under stochastic conditions, the model suggests that ARM becomes a more advantageous strategy when the probability of hazard occurrence is low. This implies that, in uncertain environments, biological systems might prefer less energy-intensive ARM processes to conserve energy. An algorithm that would only express LTM in response to consistently repeating associations, might perform better under these circumstances.

A key assumption in our model performance comparison is that the threshold controlling the ARM/LTM pathway selection is optimized to maximize memory lifetime. In the FT model, this optimal threshold is determined by the stimulus hazard level (h_s). In the MT model, the threshold is proportional to the influence factor of DA (α), which is also optimized for the given h_s. Future research could explore adjusting the FT threshold or MT α in response to varying hazard levels (h_s) or environmental stability, such as by developing models that dynamically adjust these parameters based on real-time assessments of environmental conditions.

Future research directions include expanding the model to incorporate biological factors such as age, genetic variations, and environmental stressors. This model could also be extended to other organisms, including humans, to yield valuable insights into memory regulation across species. Lastly, empirical validation could test and refine the model's predictions through experimental studies, thereby enhancing its biological plausibility.

Chapter 3

Weight Update Efficiency of the MB-based Bandit

3.1 Synopsis

This chapter assesses the effectiveness and efficiency of the MB's intrinsic learning mechanism and delves deeper into the varied functions of Dopamine Neurons (DANs) in RL. Employing a methodology similar to Chapter 2, we utilize a two-armed bandit model based on MB architecture to assess the efficiency of various synaptic weight update strategies, defining "Weight Update Efficiency" as the change in behavior per unit of synaptic weight adjustment. Our initial investigation into single-trial learning reveals a highly efficient method that reduces the strength of synapses associated with incorrect responses. In olfactory learning, DANs inhibit the activity of MB Output Neurons (MBONs) linked to incorrect actions, the finding of this single-trial learning suggesting that this biologically informed approach to synaptic weight adjustment is particularly effective. Subsequently, in a multi-trial learning context, we compare three synaptic update methods: one that depresses synapses for incorrect responses, another that potentiates synapses for correct responses, and a combined method that both potentiates correct synapses and depresses incorrect ones. Since DANs not only modulate synaptic strength in response to reinforcement signals but also integrate energy status into the decision-making context, we include energy signals as contextual information received by DANs, biasing the behavioral choices. The findings reveal that the weight update strategy mimicking the natural learning mechanisms within the MB outperforms the alternative strategies in weight update efficiency. This advantage is maintained when incorporating energy

considerations as a contextual bias of decision-making.

3.2 Introduction

As introduced in Section 1.1, within the MB, Kenyon Cells (KCs) receive signals from multiple olfactory glomeruli via the MB calyx. These KCs relay this odor sensory information through the MB lobes to the MBONs, which subsequently influence odor-induced behavior (Turner et al. 2008, Campbell et al. 2013, Cohn et al. 2015, Hige et al. 2015, Cognigni et al. 2018). DANs modulate the KC-MBON synapses in response to reinforcement signals, playing a pivotal role in olfactory conditioning (Waddell 2013, Owald & Waddell 2015).

A distinct configuration in the MB lobe involves DANs and MBONs paired based on opposing valence. Reward-associated DANs are linked to avoidance MBONs and vice versa (Aso et al. 2014*b*, Amin & Lin 2019). In appetitive conditioning with positive reinforcement, the engagement of DANs suppresses the KC-MBON connection which drives avoidance behavior (Owald et al. 2015). In contrast, the aversive conditioning with negative reinforcement dampens the feedforward inhibition in an MBON, which, in the context of appetitive conditioning, typically facilitates approach behaviors (Perisse et al. 2016). Upon receiving reinforcement signals, DANs inhibit the KC-MBON synapses inducing the wrong behavior, thereby increasing the tendency to select the right behavior (Cohn et al. 2015, Hige et al. 2015, Owald et al. 2015).

Ideally, both enhancing synapses associated with correct behaviors and weakening those linked to incorrect ones enable fruit flies to make accurate decisions after learning. We propose that the efficiency of different weight update strategies varies, with this efficiency assessed by examining the behavioral change per unit of synaptic weight change. Our hypothesis suggests that fruit flies preferentially weaken incorrect synaptic connections due to higher efficiency. To explore this, we implement a two-armed bandit model mirroring the MB structure, initiating with a simple scenario where learning concludes in a single trial. Here, a predetermined amount of synaptic weight change is available, and this change can adjust weights associated with all behaviors. The optimal learning is defined as achieving the highest performance using this fixed weight change. Our findings suggest optimal learning efficiency is achieved when the entire weight change is allocated to depressing incorrect weights. In a more complex learning scenario involving multiple trials,

we assessed three synaptic modification strategies: depressing incorrect KC-MBON synapses, potentiating correct KC-MBON synapses, and a dual approach modifying both types of synapses simultaneously. Our results indicate that focusing on depressing incorrect synapses offers an advantage in efficiency.

Moreover, in the behavioral adaptation to olfactory conditioning in Drosophila, DAN neurons respond to more than just reinforcement signals, they also account for contextual information derived from internal state and external factors (Cohn et al. 2015), such as reproductive status (Lin 2023), mating (Boehm et al. 2022), satiety (Kim et al. 2007, Tsao et al. 2018, Zolin et al. 2021) and airflow (Zolin et al. 2021).

Of all contextual factors influencing decision-making, energy stands out as a pivotal element, explored across molecular to behavioral dimensions. The metabolic energy is deeply linked with Drosophila's hunger and satiety states, influencing both sensory systems and memory formation (Kim et al. 2007, Tsao et al. 2018, Cognigni et al. 2018). Specifically, starvation enables flies to quickly associate specific odors with sugar rewards, forming appetitive olfactory memories (Kim et al. 2007, Tsao et al. 2018, Cognigni et al. 2018, Zolin et al. 2021). Also, during appetitive memory retrieval, the fly's energy state shapes odor-driven behavior by routing the hunger-indicative dNPF signal via MP1 DANs. Hungry flies see an activation of the positive-valence MVP2 MBON under MP1 DANs' influence. In contrast, satiated flies experience suppressed MVP2 MBON activation, with an increased activity in the opposing negative-valence M4/6 MBONs, resulting in avoidance behaviors (Cognigni et al. 2018). This complex neural network ensures that flies respond to food-associated memories appropriately, especially when hungry, showcasing the fly's inherent mechanism to align behavior with physiological needs.

Hence, we propose that DAN neurons consider energy-related signals as a key contextual cue, incorporating energy level as a contextual influence that biases the decision-making in the learning model. Initially, we employ single-trial learning with a fixed weight change to evaluate the efficiency of weight adjustments under different biases. The results demonstrate that depressing the weight associated with incorrect responses continues to be the most effective strategy when the bias increases the probability of making the desired decision. Moreover, as the likelihood of selecting the correct option increases with decreasing energy levels, employing

multiple-trial learning with three distinct synaptic weight update strategies further indicates that weakening the incorrect synaptic connections yields the highest efficiency in weight adjustment. These findings suggest that the biologically plausible strategy for updating synaptic weights remains most effective when decision-making is influenced by energy considerations.

3.3 Basic Model

Drawing inspiration from the Drosophila MB's architecture, we designed an olfactory learning model suitable for both appetitive and aversive conditioning. Herein, the sensory inputs are encoded in KCs, which interface with MBONs governing approach and avoidance behaviors. The synaptic strengths between KCs and MBONs are modulated by DANs of opposing valence. We denote MBONs associated with approach and avoidance as $MBON^{+}$ and $MBON^{-}$, respectively. Similarly, DANs responding to reward and punishment are represented as DAN^{+} and DAN^{-}. The intricate network structure is detailed in Figure 3.1.

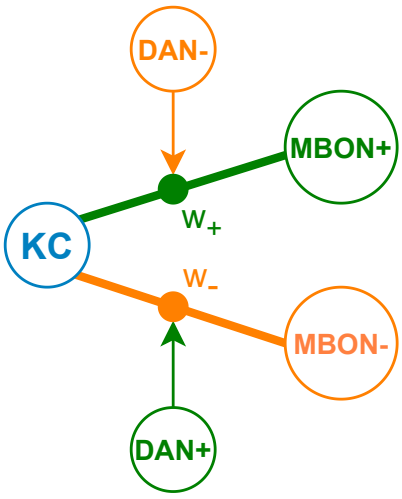

Figure 3.1: The structure of the basic olfactory learning in MB. The odor inputs stimulate the KCs. Behavior, whether approach or avoid, is modulated via the synapse connecting the KCs to $MBON^{+}$ or $MBON^{-}$. The strength of these synapses is influenced by DAN^{-} and DAN^{+} respectively, which correspond to punishment and reward.

Upon receiving an odor input, the KCs are activated. We model the KC signals using a normal distribution for each action, characterized by a mean $\mu = 1$ and standard

deviation $\sigma = 0.2$. By introducing this stochastic component, the decision-making process achieves a balance between the exploration and exploitation trade-offs during learning. The possible actions, denoted as a^+ for approach and a^- for avoid, are represented in $\mathbf{a} = [a^-, a^+]$. The chosen action a from this vector is determined by

$$a = \arg\max_i \mathbf{Q} \tag{3.1}$$

where the activities of the MBON neurons, $\mathbf{Q} = [Q^-, Q^+]$, are driven by the sensory inputs $\mathbf{x} = [x^-, x^+]$ from the KCs, and the synaptic weights between KCs and MBONs, $\mathbf{w} = [w^-, w^+]$

$$\mathbf{Q} = \mathbf{w} \circ \mathbf{x} \tag{3.2}$$

where $\circ$ denotes the element-wise product of the vectors. w^- and w^+ denote the strengths of the KC-MBON synapse for avoiding and approaching behaviors respectively, with an initial value set to 0.5 for both. In this research, the synaptic weight values are constrained within a range of 0 to 1.

In this model, it is hypothesized that the KCs introduce a noisy input, where both (x^-) and (x^+) are independently sampled from a Gaussian distribution, specifically $x^i \sim \mathcal{N}(\mu, \sigma)$ for $i \in \{-, +\}$. Note, unlike the approach in Chapter 2, this chapter does not account for memory retention or the selection among different memory pathways; here, memories are assumed to be non-decaying.

When the weights inducing approach w^+ and avoid w^- behavior are initialized at 0.5, the probability of approach/avoid without learning is 50%. Similar to the method used in Section 2.3.1, the probability of approaching given the weight values can be found as $P(w^- x^- < w^+ x^+)$. Let $Z = w^- x^- - w^+ x^+$, its mean μ_Z and variance σ_Z^2 are given by $\mu_Z = (w^- - w^+)\mu$, and $\sigma_Z^2 = [(w^-)^2 + (w^+)^2]\,\sigma^2$. Then the probability of approaching becomes $P(Z < 0)$. Since Z follows a standard normal distribution, we can express this probability as:

$$\begin{aligned} P(Z < 0) &= P\left(\frac{Z - \mu_Z}{\sigma_Z} < -\frac{\mu_Z}{\sigma_Z}\right) \\ &= \frac{1}{2}\left[1 + \operatorname{erf}\left(-\frac{\mu_Z}{\sqrt{2}\sigma_Z}\right)\right] \end{aligned}$$

Substituting the values for μ_Z and σ_Z, the probability of taking approaching action given the weight values can be found by

$$\begin{aligned} P_+(w^+, w^-) &= P(Z < 0) \\ &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{-\mu(w^- - w^+)}{\sigma\sqrt{2}\sqrt{(w^+)^2 + (w^-)^2}}\right)\right] \end{aligned} \tag{3.3}$$

3.3.1 Weight Update Efficiency of Single-trial Learning

To understand the way to produce the desired behavior changes with minimal modifications to the neural network, we evaluate the learning efficiency with different weight updating methods. In the context of this study, the "Weight Update Efficiency" is defined as the change in behavior (measured by the percentage of taking optimal actions) per unit change in the weight.

We first analyze a simplified scenario where learning is completed after a single trial, with a desired behavior of "approach". Initially, synaptic weights before this one-step learning are indicated as w_0^+ for approach activities and w_0^- for avoidance activities. The magnitude of the synaptic weight change in this one-step learning is denoted by a non-negative value, Δw^{Σ}. This model posits that the change in synaptic weights contributes to modifications in both approach (w^+) and avoidance (w^-) behaviors. Given a fixed total weight change Δw^{Σ}, the division of this change into w^+ and w^- is represented by $\alpha\Delta w^{\Sigma}$ and $(1-\alpha)\Delta w^{\Sigma}$, respectively. Where α represents the synaptic adjustment ratio, ranging from 0 to 1, it quantifies the proportion of the overall weight modification allocated to adjusting the weight associated with the desired behavior.

In this scenario, the magnitude of weight adjustment contributes to either synaptic potentiation or depression. When putting w^+ and w^- in a Cartesian coordinate system, the total weight change Δw^{Σ} can be conceptualized as the Manhattan distance (i.e. the sum of the absolute differences of their coordinates) between the initial and adjusted weights. As illustrated in Figure 3.2 (a), when the initial weight values are fixed, all the coordinates with a certain amount of Manhattan distance of Δw^{Σ} form a diamond-like shape with four edges, where each edge leads to a distinct

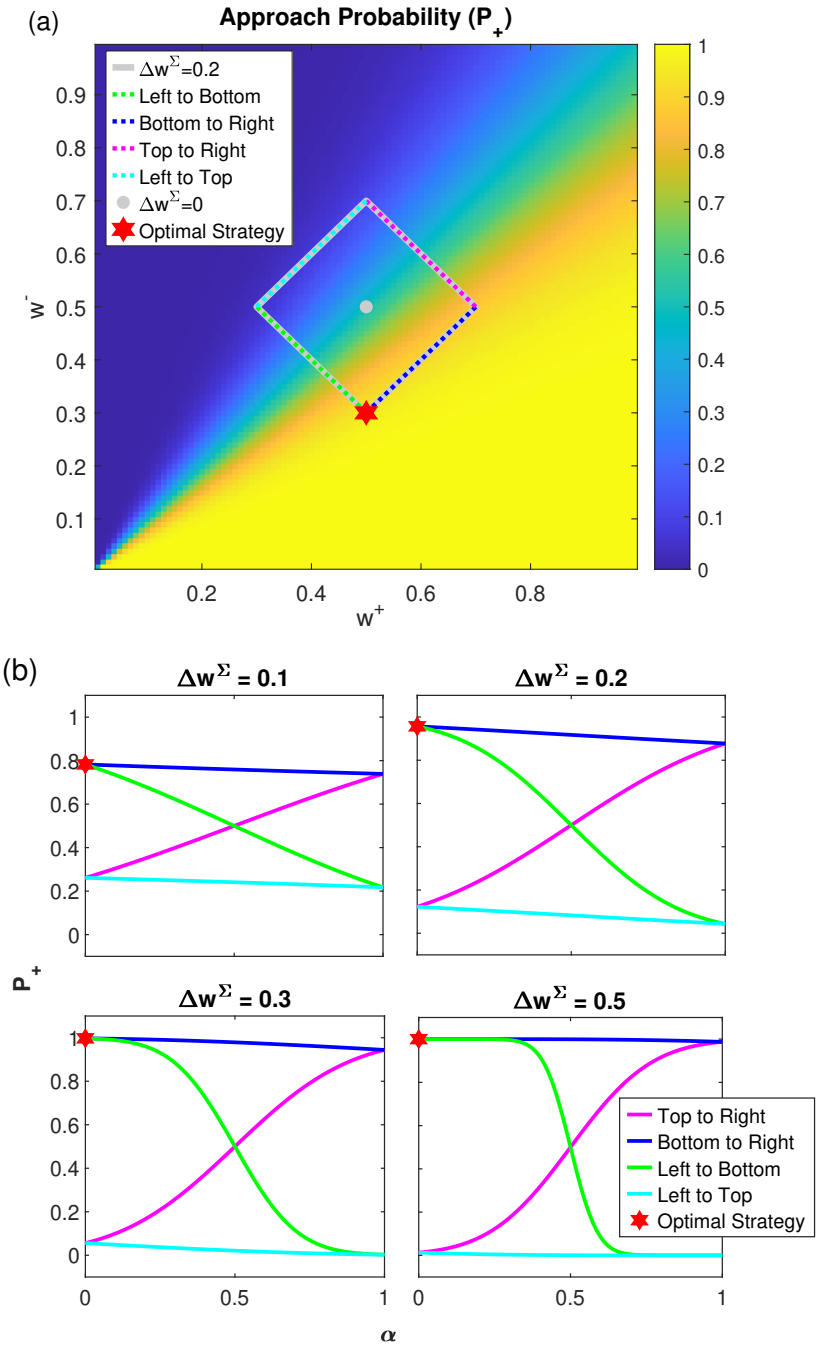

Figure 3.2: Approach probability P_+ and learning efficiency with fixed weight change. (a) Approach probability P_+ for varying values of w^+ and w^-. The gray contour delineates the region where the weight change Δw^{Σ} remains constant at 0.2, with the initial weight value denoted by a gray dot. The optimal weight update strategy, yielding the highest approach probability, is indicated with a red star. (b) Evolution of the approach probability along each edge of the contour from (a), where $\Delta w^{\Sigma} = 0.1$, 0.2, 0.3, and 0.5. The x-axis indicates the ratio α of the weight change directed towards the desired behavior. The optimal weight update strategies with the highest approach probability are emphasized by red stars, these strategies are characterized by the exclusive allocation of all weight modification efforts towards depressing the synaptic weight associated with undesirable behavior, w^-.

combination of synaptic adjustments:

- Top to right edge: Potentiate both w^+ and w^-.
- Bottom to right edge: Potentiate w^+ and depress w^-.
- Left to bottom edge: Depress both w^+ and w^-.
- Left to top edge: Depress w^+ and potentiate w^-.

After learning, the weight values on these four edges can be found as:

- Top to right edge:

$$w^+ = w_0^+ + \alpha \Delta w^{\Sigma}$$

$$w^- = w_0^- + (1-\alpha)\Delta w$$

- Bottom to right edge:

$$w^+ = w_0^+ + \alpha \Delta w^{\Sigma}$$

$$w^- = w_0^- - (1-\alpha)\Delta w$$

- Left to bottom edge:

$$w^+ = w_0^+ - \alpha \Delta w^{\Sigma}$$

$$w^- = w_0^- - (1-\alpha)\Delta w$$

- Left to top edge:

$$w^+ = w_0^+ - \alpha \Delta w^{\Sigma}$$

$$w^- = w_0^- + (1-\alpha)\Delta w$$

When the total weight change varies, the probability of approach with different weight update strategies on these four edges are shown in Figure 3.2 (b).

To explore the weight update that optimally distributes changes, thereby effectively improving the probability of choosing the correct action, we adapt Equation (3.3) to align with the dynamics at the four edges of the diamond-shaped contour. This approach allows us to examine the variations in P_+, the probability of taking the desired action, in correlation with the initial weight values, the overall weight modification Δw^{Σ}, and the synaptic adjustment ratio α indicating the fraction of synaptic modifications directed towards the desired behavior. A mathematical expression for

this relationship positioned at the top right edge is detailed in Equation (3.4), and the formulas for the four edges are outlined in Appendix B.

$$P_+(\alpha, \Delta w^{\Sigma}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu\left((2\alpha-1)\Delta w^{\Sigma} - w_0^- - w_0^+\right)}{\sqrt{2}\sigma\sqrt{((1-\alpha)\Delta w^{\Sigma} + w_0^-)^2 + (\alpha\Delta w^{\Sigma} + w_0^+)^2}}\right)\right] \tag{3.4}$$

By undertaking this analytical approach, we can investigate the impact of diverse patterns of weight adjustment on decision-making. To understand the learning efficiency across different magnitudes of weight updates, the probability of approaching the desired outcome P_+ as a function of the total weight change Δw^{Σ} is depicted in Figure 3.2 (a, b), the optimal weight update strategy is allocating all the weight change to depress the weight associate with the wrong action, w^-, yields the highest P_+, and this remains consistent across various total weight change.

The observation that solely depressing, the synaptic weight associated with undesirable outcomes, consistently results in the highest probability of achieving the desired 'approach' behavior is significant. This effect remains robust across variations in the mean and standard deviation of the sensory input signal $\mathbf{x}$, as illustrated in Figure 3.3.

Notably, in Figure 3.2 (b), when depressing only the weight linked to the undesired action yields the best learning performance, the approach probability on the bottom-to-right edge—where w^+ is potentiated and w^- is depressed—shows a very similar probability of approach. This similarity fluctuates with changes in the mean and variance of the input signal distribution.

Based on these findings, we propose a hypothesis within the framework of associative conditioning: a learning strategy that focuses exclusively on depressing the synaptic weight associated with undesirable outcomes may represent the most effective approach to facilitating learning.

Moreover, when the initial weight values w_0^+ and w_0^- are set to 0.5, for the four synaptic adjustment strategies under examination, the sign of $\frac{\partial P_+}{\partial \alpha}$ remains unchanged

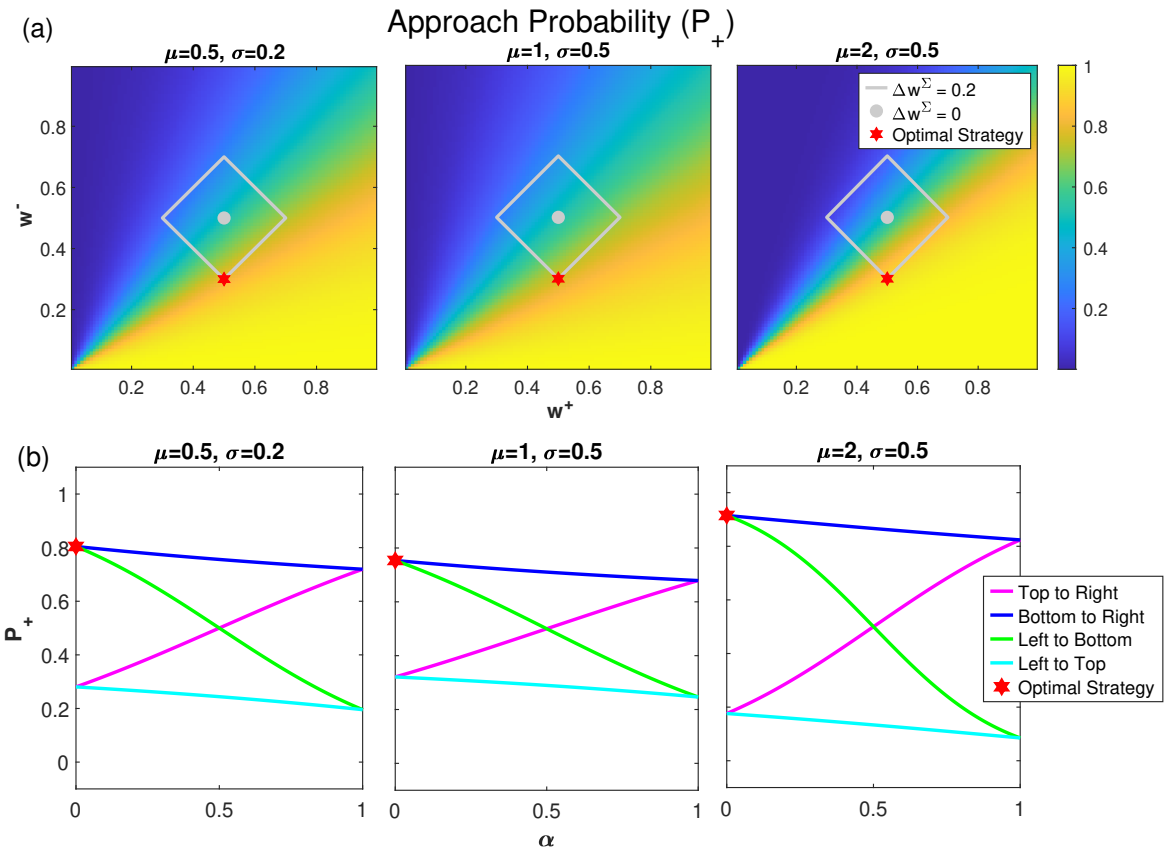

Figure 3.3: Variation in approach probability P_+ and learning efficiency across sensory input $\mathbf{x}$, characterized by differing means and standard deviations. (a) Approach probability P_+ for varying values of w^+ and w^-. The gray contour highlights the region where the weight change Δw^{Σ} remains at 0.2, with the initial weight value denoted by a gray dot. (b) The approach probability along each edge of the contour. Where the x-axis indicates the ratio α of the weight change directed towards the desired behavior. The optimal strategies are signified by red stars, these strategies allocate all the weight change to depress w^-.

across the stipulated ranges of α. This consistency reveals that, as illustrated in Figure 3.2 (a), the extrema consistently occurs at the corners. Consequently, in scenarios without prior bias before this single-trial learning, the optimal learning strategy exclusively employs the total weight change for the potentiation or depression of a singular synapse. This analysis is documented in Appendix B.

3.3.2 Weight Update Efficiency of Multi-trial Learning

Unlike the single-trial learning scenario that exclusively focused on approach as the desired behavior, this comprehensive analysis extends to include both appetitive and aversive conditioning in a multi-trial learning context.

Weight update Strategies

In the MB, the synaptic connections from KCs to the MBONs that induce approach behavior (termed here as MBON^{+}) are modulated by DANs that receive aversive signals (termed here as DAN^{-}). When flies receive punishment concurrent with odor presentation, the inclination to approach the odor diminishes. Conversely, when flies are rewarded with the odor presentation, the synaptic strength between KCs and the avoidance-inducing MBON^{-} reduces. This process primarily depresses the synapses responsible for the undesired behavior. We term this the "Depression" strategy in this study.

We introduce two alternative strategies for comparison. The first, named the "Potentiation" strategy, deviates from merely suppressing the "wrong" connections and focuses on enhancing the synapses that guide beneficial behavior. The second, the "Mixed" strategy, simultaneously potentiates the synapses responsible for the correct behavior and depresses those leading to the erroneous action, with the weight changes equally distributed between depression and potentiation. The conceptual framework for these methods is visualized in Figure 3.4.

In each trial, the weight change Δw is determined by the discrepancy between the actual reward received, R_t, and the estimated reward for the approach action, $\hat{R}_+$

$$\Delta w = \eta\mu(R_t - \hat{R}_+) \tag{3.5}$$

Where η is the learning rate, fixed at 0.1. Note, that the synaptic plasticity typically also relates to the pre-synaptic signal, here we suppose the weight change is proportional to the mean value of the sensory input μ, in this simulation, this value is 1. The reward estimation for approaching, initialized to zero, is incrementally updated each time the "approach" action is taken:

$$\hat{R}_+ \leftarrow \hat{R}_+ + \frac{r_+ - \hat{R}_+}{n_{a+} + 1} \tag{3.6}$$

Here, n_{a+} denotes the count of instances in which the fly opts to approach the odor. Given that the weight change for each trial is denoted by $\Delta w(t)$, when the learning

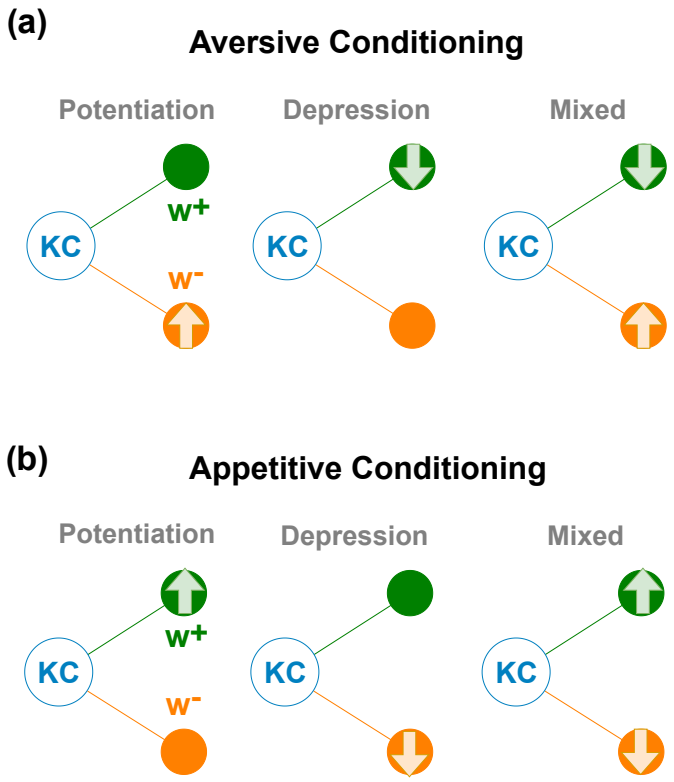

Figure 3.4: Weight update strategies for aversive and appetitive conditioning. (a) Aversive conditioning strategies. (b) Appetitive conditioning strategies. In both plots, the green and orange circles symbolize the KC-MBON weights influencing approach and avoidance behaviors, respectively. Arrows pointing upwards (downwards) signify potentiation (depression).

scenario involves multiple trials, the cumulative weight change ΔW_{cum} up to trial t is given by

$$\Delta W_{cum}(t) = \sum_{i=1}^{t} |\Delta w(i)| \tag{3.7}$$

By analyzing the cumulative weight change $\Delta W_{cum}(t)$ in relation to the percentage of desired actions (referred to as Performance, P^*) selected up to trial t, we establish a method to evaluate learning efficiency.For aversive conditioning, the desired action is "avoid," represented as $P^* = 1 - P_+$. For appetitive conditioning, the desired action is "approach," represented as $P^* = P_+$, where P_+ is determined by Equation (3.3). Every time the fly receives a reinforcement signal, the fixed values applied to the weight change will be obtained by Equation (3.5), regardless of the weight change strategy used. This standardized adjustment provides a direct comparison of the learning enhancements made by a consistent weight modification across diverse strategies. Furthermore, it facilitates an assessment to determine whether the MB's

biologically-inspired method indeed offers the highest efficiency in terms of learning.

3.3.3 Learning Performance

In this study, we conducted simulations of the learning processes for 5,000 fruit flies across 100 trials to evaluate the learning efficiency of a basic olfactory learning network employing various synaptic weight update strategies.

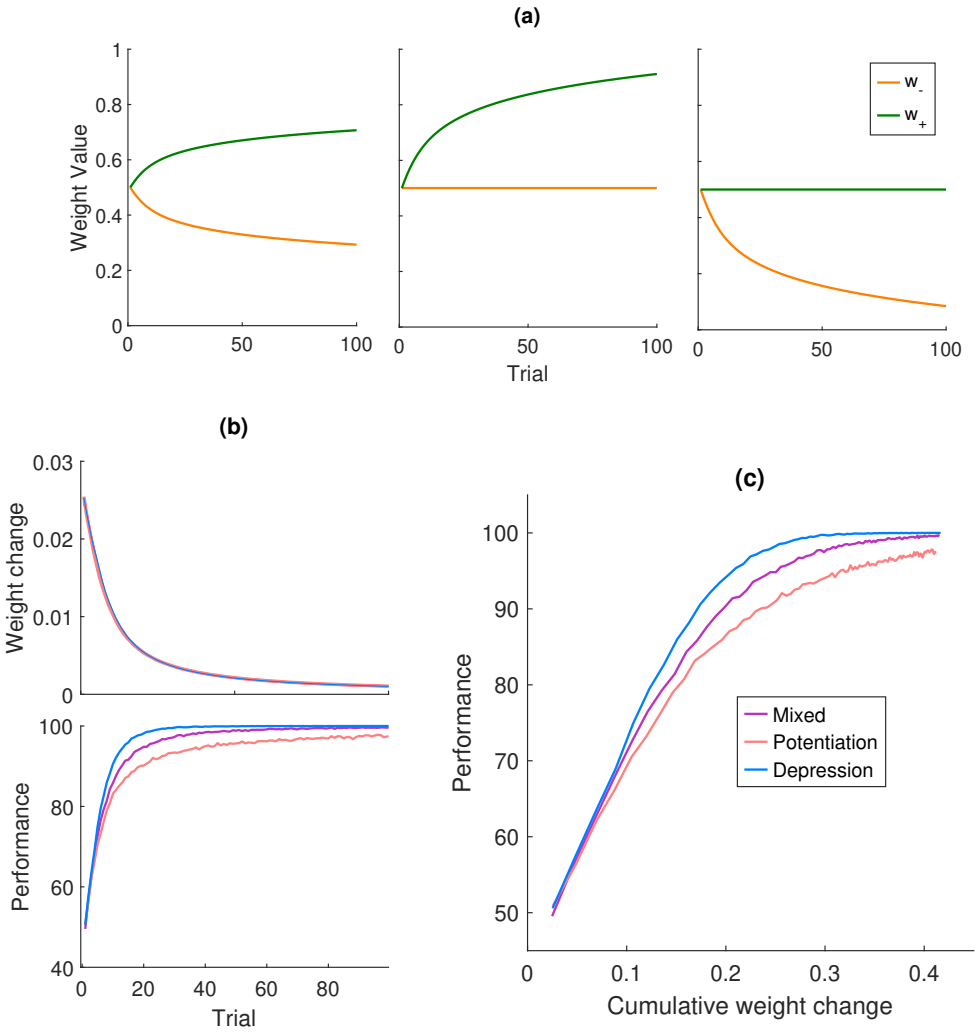

Figure 3.5: The Learning Efficiency for Appetitive Learning. (a) Average weight value over trials for different strategies: "mixed" (left panel), "potentiation only" (middle panel), and "depression only" (right panel). (b) Average weight change and the percentage of taking the optimal action, denoted as Performance, across trials. (c) Relationship between the performance and cumulative weight change.

In appetitive learning, the fly receives positive reinforcement upon approaching the

odor. To learn the desired behavior, flies either potentiate the approach weight or depress the avoid weight. The weight changes associated with the three distinct learning strategies are depicted in Figure 3.5 (a). Despite differences in potentiation/depression strategies, the magnitude of weight changes per trial is consistent across the strategies, as mirrored by the weight change over time in Figure 3.5 (b).

Given the uniform weight alteration per trial, the depression-only method emerges as the most efficient, yielding the highest approach percentage. This is evident both in the approach percentage over time, shown in Figure 3.5 (b), and the comparison between the approach percentage and cumulative weight change in Figure 3.5 (c). These findings underscore the superior weight change efficiency of Drosophila's innate weight-updating strategy.

In aversive learning, flies acquire negative reinforcement upon approaching the odor. Consequently, learning the optimal behavior entails either potentiating the avoid weight or depressing the approach weight. These weight changes for the three strategies are illustrated in Figure 3.6 (a). Similarly to appetitive learning, the weight change per trial is consistent among strategies, evident from the weight change progression in Figure 3.6 (b).

Results for aversive conditioning are presented in Figure 3.6. With constant weight changes in every trial, the depression-only technique registers the maximum avoidance percentage. This is observable both in the avoidance percentage over time, presented in Figure 3.6 (b), and the relationship between the approach percentage and cumulative weight change in Figure 3.6 (c).

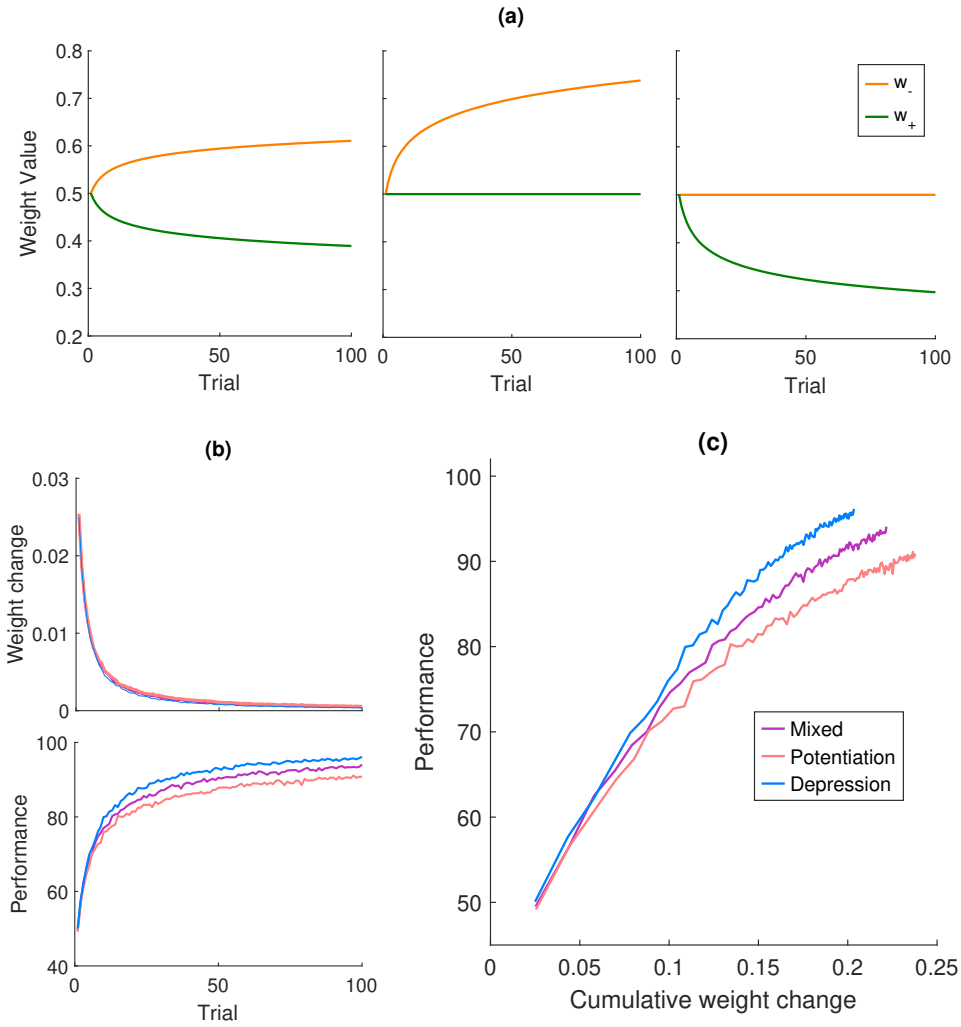

Figure 3.6: The Learning Efficiency for Aversive Learning. (a) Average weight value over trials for different strategies: "mixed" (left panel), "potentiation only" (middle panel), and "depression only" (right panel). (b) Average weight change and the percentage of taking the optimal action, denoted as Performance, across trials. (c) Relationship between the performance and cumulative weight change.

3.4 Contextual Bandit Model

In a state of hunger, fruit flies demonstrate an increased possibility to approach food-related odors and exhibit a reduced response to unpleasant odors (Inagaki et al. 2014, Cohn et al. 2015, Lin et al. 2019). This implies that hunger effectively modifies decision-making when the fruit flies are upon starvation. Also, in the intricate neural dynamics of the Drosophila brain, DAN neurons don't solely modulate based on reinforcement signals. They also factor in contextual information originating from both the internal physiological state and the external environment (Zolin et al. 2021, Lin 2023, Kim et al. 2007, Tsao et al. 2018, Zolin et al. 2021). Motivated

by biological observations, this section firstly evaluates the learning efficiency with different biases, then augments the basic model to incorporate energy (hunger) signal as the contextual dynamics that bias the decision-making, and investigates the impact of energy during this learning process.

3.4.1 Single-trial Learning with Bias

In this research, we propose that contextual biases significantly shape decision-making preferences before any learning occurs, specifically focusing on the approach and avoidance behaviors modulated by DANs and MBONs. A bias that favors approach behavior might emerge from either an enhancement of the synaptic weight that facilitates approach w^+ or a reduction in the synaptic weight that facilitates avoidance w^-. Conversely, a bias towards avoidance behavior could arise from an increase in w^- or a decrease in w^+.

To investigate how these initial biases affect the efficiency of learning, we utilize the evaluation framework outlined in Section 3.3.1. Assuming the correct action is to approach, and considering that the fly applies a predetermined amount of weight change in a single trial, the impact of pre-learning biases is considered by starting with varied initial weights. We introduce variability in the initial weights through biases b_+ and b_-, varying from -0.5 to 0.5, which adjust the starting values of w_0^- and w_0^+, respectively. In the absence of any bias, both w_0^- and w_0^+ are set to 0.5. Hence, the initial synaptic weights can be modified as $w_0^+ = 0.5 + b_+$ and $w_0^- = 0.5 + b_-$.

In this context, we categorize pre-learning biases into two distinct types. A "beneficial" bias refers to a predisposition that increases the probability of choosing the preferred action. Conversely, a "detrimental" bias denotes a predisposition that diminishes the likelihood of selecting the preferred action. Figure 3.7 presents the impact of different biases on the learning performance when a consistent weight change value of 0.2 is applied. The findings suggest that when the initial bias is beneficial, as depicted in the top three panels of Figure 3.7 (a) and (b), dedicating the entire weight change budget to reduce the synaptic strength associated with the incorrect action (avoidance) secures the most favorable learning results. This method of synaptic adjustment, which focuses on decreasing the synaptic weights linked to incorrect actions, proves to be effective across a wide range of biases, from

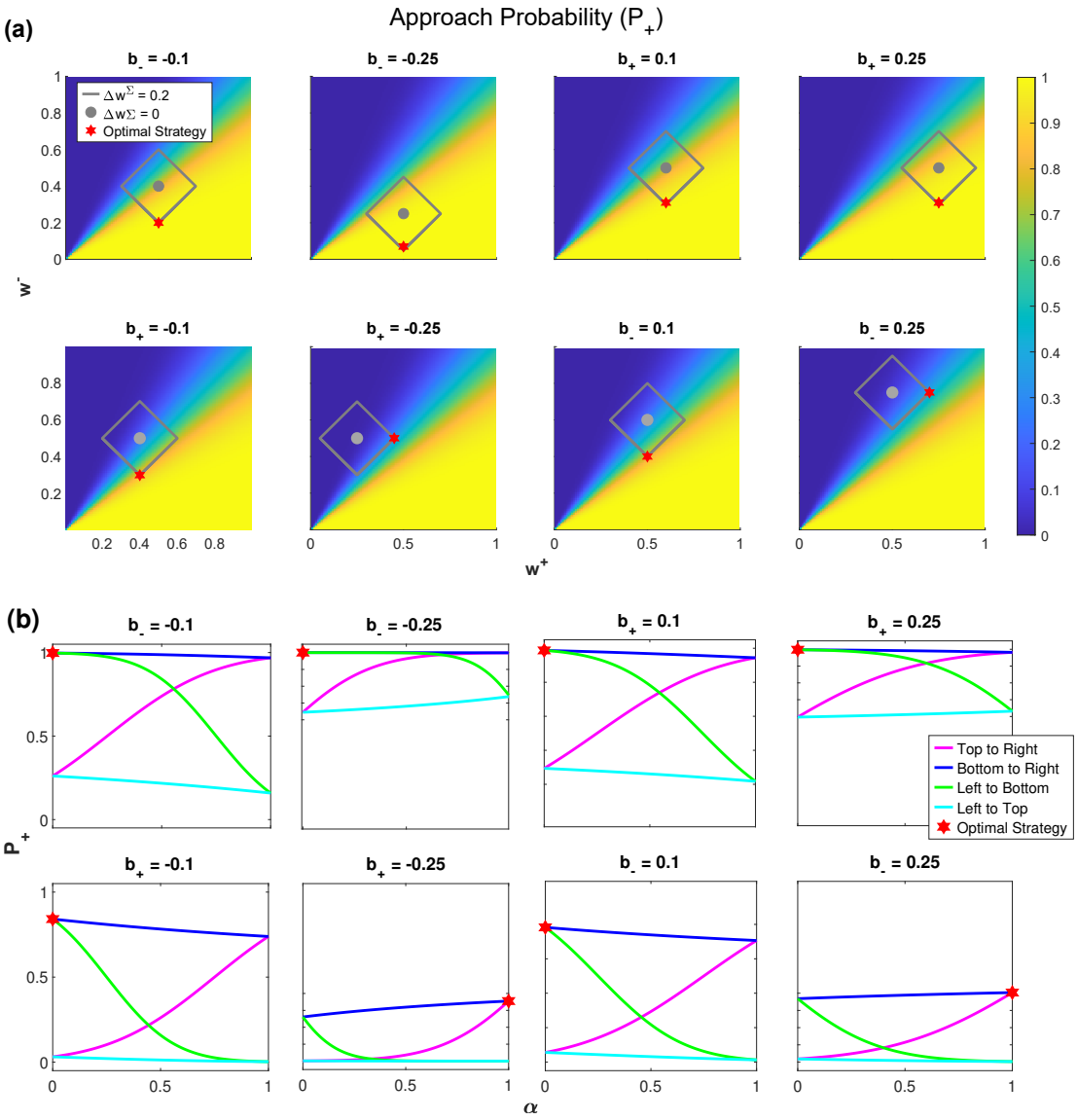

Figure 3.7: Approach probability P_+ and learning efficiency with fixed weight change, when there's a bias prior to the learning. (a) Approach probability P_+ for varying values of w^+ and w^-. The gray contour delineates the region where the weight change Δw^{Σ} remains constant at 0.2, with the initial weight value denoted by a gray dot. The optimal weight update strategy, yielding the highest approach probability, is indicated with a red star. The first and second row indicates the initial weight values leading to the beneficial and detrimental bias respectively. (b) Change of the approach probability along each edge of the corresponding diamond-shaped contour from (a), where the optimal weight update strategies are highlighted via the red stars.

minor to significant.

In scenarios where there is a detrimental bias, as shown in the bottom three panels of Figure 3.7 (a) and (b), the optimal strategy for employing synaptic weight change

varies with the degree of bias. For a small bias towards incorrect decisions, utilizing the weight change for the depression of synaptic weights associated with these incorrect actions remains the optimal strategy. However, as the increase of the bias towards incorrect decision-making, a shift in strategy becomes necessary. In these instances, potentiating the synaptic weights that lead to the correct action becomes more effective than simply depressing the weights associated with incorrect actions.

3.4.2 Incorporating Energy as Contextual Information

From the energy perspective, DANs can respond to the hunger signals, such as dNPF (Krashes et al. 2007, 2009), sNPF and 5HT (Lin et al. 2019), which bias the approach/avoid behavior. Specifically, we've considered the role of energy as significant contextual information, hypothesizing that DAN neurons, upon receiving energy signals, convey this as a contextual signal. This transmission subsequently biases the behavioral response towards an odor, as visualized in Figure 3.8.

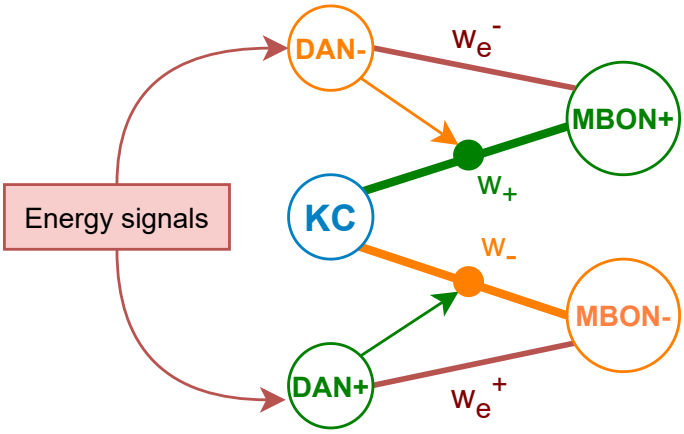

Figure 3.8: The olfactory learning network incorporating energy as contextual information. Energy-related signals (e.g., dNPF and 5HT) are transmitted to DANs (e.g., MP1 DANs), which then utilize these signals to modulate behavioral responses to odors. Notably, in this configuration, the DAN-MBON synaptic weights, w_e^+ and w_e^-, associated with behavioral biases due to the contextual energy signal, remain constant and are not subject to learning variations.

The computation of the action values in this revised model, denoted by $\mathbf{Q_e}$, not only factors in learned behaviors but also integrates the fly's energy signals. The action values are obtained through a multiplication of sensory inputs, $\mathbf{x} = [x_-, x_+]$, with the corresponding synaptic weights, $\mathbf{w} = [w^-, w^+]$. Furthermore, the hunger

signal, represented as $H = 1 - E$ (with E being the energy level ranging from 0 to 1), modulates an energy-centric term governed by $\mathbf{w_e} = [w_e^+, w_e^-]$, the weights connecting DANs and MBONs. The value of $\mathbf{Q_e}$ can be found by

$$\mathbf{Q_e} = \begin{bmatrix} w^-, & w^+ \end{bmatrix} \circ \begin{bmatrix} x_-, & x_+ \end{bmatrix} + H \begin{bmatrix} w_e^+, & w_e^- \end{bmatrix} \tag{3.8}$$

In this setup, the value of the hunger-dependent bias remains constant during the learning process, reflecting the physiological state of the organism without fluctuation during individual learning episodes.

Since hunger increases responses to food odors while reducing the response to unpleasant odors (Inagaki et al. 2014, Lin et al. 2019), this model posits that starvation induces a bias favoring approach behaviors. This is achieved by setting w_e^+ to zero, and assigning w_e^- a positive value. This configuration implies that as energy depletes, a higher hunger level is conveyed by the DAN-neuron, thereby increasing the g possibility of approach P_+. Figure 3.9 illustrates the probability of flies approaching an odor, given varying energy levels and w_e^- values. With diminishing energy levels, a greater proportion of flies approach the odor. Additionally, as w_e^- rises, the approach percentage becomes increasingly sensitive to changes in energy.

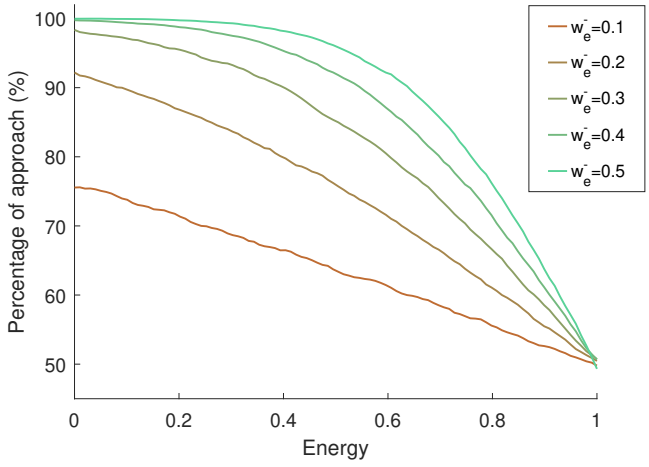

Figure 3.9: Percentage of odor approach without prior learning, as influenced by varying energy levels and w_e^- values, where w_e^+ is set to zero.

3.4.3 The Impact of Energy during the Learning Process

Utilizing the contextual bandit setup, the possibility of the flies approaching the odor and collecting the reinforcement signal increases as energy diminishes. Similar to the experiment demonstrated in Section 3.3.3, we simulate the learning processes of 5,000 flies over 100 trials, and set w_e^- to a moderate value of 0.3. The percentage of correctly choosing an action relative to cumulative weight changes is illustrated in Figure 3.10 (a) for appetitive learning and Figure 3.10 (c) for aversive learning. In appetitive learning, the initial proclivity to approach the odor is modulated by energy levels. The bias induced by energy signals enables flies to achieve higher approach percentages with minimized weight adjustments. An instance of cumulative weight change, when the approach rate attains 90%, is shown in Figure 3.10 (b). At reduced energy levels, flies can reach a 90% approach rate with fewer weight adjustments.

For aversive learning, weight change efficiency reduces with decreasing energy. As energy levels go down, the fly's inclination to approach increases, requesting more weight adjustments to counteract this bias. This trend is evident in Figure 3.10 (d), where weight changes increase as energy decreases to achieve a 70% avoidance rate.

When comparing the learning performance between appetitive and aversive learning, as illustrated in Figure 3.10 (a) and (c), our findings show that achieving a high percentage of correct actions is generally easier in appetitive learning than in aversive learning. This is because reinforcement signals are provided when flies approach the odor. In contrast, aversive learning requires a longer learning process to reach a better performance; as learning progresses, flies tend to avoid the odor, resulting in fewer aversive reinforcements. This trend is evident in the absence of contextual bias, as seen in Figure 3.10, where the energy is full ($E = 1$), indicating no initial bias prior to learning. In this case, the learning process for appetitive learning (shown in Figure 3.10 (a)) is faster than for aversive learning (shown in Figure 3.10 (c)). This trend persists even when energy introduces a bias before learning (i.e., when $E < 1$). Specifically, when the inherent bias favors approaching the odor, this inclination further slows down the aversive conditioning process.

Remarkably, the efficiency of the depression-only method surpasses both the "potentiation-only" and "mixed" strategies, independent of how energy states affect initial odor

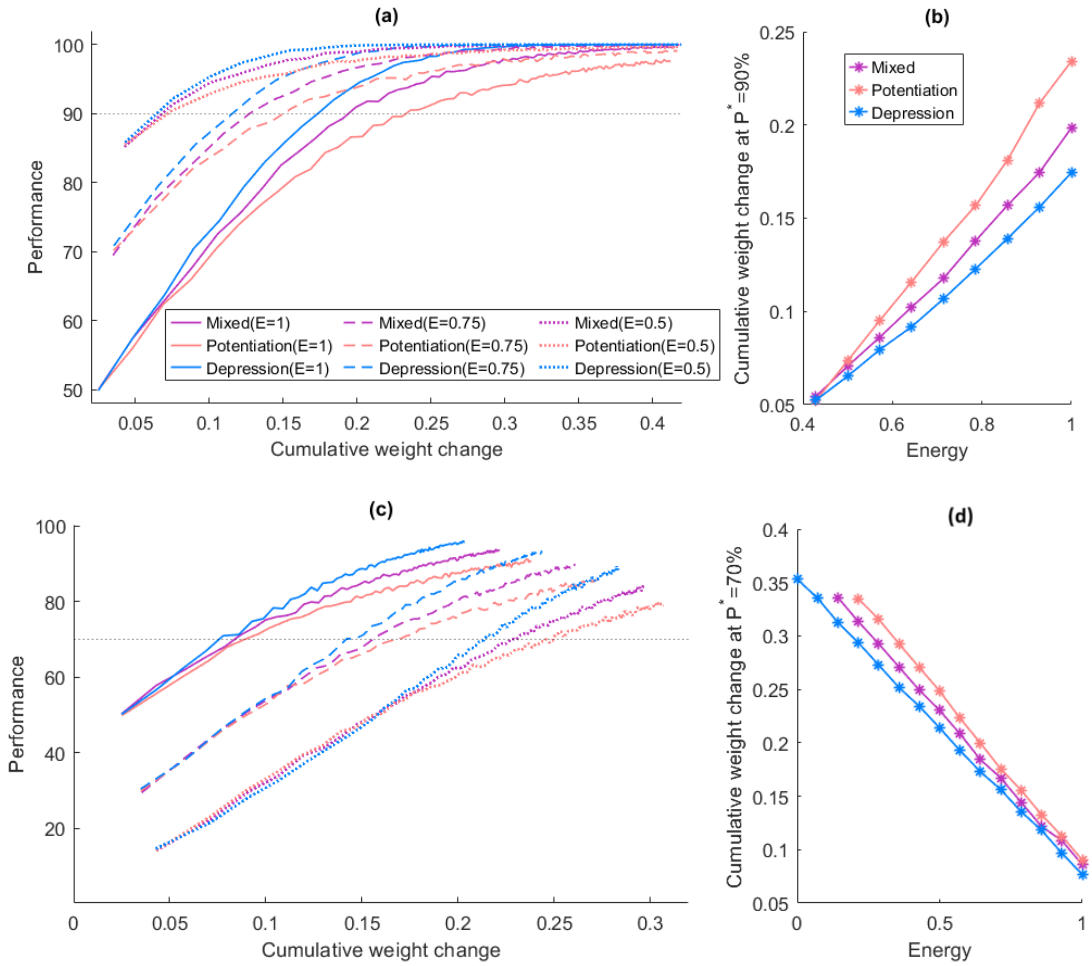

Figure 3.10: Impact of Contextual Energy Signal in Appetitive and Aversive Conditioning. (a) Relationship between the percentage of optimal action selection and cumulative weight change across various energy levels during appetitive conditioning. (b) Cumulative weight change required for the percentage of correct action selection P^* to reach 90% during appetitive conditioning. (c) Relationship between the percentage of optimal action selection and cumulative weight change across different energy levels during aversive conditioning. (d) Cumulative weight change required for the percentage of optimal action selection P^* to reach 70% during aversive conditioning.

preferences. This suggests that MB's weight update strategy maintains high efficiency when accounting for biases introduced by hunger signals as contextual information. During appetitive conditioning experiments, changes in energy levels introduce varying degrees of beneficial bias, the result here is consistent with findings from single-trial learning, where a bias towards the desired action always benefits the "depression-only" strategy.

In the context of aversive conditioning, biases induced by energy-related signals in-

crease the likelihood of selecting the undesired action. Here, the "depression-only" mechanism achieves the highest weight change efficiency, irrespective of the bias size. This contrasts with the analysis of single-trial learning, where the "potentiate-only" method is more effective with a significant bias towards making an incorrect choice. The potential reasons for this discrepancy are discussed in Section 3.5.3.

3.5 Discussion and Conclusion

3.5.1 Efficiency of "Depression-Only" Learning

The study highlights the superior learning efficiency of a "depression-only" mechanism, an approach that is biologically analogous to Drosophila's neural processes (Cohn et al. 2015, Hige et al. 2015, Owald et al. 2015). We introduce a term named "Weight Update Efficiency," designed to quantify the behavioral modification per unit of synaptic weight change. This efficiency is assessed within the frameworks of both a straightforward single-trial learning scenario and a more intricate multi-trial learning context. In each of these learning environments, a learning mechanism that exclusively involves synaptic depression upon the weight related to the incorrect actions demonstrates remarkable weight update efficiency. Consequently, we suggest the hypothesis that the fruit fly may employ this "depression-only" mechanism due to its superior efficiency.

This chapter evaluates the "Weight Update Efficiency" by assessing performance enhancements in comparison to the synaptic weight modifications. Such an evaluation potentially offers insights into the estimation of energy efficiency throughout learning, building on existing approaches that establish a correlation between synaptic plasticity's energy demands and the magnitude of weight change (Li & Van Rossum 2020) and as demonstrated in Chapter 2. Assuming equal energy costs for both potentiation and depression, the superior weight update efficiency observed in a "depression-only" learning paradigm suggests its high energy efficiency.

When accounting for the differential energy costs of synaptic potentiation and depression, depression may emerge as a more energetically economical form of plasticity. Potentiation, such as NMDA receptor-dependent LTP, requires the formation of

additional AMPA receptors into the postsynaptic membrane, which escalates synaptic energy consumption, afterward, the ATP demand at potentiated synapses also increases (Wieraszko 1982). Moreover, neuronal activity, such as the firing rate of cerebellar granule cells, is proportionally related to energy usage (Howarth et al. 2010). Synaptic depression contributes to an increase in silent neurons, thereby providing a potential for energy saving (Harris et al. 2012). To regulate the progressive synaptic strengthening and associated energy expenditures, the brain implements a negative feedback loop that suppresses the sustenance of potentiation under conditions of energy scarcity (Potter et al. 2010). Therefore, even when weight update efficiency is equivalent across different learning mechanisms, the "depression-only" approach may still manifest superior energy efficiency.

This research, in line with other studies on olfactory learning in Drosophila, predominantly considers neuronal activities based on binary valence: approach and avoid. Consequently, the bandit model derived from MB structure is also binary, featuring just two actions—one designated as "correct" and the other as "incorrect." Whereas in multi-armed bandit problems, it is often encountered in decision-making scenarios with multiple "correct" and "incorrect" choices. The two-armed model may not fully capture the character of scenarios with more options. Therefore, when extending the principles of "potentiating correct actions" and "depressing incorrect actions" to multi-armed contexts, the assessment of learning efficiency may need to account for the varied ratios of "correct" to "incorrect" options, which could influence the dynamics and outcomes of the learning process.

The finding that depressing the weight associated with incorrect actions results in the highest weight update efficiency is based on the assumption of multiplicative noise at the KC-MBON synapse. Specifically, this is expressed as $\sigma_Z^2 = [(w^-)^2 + (w^+)^2]\,\sigma^2$, as shown in Section 3.3, where $Z = w^- x^- - w^+ x^+$, w^+ and w^- represent the weights for approach and avoidance behaviors respectively. Future research could investigate whether this finding holds under different conditions, such as varying levels of noise or alternative synaptic models.

Apart from the notable weight change efficiency, previous studies indicate that a learning mechanism exclusively focused on synaptic depression in response to incorrect actions can outperform alternative approaches under certain conditions. For

instance, research from Abdelrahman (2023) suggests that within a decision-making framework with divisive normalization based on MB, a depression-only learning model excels over one based on potentiation.

3.5.2 Influence of Metabolic Energy on Behavior

When incorporating energy signals as contextual information, we adopted a contextual bandit framework and developed a model that increases the probability of flies approaching odors under conditions of starvation. This model is inspired by empirical evidence indicating that starvation enhances the likelihood of flies being attracted to food odors (Lin et al. 2019). This adaptive behavior likely acts as a survival strategy, optimizing food-seeking behaviors in starved flies by improving learning efficiency in response to hunger signals.

In this study, energy levels differentially modulate learning based on the nature of the reinforcement. In the context of appetitive learning, diminished energy levels—indicative of hunger or starvation—amplify learning efficiency. Conversely, during aversive learning, a heightened propensity to approach odors in energy-deprived states impedes effective learning. Such differential responses to energy conditions have been echoed in empirical studies. For example, starvation has been shown to impede LTM formation for aversive conditioning (Plaçais & Preat 2013), while facilitating LTM formation in appetitive scenarios (Krashes & Waddell 2008).

3.5.3 Decision-Making Bias from Contextual Information

When analyzing the pre-learning biases, we distinguish between "beneficial bias", which enhances the probability of correct actions, and "detrimental bias", which increases the likelihood of incorrect actions. The "depression-only" strategy is highly efficient under conditions of beneficial bias in both single-trial and multi-trial learning environments. This approach retains its superior efficiency in the face of detrimental biases within multi-trial settings; however, in single-trial learning scenarios with significant detrimental biases, potentiating synaptic weight for correct actions becomes the preferred method. This discrepancy might be due to multi-trial learning spreading total weight adjustments across several trials, resulting in small weight

changes per trial. Such gradual weight changes across trials might also cause variations in weight update efficiency patterns between single-trial and multi-trial learning scenarios.

In addition to energy, various contextual elements influence odor response in MB. External factors, such as environmental conditions like airflow (Zolin et al. 2021), and internal states, such as reproductive status (Lin 2023), mating (Boehm et al. 2022), and satiety (Kim et al. 2007, Tsao et al. 2018, Zolin et al. 2021), all bias behavioral outcomes. An investigation into how DANs interact with factors like mating status or the novelty of information could further enhance the contextual bandit framework.

Chapter 4

Energy-Adaptive Reinforcement Learning for Foraging

4.1 Synopsis

The exploration-exploitation trade-off, fundamental to neurobehavioral decision-making and reinforcement learning, is a subject of interest across both human and non-human organisms. It is also of interest in marketing and medical testing. This chapter focuses on the exploration-exploitation trade-off in the context of foraging. Existing algorithms, such as Upper Confidence Bound (UCB) based algorithms and Bayesian-based algorithms, strive to optimize regret under the assumption of infinite agent lifespan. They have been used in real-world foraging research, considering the costs of switching options and information gathering, and evaluating immediate and future rewards. However, these models often neglect survival in their foraging strategy considerations. Our study addresses this gap by incorporating the agent lifetime into a multi-armed bandit model for foraging behavior. The primary objective of the agent within this model is to optimize the mean lifetime, thereby providing a novel perspective on decision-making strategies in the context of foraging. We find that models promising minimal regret may reduce agents' lifetime due to extensive exploration. To resolve this issue, we propose energy-adaptive algorithms that not only extend agents' lifetime, but also maintain comparable regret to the baseline models. We hypothesize that these models possess the potential to elucidate the underlying mechanisms governing animal foraging.

4.2 Introduction

The exploration-exploitation trade-off plays a significant role in the domains of neurobehavioral decision-making and reinforcement learning, where an agent faces the choice of either exploiting the currently perceived optimal option or exploring new options in the quest for potentially higher rewards. A disproportionate focus on exploration could lead to wasting resources by committing to suboptimal decisions, whereas an over-emphasis on exploitation could potentially overlook superior strategies (Cohen et al. 2007). This fundamental trade-off has been extensively studied across various species, both human and non-human (Krebs et al. 1978, Daw et al. 2006, Pearson et al. 2014, Addicott et al. 2017). Unraveling the complexities of this trade-off is essential for both the improvement of theoretical models and the comprehension of decision-making systems in animal brains.

This dilemma of exploration versus exploitation is likewise crucial in the context of animal foraging. When searching for new food patches, foragers must navigate a trade-off between the discovery of nutrient-rich resources and the associated expenditures of time and energy (Krebs et al. 1978, McNamara & Houston 1985, Bell 2012, Katz & Naug 2015). This predicament is often referred to as the patch-leaving problem (Charnov 1976). Research on the patch-leaving problem has revealed not only consistent neural activities across various species (Pearson et al. 2014) but also comparable near-optimal behaviors describable by shared algorithms (Adams et al. 2012, Pearson et al. 2014). Such findings suggest the potential existence of a universal foraging algorithm applicable across different species.

Evidence suggests that the level of satiety significantly regulates the explore/exploit decision in foraging contexts. Empirical findings have demonstrated a correlation between foraging behavior and levels of hunger. For example, in Drosophila foraging, a period of starvation leads to an increase in local search duration, a behavior characterized by staying close to the food source (Bell et al. 1985). Similarly, starved honeybees reduce preference for novel rewards and uncertain rewards during foraging tasks (Katz & Naug 2015). Hence, it can be inferred that animals with lower energy reserves tend to emphasize exploitation during foraging.

Several algorithms for solving the Multi Armed Bandit (MAB) problem have been

proposed to balance exploration and exploitation. A notable example is the Upper Confidence Bound (UCB) algorithm. It bases its decisions on the empirical mean reward estimation coupled with a confidence radius. A classic representation is UCB1, which employs a straightforward confidence radius contingent on both the number of times an arm has been selected and the total trials conducted (Auer et al. 2002). The calculation of this confidence radius can be refined for various scenarios, such as KL-UCB, who leverage the Kullback-Leibler divergence to determine a narrower confidence interval (Cappé et al. 2013). From a Bayesian perspective, strategies like Thompson Sampling (TS) derive decisions from the posterior distribution of the arm expected reward (Chapelle & Li 2011). Another approach is the Gittins index, which formulates an optimal stopping boundary from the present posterior and the expected value of exploring alternative options (Gittins et al. 2011). These algorithms primarily target the optimization of regret, defined as the disparity between the chosen and optimal rewards. This optimization generally assumes an agent's infinite lifespan and, often, an infinite number of trials.

As the exploration-exploitation dilemma is central to both MAB problems and foraging, these models have found application in foraging tasks. For example, the application of the two-armed bandit model in foraging studies, such as those conducted with great tits (Krebs et al. 1978) and bumblebees (Keasar et al. 2002), demonstrates its utility in examining how these species make foraging decisions in environments characterized by uncertainty and variably rewarding feeding sites. Other studies have considered the cost of switching from one option to another (Agrawal et al. 1990) or the trade-off between immediate and future rewards while factoring in the cost of information gathering (Averbeck 2015). Morimoto (2019) utilized the UCB and TS algorithm to interpret data on foraging behavior gathered from fly larvae. However, the existing models generally presume the continuous survival of the foraging agent under various strategies, thereby overlooking scenarios where the animal might perish during the foraging endeavor. Given that foraging activities are integral to an animal's survival and that animals are constrained by finite energy and lifetime resources (Addicott et al. 2017), the factor of survival should not be decoupled from the pursuit of reward optimization.

Our study advances this perspective by modeling foraging behavior using a multi-armed bandit approach, incorporating energy cost, energy intake, and agent lifetime

estimation based on energy reservation during the learning process. We implement the lifetime evaluation into standard models and discover that models with favorable regret could have shorter lifetimes due to excessive exploration, suggesting that strategies maximizing reward collection might not always correspond to optimal foraging. To address this issue, we introduce an energy-regulated exploration/exploitation parameter into three approaches, the ϵ-greedy, the UCB algorithm, and the Bayesian-based TS algorithm. Our findings suggest that this energy-adaptive UCB and TS extends the agent's lifetime and enhances overall reward collection during the agent's lifetime, while maintaining a comparable regret level with the well-established baseline algorithms.

In contrast to Chapter 2 and Chapter 3, we take an algorithmic approach, and leave the question of how the algorithms would be implemented biologically for future investigations. As a result, importantly, we also don't include a metabolic cost for memory formation.

4.3 Multi-Armed Bandit problem with Lifetime Evaluation

We construct a framework based on the Stochastic MAB problem to simulate the learning processes in Drosophila. In the standard MAB framework, as depicted in Figure 4.1, an agent is presented with multiple arms, each offering rewards from distinct, unknown distributions. At each trial, the agent selects an arm based on its current reward estimate. Subsequently, upon receiving the actual reward, the agent updates its estimate for the chosen arm. The standard challenge is devising a strategy that guides the agent's arm selection, aiming to maximize the cumulative reward over a sequence of trials.

In traditional MAB settings with a finite horizon, learning ceases when the number of trials reaches a predetermined (large) maximum. Thus, if multiple agents are operating under the same configuration, each agent will engage in a fixed number of learning trials.

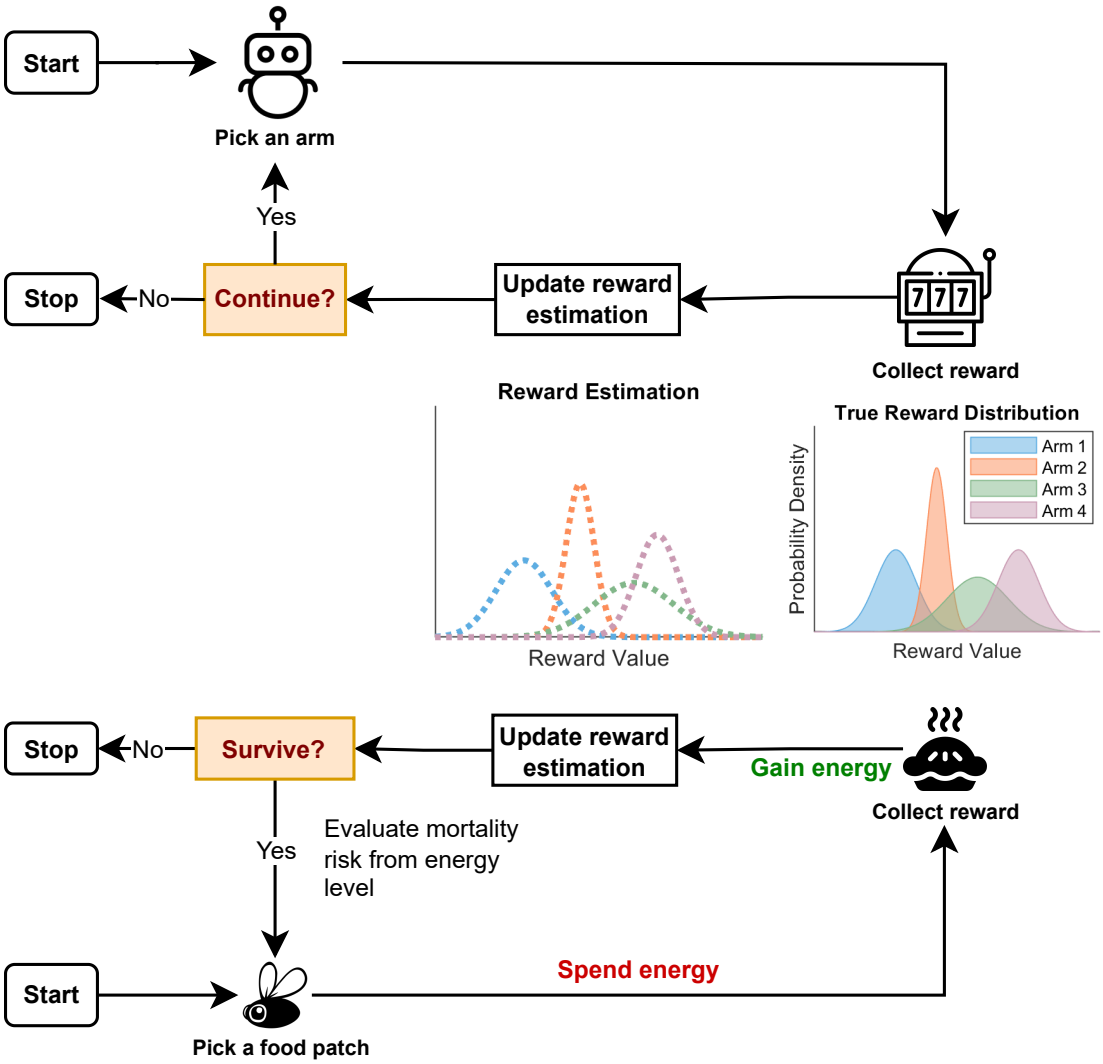

Figure 4.1: The learning process of MAB problem (upper panel), and its application to foraging behaviors in animal agents (lower panel).

While the conventional MAB framework operates under the assumption of an equal number of trials, such is not the case for foraging animals. These animals often encounter various risks, such as the threat of energy depletion during the learning process. This implies that each individual might not undergo the same number of learning trials. To address this discrepancy, we introduced an energy-centric learning protocol. Herein, each arm symbolizes a nutrient-rich patch, each distinguishable, for example, by a unique odor or color of a flower. During every trial, the agent expends energy to obtain the reward. This reward, in turn, is converted into an energy intake, as illustrated in Figure 4.1. Furthermore, we instituted an energy-based lifetime evaluation for each trial. Through this evaluation, we can ascertain the agent's lifespan when factoring in energy consumption.

4.3.1 Learning Task Incorporating Energy Considerations

As mentioned earlier, the model has K possible arms to choose from, where each arm (a) offers a reward drawn from an arm-specific distribution $\mathcal{D}_a$. Of course, a deterministic reward is included in such a model. We assume that the reward distribution is stationary. The energy reserve during the learning process is denoted by M, ranging from 0 to 1. Initially, at the outset of learning, each agent's energy level, M_{t_0}, is at its maximum value of 1. Each trial compels the agent to select a certain arm, which incurs an energy cost, M_f, associated with foraging and basal costs. Concurrently, an energy intake, $M_r(t)$, arises from the food reward. Agents cannot abstain from decision-making to avoid energy expenditure.

For simplicity, we equate the energy intake from the food reward with the reward value itself, $M_r(t) = r_t$. The agent's energy level following each trial is then updated by adding the reward value and subtracting the foraging cost,

$$M(t+1) = M(t) + r_t - M_f \tag{4.1}$$

Which M is rectified to lie between 0 and 1. The primary objective for each agent here is the prolongation of their survival duration. The value of the energy cost of foraging M_f is introduced in Section 4.5.

4.3.2 Energy Based Lifetime Prediction

In analogy to the survival analysis technique applied in Chapter 2 (refer to Section 2.4.2), we predict the lifetime of an agent through a similar method. We formulate this prediction under four assumptions:

- In each trial, the agent is subjected to a probability of 'death', denoted by the hazard function $h(t)$.
- The hazard, in the context of this learning protocol, arises from energy deficiency. The functional relationship between the hazard and energy can be expressed as:

$$h(t) = \exp\left(-c_m M(t)\right) \tag{4.2}$$

  $M(t)$ denotes the energy reservation at trial t, where $0 \leq M(t) \leq 1$. c_m is the

steepness of the hazard.

- When the energy level is maintained at its upper limit throughout the entire learning process, the agents exhibit a maximum lifetime l^*, and this value determines the steepness of the hazard c_m in Equation (4.2), via

$$c_m = \log(l^*) \tag{4.3}$$

The derivation of this equation is explained in Section 2.4.2, in this study, we let $l^* = 50$.

- In our simulations, T represents the upper limit of trials. We assume this limit is sufficiently large such that no agent can exceed it.

The agent's predicted lifetime, l, can be calculated from the hazard values extending from the initial to the final trial. In this context, "*lifetime*" denotes the predicted number of trials the agent remains active. To compute l, we begin by evaluating the survival function, $S(t)$, denoting the probability that an agent survives up to time t.

$$S(t) = \prod_{i=1}^{t} (1 - h(i))$$

Then we can find the probability that an agent has a lifetime of t trials. That is, the likelihood that the agent survives up to the $(t-1)^{th}$ trial and dies at t^{th} trial.

$$\begin{aligned} P(t) &= h(t)S(t-1) \\ &= h(t) \prod_{i=1}^{t-1} (1 - h(i)) \end{aligned} \tag{4.4}$$

Then we let T denote the upper limit of the trial number. Given that this upper limit is sufficiently large, we assume no agent's lifetime exceeds T, the predicted lifetime of the agents can be represented by:

$$\begin{aligned} l &= \sum_{t=1}^{T-1} tP(t) + T\left(1 - P(T)\right) \\ &= \sum_{t=1}^{T-1} th(t) \prod_{i=1}^{t-1} (1 - h(i)) - Th(T) \prod_{i=1}^{T-1} (1 - h(i)) + T \end{aligned} \tag{4.5}$$

For the most reliable estimation of the lifetime, T should ideally tend towards in-

finity. However, to strike a balance between robustness and practical experimental considerations, we set T as 500.

It's important to note, given this evaluation approach, that the deduced predicted lifetime is constrained by the predefined maximum threshold, denoted as $l^* = 50$. Yet, in practical scenarios and the original learning setup demonstrated in Figure 4.1, the actual lifetime of the animal/animal agent has the potential to surpass this maximum threshold l^*.

4.3.3 Evaluation Metrics

We compare model performance with two distinct evaluation metrics: the traditional matrices utilized for bandit models, and those incorporating considerations of energy and lifetime.

Initially, we employ a selection of commonly used conventional performance metrics for MAB problem, assuming an 'immortal' agent that disregards any potential mortality scenarios. Recognizing that such metrics may not accurately reflect biological realities for the animal agents, given that animals have a finite lifetime, during which they can learn and accumulate rewards, we employ the ensuing metrics to evaluate the performance of these animal agents within their living span, thereby enhancing the biological plausibility of our model performance.

In practice, we run all agents until time T (as if they were immortal), the actual lifetimes are then calculated at the end of the simulation.

Conventional Evaluation Metrics

Regret: One of the standard approaches to evaluating the performance of stochastic MAB problems is by checking the regret. Regret is defined as the difference between the cumulative reward of the optimal action with full prior knowledge of the rewards and the cumulative reward of the action taken by the algorithm. Denoting the mean

reward of the best arm μ^*, at trial T, the regret for the immortal agents is

$$R^c(T) = \mu^* T - \sum_{t=1}^{T} r_t \tag{4.6}$$

Exploration/Exploitation: To assess the models' proficiency in handling the exploration-exploitation trade-off, we monitor the Exploration/Exploitation behavior throughout the learning phase. In each trial, the algorithm generates a mean reward estimation, $\bar{\mu}_a$. If the selected arm, a, corresponds to the arm with the maximum $\bar{\mu}_a$, it indicates the model is in the exploitation phase. Conversely, if the chosen arm isn't the one with the highest $\bar{\mu}_a$, it suggests the model is in the exploration phase. Notably, as every model encounters an unseen arm during the initial trial, we universally classify the first encounter with an unseen arm as exploration.

Evaluation Metrics with the Consideration of Energy and Lifetime

Lifetime: The lifetime l represents an estimate of the total number of trials an agent is expected to participate in during the course of the experiment, as delineated by Equation (4.5).

Hazard Trajectory: This represents the hazard alterations through the learning phase, spanning from the initial trial to the maximum trial, T.

The simulation is outlined in Learning Protocol 1. In the scope of this research, we formulated an energy-adaptive algorithm, and its efficacy was assessed against three benchmark algorithms, detailed in Section 4.4.

Notably, the exploration-exploitation trade-off still exists in this protocol. An exploitation strategy prompts the agents to gravitate towards arms with the highest expected reward, though this potentially risks overlooking the most rewarding arm. In contrast, an exploration strategy empowers the agents to accumulate extensive knowledge of the reward distributions of each arm. However, this introduces the risk of energy exhaustion by accessing low-reward arms only.

Learning Protocol 1: Stochastic MAB Incorporates Energy Considerations

1: **Known parameters:** arm number K, maximum trials T, initial energy M_{t_0}, foraging cost for every trial M_f;
2: **Unknown parameters:** reward distribution $\mathcal{D}_a$ for each arm a.
3: **Initialize :** Energy level, $M \leftarrow M_{t_0}$.
4: Base on the **Algorithm**, selects arms, collects reward r_t, and updates the energy level $M \leftarrow M + r_t - M_f$ iteratively.
5: Find the **Conventional Evaluation Metrics**: $R^c(T)$.
6: Evaluate **Metrics with the Consideration of Energy and Lifetime**: l, h.

4.4 Model Design

In this research, we introduce three Energy Adaptive (EA) algorithms grounded in the ϵ-greedy, UCB, and TS algorithms. These EA models are benchmarked against established standards known for optimally managing the exploration-exploitation trade-off.

4.4.1 ϵ-greedy

The ϵ-greedy bandit algorithm is a simplistic yet effective approach to maintaining an equilibrium between exploration and exploitation. The mechanism predominantly selects the optimal arm with the highest expected average reward (with a probability of $1 - \epsilon$), and sporadically selects a random arm (with a probability of ϵ) (Sutton & Barto 2018). This strategy ensures that the algorithm capitalizes maximally on the existing knowledge, interspersed with random explorations to potentially discover superior options. For the baseline algorithms in this study, we set ϵ at 0.2. Note, that the continued exploration of this algorithm is also useful if the reward distribution is non-stationary. Which is for instance relevant if the reward gets exhausted (not modelled here).

4.4.2 Energy dependent ϵ-greedy

Given that animals with reduced energy reserves are less inclined to explore (Keasar et al. 2002), we conceptualized an energy-adaptive variant of the standard ϵ-Greedy algorithm. In this adaptation, the value of ϵ_{EA} is modulated based on the animal's energy reserves, see Equation (4.7). The details of the benchmark and EA model

based on ϵ-Greedy are shown in Algorithm 1.

$$\epsilon_{EA}(t) = \epsilon M(t) \tag{4.7}$$

We assume a simple linear dependence of the exploration rate on energy. In general, non-linear dependencies could perhaps perform better. The optimal strategy could even depend on past acquired rewards. Finding optimal energy-dependent variants has to be delegated to future work. We will make similar assumptions below.

Algorithm 1 **ϵ-Greedy** (ϵ-Greedy / EA-ϵ-Greedy)

Parameters: count of arm selections n_a, expected reward $\bar{\mu}_a$.
Initialize: For all arms, $n_a \leftarrow 0$ and $\bar{\mu}_a, \leftarrow 0$, $\epsilon \leftarrow 0.2$.
for each round $t = 1, 2, \ldots, T$ **do**
    Draw $u \sim U(0, 1)$

ϵ-Greedy	**EA-ϵ-Greedy**
$a_t = \begin{cases} \text{random arm} & \text{if } u < \epsilon \\ \underset{a}{\text{argmax}}\,(\bar{\mu}_a) & \text{otherwise} \end{cases}$	$a_t = \begin{cases} \text{random arm} & \text{if } u < \epsilon_{EA} \\ \underset{a}{\text{argmax}}\,(\bar{\mu}_a) & \text{otherwise} \end{cases}$
Collect reward r_t	Collect reward r_t
Update $(\bar{\mu}_a, n_a)$ $\leftarrow \left(\frac{n_a\bar{\mu}_a + r_t}{n_a+1}, n_a + 1\right)$	Update $(\bar{\mu}_a, n_a, M)$ $\leftarrow \left(\frac{n_a\bar{\mu}_a + r_t}{n_a+1}, n_a + 1, M + r_t - M_f\right)$

end for

4.4.3 UCB

The UCB algorithm is an optimism-oriented strategy, which balances the exploration-exploitation trade-off by having the agent operate under the presumption that the environment is as advantageous as it could feasibly be (Slivkins et al. 2019, Lattimore & Szepesvári 2020). At the start, it assumes that each arm gives an infinite reward. As exploration progresses, it curbs its enthusiasm and replaces it with the highest reward value still possible given the data. This type of model generally selects the arm with the highest UCB value, which incorporates both the average reward of each arm and its uncertainty (confident radius) (Slivkins et al. 2019). Because each arm is initialized with an infinite reward, the algorithm always starts by sampling all arms.

In the realm of predictive decision-making, algorithms based on UCB have demonstrated their applicability across a diverse array of fields, including the study of foraging behavior. Under quite general conditions it is optimal in the limit of infinitely many trials. That is, the regret approaches that of the optimal bandit. It was applied in the models for both human and animal foraging. For instance, in the context of an MAB information foraging task with human participants in a controlled laboratory environment, UCB-based models yielded a commendable fit to paired decision-making behaviors (Naito et al. 2022). Also, in modeling fly larvae distribution across foraging patches, UCB significantly outperformed random algorithms and exhibited greater accuracy than TS (Morimoto 2019).

It was applied in the models for both animal and human foraging. For instance, when fitting the data about the fruit fly larvae distribution in foraging patches over time, UCB performed significantly better than the random algorithm and it's more accurate than TS (Morimoto 2019). Also, when fitting data from the MAB-based information foraging task for humans in controlled laboratory environments, UCB-based models provided a reasonable overall fit to the participants' choices, specifically when participants deciding between pairs (Naito et al. 2022).

In particular, the UCB1 algorithm proposed by Auer et al. (2002) is frequently applied as a baseline model in the context of MAB problems, owing to its theoretical assurances and robust performance in practice. Here, we also employ UCB1 as one of our benchmark algorithms, where its details are demonstrated in Algorithm 2. In UCB1 model, the decision-making among the K available arms in this model is guided by an upper confidence bound, symbolized as UCB^1, refer to Equation (4.8). Each arm a possesses a UCB^1_a value, which is composed of an estimated mean reward $\bar{\mu}_a$ and its associated confidence radius c_a.

$$UCB^1_a(t) = \bar{\mu}_a(t) + c_a(t), \quad \forall a \in [K] \tag{4.8}$$

Where t denotes the trial number, the estimated mean reward value for arm a is updated as

$$\bar{\mu}_a(t+1) = \frac{n_a(t)\bar{\mu}_a(t) + r_t}{n_a(t)+1}, \quad \forall a \in [K], \tag{4.9}$$

Where $n_a(t)$ denotes the cumulative count of selections made for arm a up to the

t^{th} trial. And the confidence radius $c_a(t)$ can be found by

$$c_a(t) = \sqrt{\frac{2\ln(t)}{n_a(t)}}, \quad \forall a \in [K] \tag{4.10}$$

This confidence radius is derived from **Hoeffding's Inequality** (Hoeffding 1994): Given n independent random variables *, $X_1, X_2, \ldots, X_n$, such that $a' \leq X_i \leq b$ for all i,

$$P\left(\left|\sum_i X_i - nE[X_i]\right| \geq nc\right) \leq 2\exp\left(-\frac{2n^2c^2}{\sum_i (b-a')^2}\right) \tag{4.11}$$

In this context, $E[X_i]$ represents the expected value of X_i, and c stands for a positive scalar, serving as the confidence radius. This inequality expresses that with increasing observations, the deviation from the mean, as expressed by the left-hand side, becomes less and less likely.

Considering a scenario with K arms. The true and estimated expected rewards for arm a are represented by μ_a and $\bar{\mu}_a$, respectively. The confidence radius corresponding to arm a is denoted as c_a, and n_a refers to the number of pulls for the same arm. When the reward is bounded within the interval [0, 1]. Then we have $a' = 0$ and $b = 1$. Incorporating these considerations into Equation (4.11) yields:

$$P(|\bar{\mu}_a - \mu_a| \geq c_a) \leq 2\exp(-2n_a c_a^2), \quad \forall a \in [K]$$

The objective is to determine an appropriate value of c_a such that the probability $\bar{\mu}_a - \mu_a \geq c_a$ is sufficiently small. Following the UCB1 model proposed by Auer et al. (2002), we set this probability to $2/t^4$, where t represents the total number of pulls across all arms, corresponding to the trial number in our learning protocol. Then we have

$$P(|\bar{\mu}_a - \mu_a| \geq c_a) \leq \frac{2}{t^4}, \quad \forall a \in [K]$$

And

$$2\exp(-2n_a c_a^2) = \frac{2}{t^4}, \quad \forall a \in [K]$$

Solving for c_a, we obtain Equation (4.10).

*From a sub-Gaussian distribution, i.e. not heavy-tailed.

4.4.4 Energy dependent UCB

To address the exploration-exploitation dilemma characteristic of foraging tasks, specifically within an energy-regulated context, we propose a variant of the UCB1—termed as the Energy-Adaptive UCB (EA-UCB). The upper confident bound in this algorithm UCB^{EA} is scaled by the energy M introduced in Section 4.3.1, again assuming simple linear scaling

$$UCB_a^{EA}(t) = \bar{\mu}_a(t) + M(t)c_a(t), \quad \forall a \in [K] \tag{4.12}$$

Notably, at maximum energy (M=1), it equals UCB1.

In every trial, the arm with the highest UCB^{EA} value is chosen, as detailed in Equation (4.12). The EA-UCB algorithm employs the energy level M_t to strike a balance between exploitation and exploration. Hence, in this model, there are three variables to be updated during the learning process, they are:

- $\bar{\mu}_a$: The estimated mean reward.
- n_a: The number pulls for arm a.
- M: The energy level.

The update for the chosen arm a is:

$$(\bar{\mu}_a, n_a, M) \leftarrow \left(\frac{n_a\bar{\mu}_a + r_t}{n_a + 1}, n_a + 1, M + r_t - M_f\right) \tag{4.13}$$

When the energy level is at its peak, the model leans towards exploring less frequented arms, reflecting a form of 'optimistic' behavior (Slivkins et al. 2019, Lattimore & Szepesvári 2020). Conversely, as the energy depletes, the influence of the confidence radius diminishes. At zero energy, the model disregards the confidence radius entirely and defaults to a 'greedy' strategy focused exclusively on exploitation. The Algorithm of EA-UCB model is in Algorithm 2. In this algorithm, the initial value of n_a is set to a modest offset η, further elaborated in Section 4.4.6.

4.4.5 Thompson Sampling

The Bayesian methods were proved to be an effective model for the estimation of animal decision-making across species (Arganda et al. 2012, Morimoto 2019), and

Algorithm 2 UCB (UCB1 / EA-UCB)

Parameters: count of arm selections n_a, expected reward $\bar{\mu}_a$, confident radius c_a.
Initialize: For all arms, $\bar{\mu}_a \leftarrow 0$.

UCB1

$n_a \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  For all arms, find
  $UCB_a^1 = \bar{\mu}_a + c_a$,
  $c_a = \begin{cases} \sqrt{\frac{2\ln(t)}{n_a}} & , \text{if } n_a > 0 \\ \infty & , \text{if } n_a = 0 \end{cases}$
  Find $a_t = \underset{a}{\operatorname{argmax}} \left(UCB_a^1\right)$
  Collect reward r_t
  Update $(\bar{\mu}_a, n_a, c_a)$
  $\leftarrow \left(\frac{n_a\bar{\mu}_a + r_t}{n_a+1}, n_a + 1\right)$
end for

EA-UCB

$n_a \leftarrow \eta$
for $t = 1, \ldots, T$ **do**
  For all arms, find
  $UCB_a^{EA} = \bar{\mu}_a + Mc_a$,
  where $c_a = \sqrt{\frac{2\ln(t)}{n_a}}$
  Find $a_t = \underset{a}{\operatorname{argmax}} \left(UCB_a^{EA}\right)$
  Collect reward r_t
  Update $(\bar{\mu}_a, n_a, M, c_a) \leftarrow$
  $\left(\frac{n_a\bar{\mu}_a + r_t}{n_a+1}, n_a + 1, M + r_t - M_f\right)$
end for

foraging (McNamara et al. 2006, Morimoto 2019). Thompson Sampling (TS), is a common Bayesian approach for solving MAB problems, introduced by Thompson (1933). It has demonstrated superior empirical performance and has found widespread use across various domains such as online advertising and recommendation systems (Chapelle & Li 2011, Agarwal et al. 2014, Kawale et al. 2015). A comprehensive introduction to TS can be found in Russo et al. (2018). TS employs a Bayesian approach to address the exploration-exploitation dilemma by maintaining a posterior distribution for each arm's reward probabilities. The selection of an arm is governed by sampling from each arm's distribution of the mean, with the arm yielding the highest sample being chosen with greater likelihood. By favoring arms with higher expected rewards and greater uncertainty, this method naturally balances exploration and exploitation. As more data is accrued, the associated uncertainty diminishes, yielding distributions that more closely mirror the true reward probabilities.

To upkeep the reward distribution for each arm, TS relies on Bayesian updating. The process commences with a prior distribution, which is then updated to a posterior distribution every time a reward is received. TS incorporates the use of conjugate priors from Bayesian statistics to streamline the process of updating beliefs based on new data (Raiffa et al. 1961, Lattimore & Szepesvári 2020). Given a likelihood

function $p(X|\theta)$ and a prior $p(\theta)$, the posterior distribution $p(\theta|X)$ is computed via Bayes' theorem

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

When the prior is selected to be a conjugate prior for the likelihood function, it can be simplified as

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

The posterior $p(\theta|X)$ belongs to the same family (or parametric form) as the prior $p(\theta)$. For each specific conjugate prior, a distinct set of posterior hyperparameters can be updated. Given a likelihood function, a corresponding conjugate prior can be found based on different assumptions of the likelihood.

In the experiments conducted in this study, the reward distribution for the arms was normal (see details in Section 4.5), implying that the likelihood values were drawn from a normal distribution with unknown mean and precision (the reciprocal of the variance). Assuming that the rewards observed in different trials are exchangeable, that is, the random observation has exchangeability, we can use a Normal-Gamma distribution as the conjugate prior, the details are explained in Murphy (2007). The Normal-Gamma distribution is a bivariate distribution generating two variables: the mean μ, which is equivalent to the estimation of the mean reward μ_a^{TS} in our baseline TS model, and its precision τ. In its formulation, τ is modeled by a Gamma distribution, while μ, conditioned on τ, follows a Normal distribution.

The joint distribution of μ and τ, is expressed by the product of the conditional Normal distribution of μ and the Gamma distribution of τ:

$$f_{NG}(\mu, \tau|\mu_0, \lambda, \alpha, \beta) = f_N(\mu|\mu_0, \frac{1}{\lambda\tau}) f_\Gamma(\tau|\alpha, \beta) \tag{4.14}$$

Here, parameters include the prior mean of the normal component, denoted as μ_0, and the scaling factor of precision, λ. The shape of the precision τ is influenced by α and β. In the baseline TS model, λ is equivalent to the number of pulls for each arm n_a.

This prior has four variables for each arm a, including:

- $\bar{\mu}_a$: The estimated mean reward.

- n_a: The number of prior observations, that is, the number of times that the arm a has been selected.
- α_a: The shape parameter of the Gamma distribution.
- β_a: The scale parameter of the Gamma distribution.

In each trial, the algorithm starts by drawing an estimated reward expectation μ_a^{TS} for each arm from the Normal-Gamma prior. The arm yielding the highest sampled reward μ_a^{TS} is selected. Upon playing an arm a and receiving a corresponding reward r_t, we update the prior hyperparameters in accordance with the Bayesian update rules of the normal distribution

$$(\bar{\mu}_a, n_a, \alpha_a, \beta_a) \leftarrow \left(\frac{n_a\bar{\mu}_a + r_t}{n_a + 1}, n_a + 1, \alpha_a + \frac{1}{2}, \beta_a + \frac{(r_t^2 - \bar{\mu}_a^2)}{2(n_a + 1)}\right) \tag{4.15}$$

Utilizing Bayes' theorem, this likelihood couples with the prior to compute a posterior distribution, which effectively updates the parameter estimates. At the end of the trial, the posterior from the current trial transitions to the prior for the next. As further rewards r_t are accrued, the ongoing update process iteratively refines the hyperparameters, thereby enhancing the model's decision-making capability. The specifics of this learning process are outlined in Algorithm 3.

By definition, drawing samples from a Normal-Gamma distribution can be split into two steps:

1. Get the precision parameter τ from a Gamma distribution with shape α and rate β: $\tau \sim \Gamma(\alpha, \beta)$.
2. Find the mean parameter μ conditioned on the sampled τ from a Normal distribution: $\mu|\tau \sim \mathcal{N}(\mu_0, \frac{1}{\lambda\tau})$.

Incorporating the hierarchical sampling framework into the TS model with a Normal-gamma prior, the exploration intent is characterized by the term $\frac{1}{\lambda\tau}$, representing the variance of the estimated reward. To modulate the exploration intent using energy, we scale this variance by energy, yielding an energy-adaptive reward estimation μ_a^{EATS}.

$$\mu_a^{EATS}|\tau \sim \mathcal{N}(\mu_0, \frac{M}{n_a\tau}) \tag{4.16}$$

Details of the Energy Adaptive Thompson Sampling (EA-TS) can be found in Algorithm 3. Within this framework, the initial value of n_a is assigned a modest offset

η, as expounded in Section 4.4.6.

Algorithm 3 TS (TS / EA-TS)

Parameters: count of arm selections n_a, expected reward $\bar{\mu}_a$, the shape α_a and the scale β_a of the Gamma distribution.
Initialize For all arms, $\bar{\mu}_a \leftarrow 0$, $\alpha_a \leftarrow 1$, $\beta_a \leftarrow 1$.

TS

$n_a \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  For all arms, find $\mu_a^{TS} \sim \text{NG}(\bar{\mu}_a, n_a, \alpha_a, \beta_a)$
  Find $a = \underset{a}{\text{argmax}} \left(\mu_a^{TS}\right)$
  Collect reward r_t
  Update $(\bar{\mu}_a, n_a, \alpha_a, \beta_a) \leftarrow (\frac{n_a \bar{\mu}_a + r_t}{n_a+1}, n_a + 1, \alpha_a + \frac{1}{2}, \beta_a + \frac{(r_t^2 - \bar{\mu}_a^2}{2(n_a+1)})$
end for

EA-TS

$n_a \leftarrow \eta$
for $t = 1, \ldots, T$ **do**
  For all arms, find: $\tau \sim \Gamma(\alpha, \beta)$ $\mu_a^{EATS} | \tau \sim \text{N}(\mu_0, \frac{M}{n_a \tau})$
  Find $a = \underset{a}{\text{argmax}} \left(\mu_a^{EATS}\right)$
  Collect reward r_t
  $(\bar{\mu}_a, n_a, \alpha_a, \beta_a, M) \leftarrow (\frac{n_a \bar{\mu}_a + r_t}{n_a+1}, n_a + 1, \alpha_a + \frac{1}{2}, \beta_a + \frac{r_t^2 - \bar{\mu}_a^2}{2(n_a+1)}, M + r_t - M_f)$
end for

4.4.6 Novel Arm Initialization

During model initialization, a prevalent approach for both UCB and TS algorithms is to sample each arm once, see examples in Auer et al. (2002), Slivkins et al. (2019), Lattimore & Szepesvári (2020). Indeed, our baseline UCB1 algorithm mandates a single visit to each arm at the outset. When an arm remains unvisited, the number of pulls is denoted by $n_a = 0$, so that the confidence radius c_a becomes infinite (see Algorithm 2), which directs the model's attention to unexplored arms. Likewise, for the baseline TS model, the reward estimation variance, $\frac{1}{n_a \tau}$, for an unseen arm also tends towards infinity, resulting in a heightened preference for exploration.

However, incorporating energy costs introduces potential complications. An agent might face a scenario with an overwhelming number of arm choices. Consequently, a model may persistently explore unsampled arms even when energy reserves are dwindling. Such an approach could be detrimental, since prioritizing unexplored arms may amplify risks, drain energy more rapidly, and hasten the agent's death. From the biological perspective, when Drosophila encounters an unseen odor, the DANs in the $\alpha'3$ MB compartment, associated with behavioral responses to novelty,

induce an "alert" response that interrupts grooming behavior (Hattori et al. 2017).

Since there is no empirical evidence suggesting that the fruit fly possesses an exceptionally strong inclination to explore novel odors, in our EA-UCB and EA-TS algorithms, we introduce an offset, denoted by η, to the n_a value during its initialization. This modification ensures the finiteness of both the confidence radius c_a in UCB and the variance $\frac{1}{n_a \tau}$ in TS, even for arms that haven't been sampled. In our simulations, we opted for $\eta = 1$. Implementing this adjustment has proven effective in reducing excessive early exploration in UCB1 and baseline TS, consequently extending their lifetimes. However, this extension in lifetime isn't as pronounced as observed in their EA counterparts. A comprehensive analysis of the impact of novel arm initialization can be found in Section 4.5.3.

4.5 Model Evaluation

We conducted an evaluation of the conventional regret, the explore/exploit behavior, and the lifespan-associated metrics for the model under two distinct testing scenarios: first, an environment featuring a single high-reward arm among low-reward arms, and second, a setting with multiple arms each carrying varying mean rewards.

As elucidated in Section 4.3.1, both experiments have an energy expenditure for foraging M_f, we set it as 1/10th of the maximum energy, hence we have $M_f = 0.1$. The reward of each arm a is dictated by a normal distribution characterized by a mean, μ_a, and a standard deviation, σ_a. Notably, the mean reward of the optimal arm is twice the energy expenditure for foraging, $\mu_a^* = 0.2$, for all the arms, the standard deviation is fixed at 1/10th of the optimal mean reward, then we have $\sigma_a = 0.02$. When evaluating the performance of the immortal agents, we set the maxima trial T at 500.

Across both Experiment 1 and 2, we assessed the performance of 1000 agents. In presenting mean values associated with model performance, we also included the standard error for clarity.

4.5.1 Experiment 1: Single High-Reward Environment Experiment

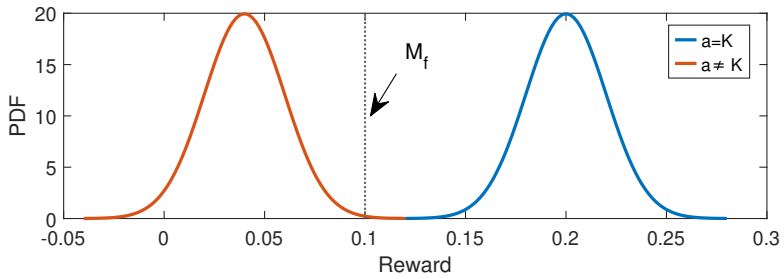

Figure 4.2: The reward distribution of Experiment 1.

In the first experiment, there are two arm categories: a high-reward arm and low-reward arms. There's only one high-reward arm with an optimal mean $\mu_a^* = 0.2$, which is twice as much as the foraging cost, all others remaining low-reward arms feature a mean value of $\mu_a^*/5 = 0.04$. The total amount of arms is denoted by K, we arbitrarily designate the last arm ($a = K$) as the high-reward arm. The probability density function (PDF) of this reward distribution is illustrated in Figure 4.2.

As the arm count K is varied, Figure 4.3 depicts both the predicted lifetime and regret at the final trial (with $T = 500$) applying distinct algorithms. The mean predicted lifetime decreases for all algorithms as the number of arms increases, as the rewarded arm becomes harder to find.

Evaluating the mean value of the predicted lifetime reveals that both the EA-UCB and EA-TS algorithms tend to exhibit enhanced survival durations relative to their baseline models. Conversely, the ϵ-greedy approach manifests negligible augmentation in lifetime.

Regarding the regret assessed at the final trial, the EA-UCB's regret closely aligns with its baseline counterparts for an arm count of less than 12. Similarly, the regret of EA-TS approximates that of the baseline model when the arm count is fewer than 6. Beyond this, the regret exhibited by both EA-UCB and EA-TS begins to surpass that of the baseline, with this discrepancy subtly intensifying with the increment of the arm count. This implies that the energy-adaptive approach incorporated within the UCB and TS algorithms can bolster the lifespan of the animal agent without significant regret sacrifices. In contrast, the EA-ϵ-greedy algorithm experiences a

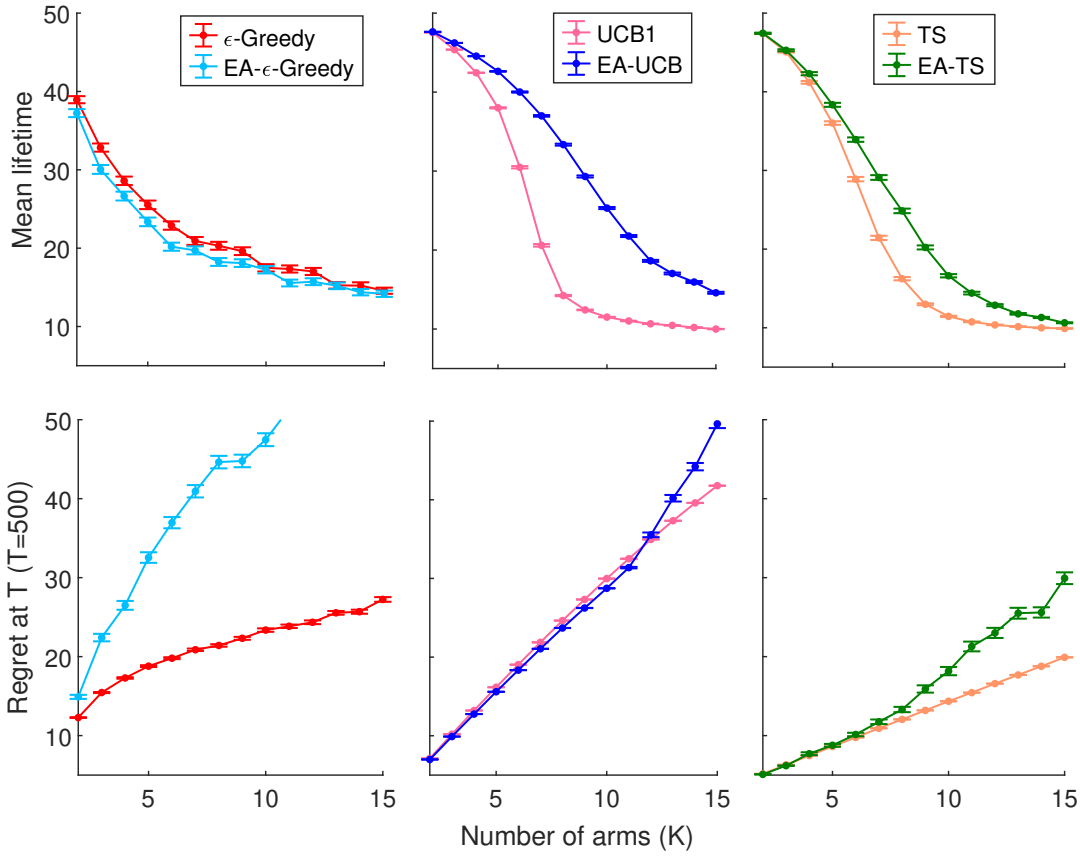

Figure 4.3: The predicted lifetime and the regret value at the final trial ($T = 500$) for testing Experiment 1.

pronounced regret relative to its baseline model, suggesting that for ϵ-greedy, curtailing exploration during periods of diminished energy neither improves lifetime nor attenuates regret.

With an increasing number of arms, discovering the arm with the optimal reward becomes more challenging due to the heightened presence of lower-reward arms serving as distractions. Across arm numbers, the EA-UCB algorithm consistently shows superior performance compared to other models. EA-TS emerges as the second-most effective model for a majority of the scenarios. Interestingly, the baseline UCB and TS models exhibit an extended lifetime when the total number of arms is limited. However, as the amount of low-reward arms grows, the performance of the ϵ-greedy algorithms begins to yield longer lifetimes.

To deepen our understanding of the learning process, we performed a pairwise com-

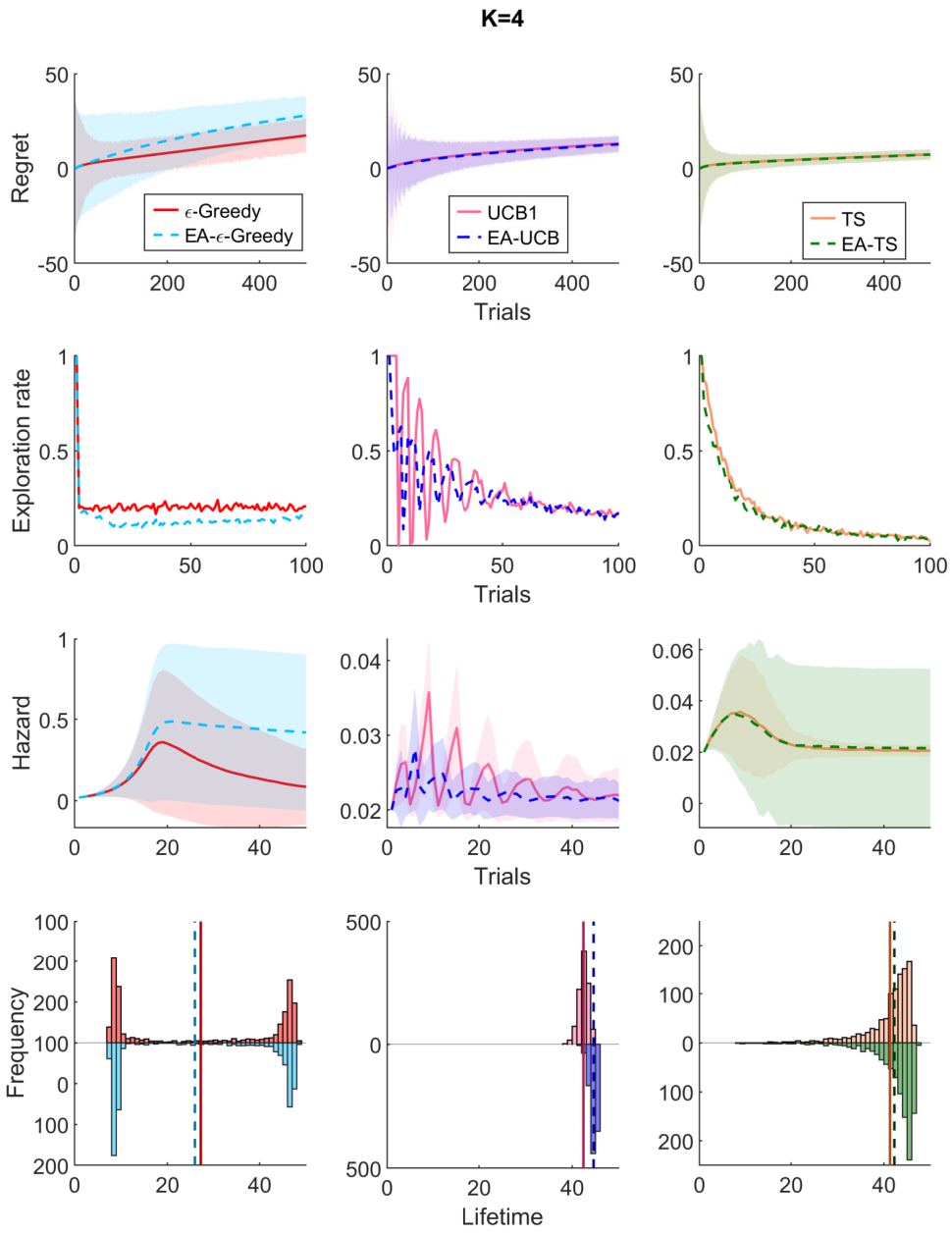

Figure 4.4: Pairwise comparison of Energy-Adaptive algorithms and their baselines in testing Experiment 1 with a setting of 4 arms. First row: Regret evolution over time. Second row: Exploration rate dynamics. Third row: Hazard trajectory over time. Fourth row: distribution of the predicted lifetime, with the mean value represented by a vertical line.

parison between the baseline algorithm and its energy-adaptive counterparts, holding the arm count constant. Our evaluation encompassed the regret over time, and the exploration rate—quantified as the proportion of agents choosing exploration relative to the total agent count across time, the hazard trajectory, and the distribution of lifetimes. As emphasized in Figure 4.3, the UCB and TS models excel over the ϵ-greedy algorithms when the arm count is low. As the number of arms rises,

the ϵ-greedy approaches surpass UCB1 and the baseline TS. To elucidate this trend, we delve into two specific scenarios: $K = 4$, showcased in Figure 4.4, and $K = 12$, illustrated in Figure 4.5.

For scenarios with low an arm number, the regret trajectories of EA-UCB and EA-TS closely align with their baseline algorithms. In contrast, the regret associated with EA-ϵ-greedy exceeds its baseline performance, where the disparity in regret widens as trials accumulate. Among the algorithms in this context, the TS variant exhibits the least regret. Additionally, UCB algorithms outpace the performance of ϵ-greedy algorithms in terms of regret.

Upon examining the exploration rate, the UCB1 algorithm exhibits pronounced oscillations in exploration across trials, indicative of periodic, intensive exploratory behaviors by the agent. As depicted in the central column plots of Figure 4.4, there's a discernible correlation for UCB1: the hazard surges when the exploration rate peaks. Notably, its energy-adaptive variant moderates these fluctuations and stabilizes the exploration rate. This stability translates to a diminished hazard, ultimately prolonging agent's lifetime.

The performance of the TS algorithms is captured in the rightmost plots of Figure 4.4. The energy-adaptive variant demonstrates a modest reduction in exploration during the initial trials, resulting in heightened variability in hazard during the early learning phase. Observing the lifetime distribution, the energy-adaptive approach enables a greater number of agents to have an extended lifespan. However, it also leads to premature death for some agents, rendering no appreciable effect on the predicted lifetime for this particular scenario.

For the ϵ-greedy algorithms, see the leftmost plots of Figure 4.4. The predicted lifespan exhibits a bimodal distribution, with significant concentrations at both a low and high lifetime, contrasting with the unimodal distribution observed in other strategies. The energy-adaptive ϵ-greedy algorithm marginally decreases the exploration rate over the entirety of the learning phase. Simultaneously, it considerably amplifies both the mean and variance of the hazard, resulting in a diminished predicted lifetime.

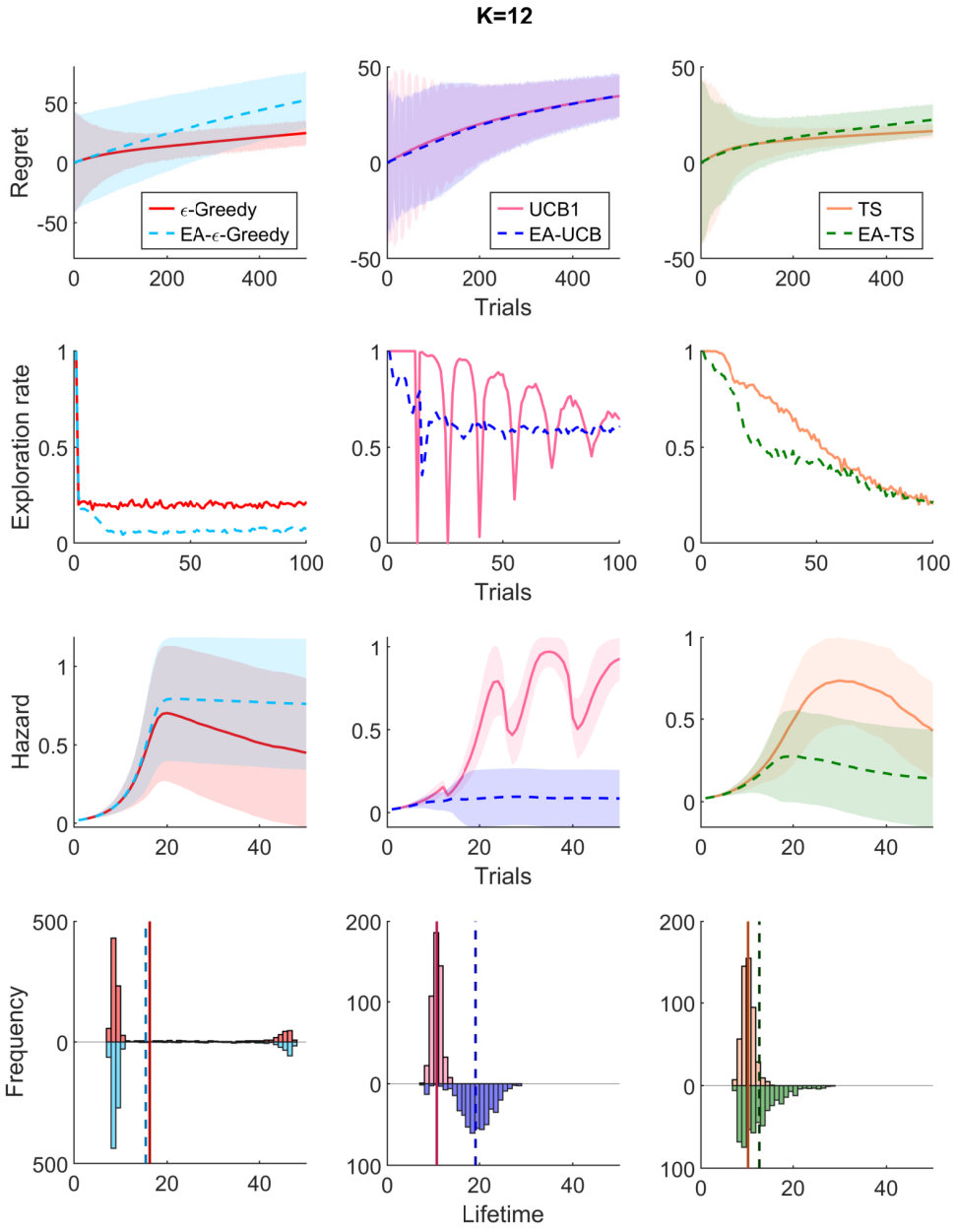

Figure 4.5: Pairwise comparison of Energy-Adaptive algorithms and their baselines in testing Experiment 1 with a setting of 12 arms. First row: Regret evolution over time. Second row: Exploration rate dynamics. Third row: Hazard trajectory over time. Fourth row: Lifetime distribution, with the mean lifetime represented by a vertical line.

Amidst the increased environmental complexity arising from a greater number of low-reward arms, the ϵ-greedy algorithms manifest patterns in regret, exploration rate, and hazard similar to those observed with fewer arm numbers, as depicted in the left column of Figure 4.5. Notably, as can be seen from the predicted lifetime time distribution of the ϵ-greedy approaches, although a majority of agents meet

their demise within the first 20 trials, a handful approach the maximum predicted lifetime prediction (as detailed in Section 4.3.2) of 50 trials. These agents' extended survival likely results from their early identification and consistent exploitation of the high-reward arm. Their presence contributes to a higher average lifetime, clarifying why the ϵ-greedy algorithms' mean lifetime with a high arm number is prominently elevated in Figure 4.3.

In scenarios with higher arm counts, the EA-TS algorithm demonstrates a more significant extension in lifetime compared to its baseline, as evidenced by the plots in Figure 4.3's right column. This improvement is primarily due to its energy-adaptive strategy, which reduces exploration, particularly in the early stages, which effectively mitigates the sharp rise in hazards.

For the UCB algorithms, Figure 4.3's middle column illustrates that UCB1 experiences heightened exploration fluctuations in the scenario with a higher arm number, leading to increased risk variation, negatively impacting the agent's longevity. In contrast, the EA-UCB model counters these rapid hazard fluctuations with two main approaches: limiting exploration in the initial learning phase, and maintaining a relatively stable exploration rate throughout. As a result, EA-UCB notably outperforms its baseline in terms of extended lifetime.

4.5.2 Experiment 2: Variable Reward Environment

In our second experiment, we transitioned from a binary mean reward system to a configuration with distinct mean rewards for each arm. To ensure a symmetrical reward distribution centered around the foraging energy expenditure, M_f, equivalent to half the optimal reward μ_a^*, we employed a logistic function for determining the mean reward of each arm a, irrespective of the arm count K.

$$\mu_a = \frac{\mu_a^*}{1 + \exp(-k \cdot (\frac{a}{K} - \frac{1}{2}))} \tag{4.17}$$

We retained the standard deviation from the first experiment, given by $\sigma_a = 0.02$. This setup ensures that upon extensive sampling of rewards from all arms, half of the rewards will exceed the basal cost, while the other half will fall below. In contrast to Experiment 1, where the count of low-reward arms increased with the arm number,

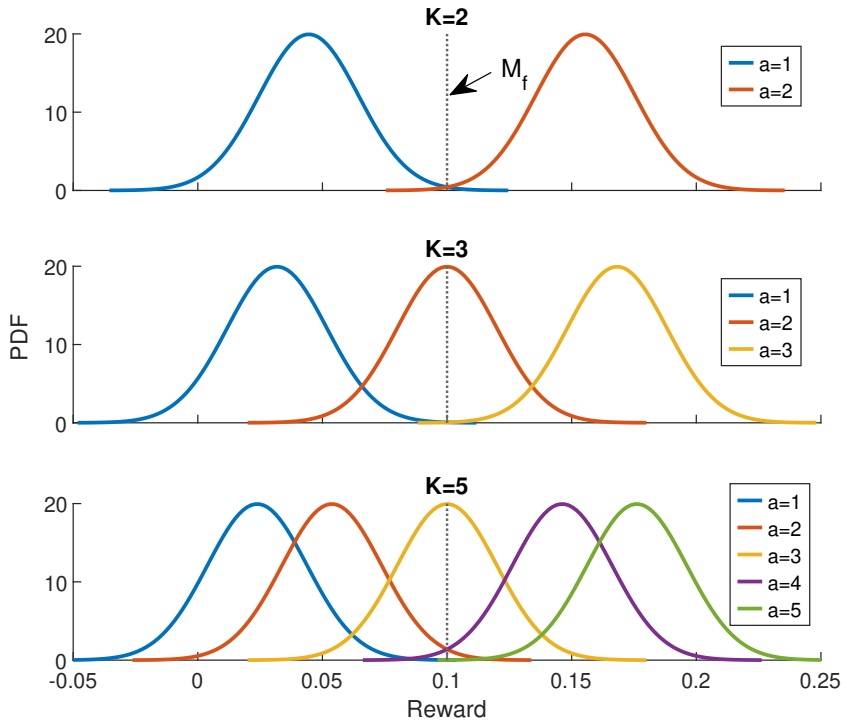

Figure 4.6: Examples of the reward distribution of Experiment 2, with different arm numbers.

the mean rewards for each arm symmetrically diverge from M_f as the increase of the arm number in Experiment 2.

The steepness of the sigmoid curve is controlled by the parameter k. A high k value results in a wider dispersion of mean rewards from the central point $\mu_a^*/2$ with the augmentation of the arm count, and the converse holds for smaller k values. For this experiment, k was preset to 10 to ensure a balanced distribution. Figure 4.6 illustrates the reward distributions for different arm counts. As the number of arms grows, finding the arm with the maximal reward becomes challenging, meanwhile, the mean reward of that arm itself rises.

Figure 4.7 displays the lifetime prediction and regret at the final trial for all considered algorithms. Although TS and EA-TS exhibit superior performance in terms of final regret, they do not necessarily have the longest lifetimes. Contrarily, the UCB1 and EA-UCB models consistently show extended lifetimes in most scenarios. Notably, the EA-UCB model distinctly outlives the other algorithms.

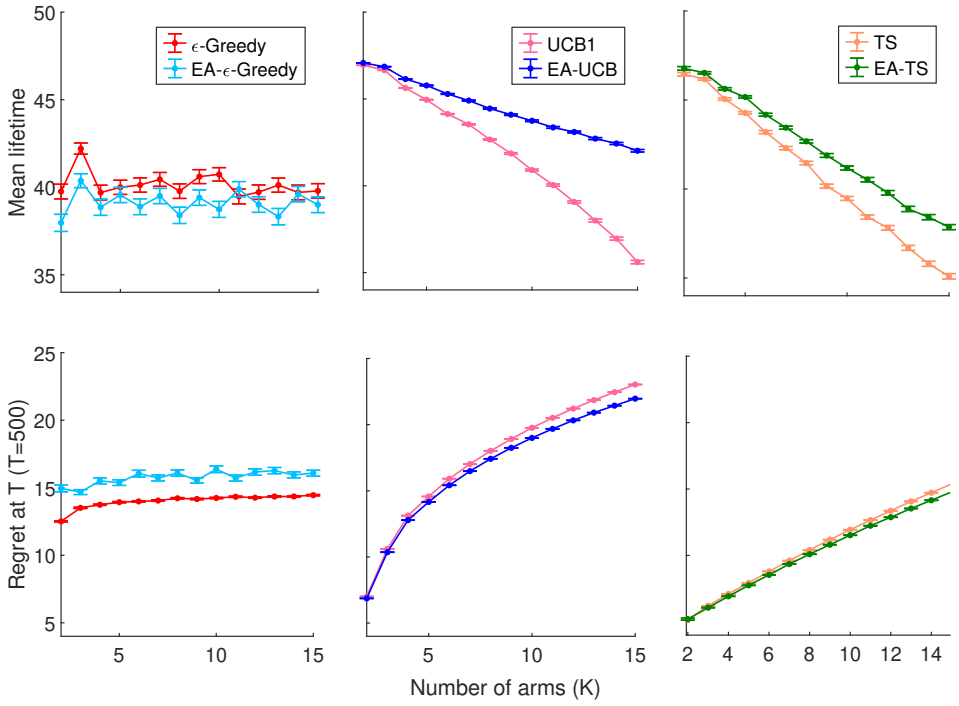

Figure 4.7: The lifetime and the regret value at the final trial ($T = 500$) for testing Experiment 2.

In Experiment 1, the regret of the energy-adaptive variations of the UCB and TS algorithms was comparable to or slightly greater than their baselines. However, in this experiment, the final regret of the energy-adaptive variations for both UCB and TS slightly outperformed their respective baselines. This suggests that EA-UCB and EA-TS not only improve survival for animal agents, but may also enhance regret minimization under specific conditions. It is important to note that when evaluating the theoretical cumulative regret for the bandit model, it is typically assumed that time is infinite. Here, with the final trial set at $T = 500$, the results for the baseline models may not fully reflect their theoretical regret.

For the ϵ-greedy algorithms, both the mean lifetime and regret demonstrate stability across varying arm numbers. In particular, when examining the final trial's regret, although the energy adaptive approach still leads to a higher regret, the EA-ϵ-greedy does not exhibit as pronounced an increase in regret as observed in Experiment 1.

Contrasting with Experiment 1—where the learning processes exhibited distinct patterns for high and low arm numbers—Experiment 2 presents relatively consistent

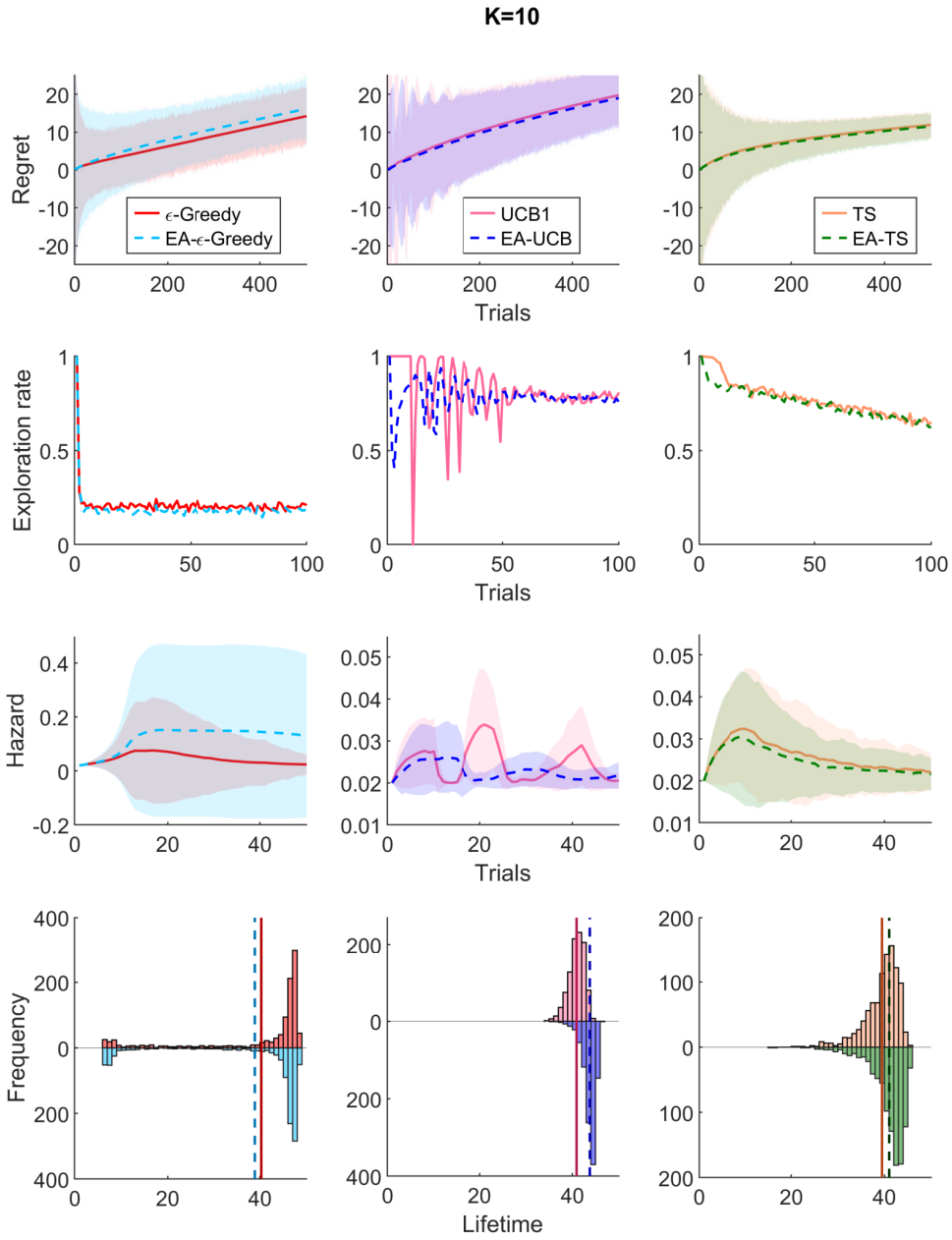

Figure 4.8: Pairwise comparison of Energy-Adaptive algorithms and their baselines in testing Experiment 2 with a setting of 10 arms. First row: Regret evolution over time. Second row: Exploration rate dynamics. Third row: Hazard trajectory over time. Fourth row: distribution of the predicted lifetime, with the mean lifetime represented by a vertical line.

performance during the learning process across different arm counts. To further investigate this, we analyzed the learning process with a fixed arm count of 10, as illustrated in Figure 4.8.

Relative to the ϵ-greedy and TS strategies, the UCB1 model displays the most sig-

nificant fluctuations in both the exploration rate and mean hazard. As observed in Experiment 1, the EA-UCB algorithm curtails the exploration rate during the initial stages, leading to dampened fluctuations. This results in a more streamlined and reduced hazard trajectory, ultimately facilitating a prolonged average lifetime. Concurrently, it achieves a similar, albeit marginally lower, regret over trials.

In its early stages, the EA-TS model curtails a minor amount of the exploration rate, allowing for a reduction in the peak value of the mean hazard trajectory. This adaptation ensures that a larger number of agents achieve prolonged survival. Mirroring the behavior observed in UCB-based algorithms, the regret over trials for the EA-TS is slightly below that of its baseline counterpart.

Echoing findings from Experiment 1, the lifetime distribution observed when deploying ϵ-greedy algorithms showcases a bimodal distribution. This indicates that a substantial proportion of agents either die early in the learning phase or endure until the latter trials. Compared with the baseline model, the EA-ϵ-greedy leads to a notable increase in both the mean and variance of the hazard throughout the learning process, consequently reducing the agents' lifetimes. This implies that the energy-adaptive approach diminishes the chance of exploring arms with adequate rewards in this context.

4.5.3 Impacts of Novel Arm Initialization

As mentioned in Section 4.4.6, to prevent the model from excessively exploring all the unseen arms, especially when the number of arms is large at the initial stages of learning, we introduce an offset value to the count of arm pulls. With an offset value set to $\eta = 1$ for the initial number of pulls n_a^{t1}, the average lifetime and regret for the UCB1 and baseline TS algorithms—both with and without the initial pull offset—are contrasted against their EA equivalents in Experiments 1 and 2. These results are depicted in Figure 4.9 and Figure 4.10. When the number of pulls is offset, the algorithms exhibit similar regret but vary in lifetime compared to their non-offset versions.

Specifically, in Experiment 1 where only one arm offers a high reward and the others yield low rewards, employing an initial pull offset results in a slight enhancement

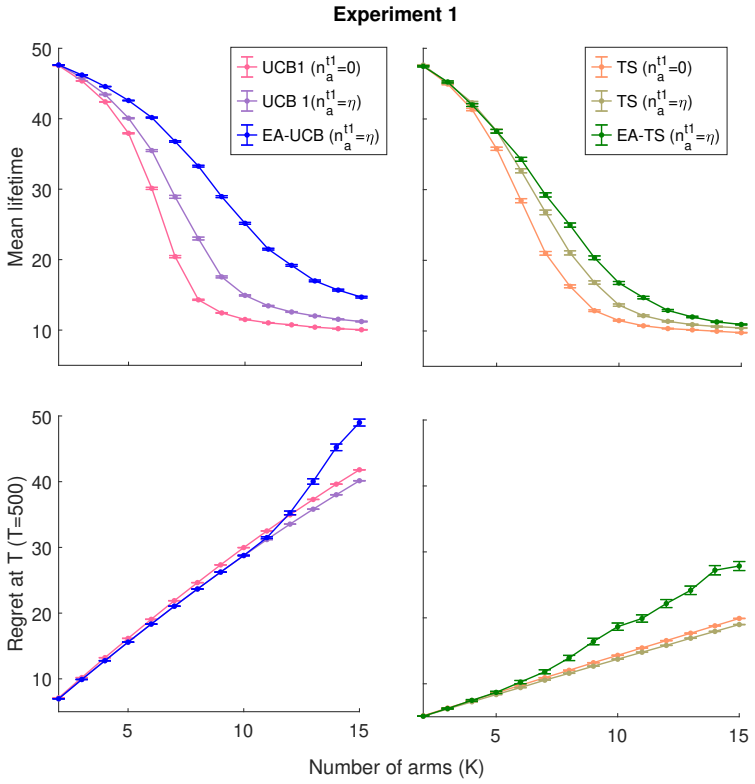

Figure 4.9: Lifetime and regret at the final trial ($T = 500$) from Experiment 1, with the offset value for the initial number of pulls n_a^{t1} set to $\eta = 1$.

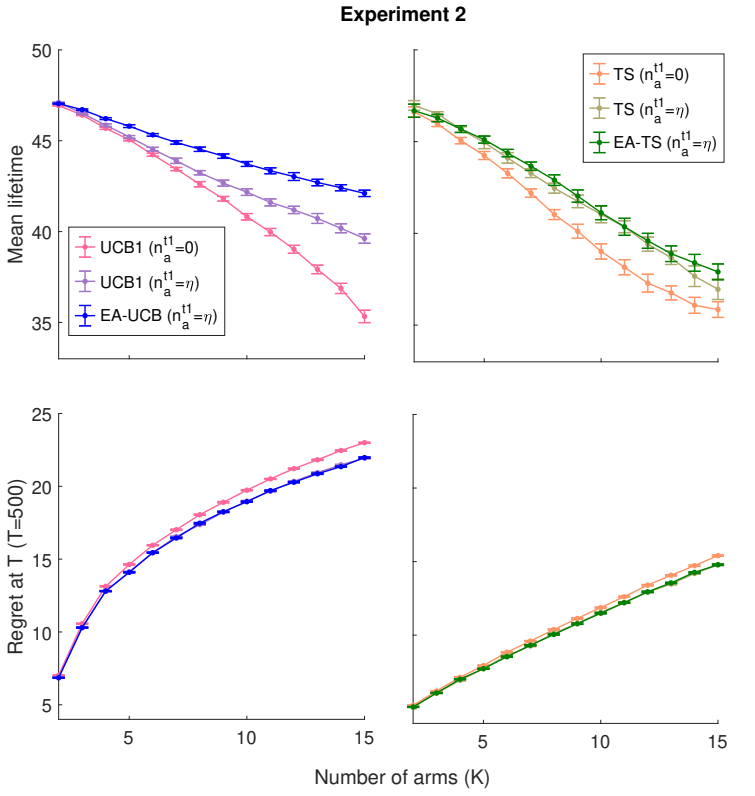

Figure 4.10: Lifetime and regret at the final trial ($T = 500$) from Experiment 2, with the offset value for the number of pulls n_a^{t1} set to $\eta = 1$.

in the mean lifetime for both UCB1 and baseline TS. However, the improvement is less pronounced than what's observed in their EA versions.

Analyzing the evolution of exploration rate and hazard, as presented in Figure 4.11, reveals that the initial pull offset slightly dampens the hazard wave amplitude, although the oscillations persist. In contrast, the EA-UCB substantially smooths the hazard curve, resulting in the most extended predicted lifetime. The exploration rate, when observed with the pull number offset, displays reduced volatility, with the EA-UCB exhibiting the most stable behavior. For TS algorithms, incorporating the initial pull offset mildly curtails the exploration rate and, in turn, the hazard. However, this hazard attenuation is notably less marked compared to its EA counterpart.

In Experiment 2, as the arm reward variety increases with the arm number, UCB1 with an arm pull number offset displays an extended lifetime, as illustrated in Figure 4.10. Regardless of whether UCB1 employs the offset or not, EA-UCB consistently achieves the longest lifetime.

For TS algorithms, introducing the arm pull number offset to baseline TS enhances its lifetime, approaching values just shy of those achieved by EA-TS. An examination of exploration rate and hazard during the learning phase, depicted in Figure 4.12, reveals that the exploration rates of baseline TS and UCB1 using the arm pull number offset closely mirror those of their EA versions. The hazard associated with the baseline TS offset closely aligns with that of EA-TS. In contrast, for UCB-based models, the hazard observed in UCB1 with the offset remains slightly elevated compared to EA-UCB, resulting in a marginally reduced lifetime for the former.

In sum, the augmented lifetimes observed in the EA variants of both UCB and TS algorithms can be attributed to two primary factors: energy-regulated exploration and the initialization of novel arms. For UCB-based algorithms, the energy-regulated exploration in the EA method consistently resulted in a significant extension of lifetime compared to UCB1 with an arm pull number offset. On the other hand, for the TS algorithms, the EA approach's energy-regulated exploration yielded a pronounced lifetime extension in Experiment 1, characterized by a binary reward distribution where only one arm bore a high reward. However, in Experiment 2, which featured

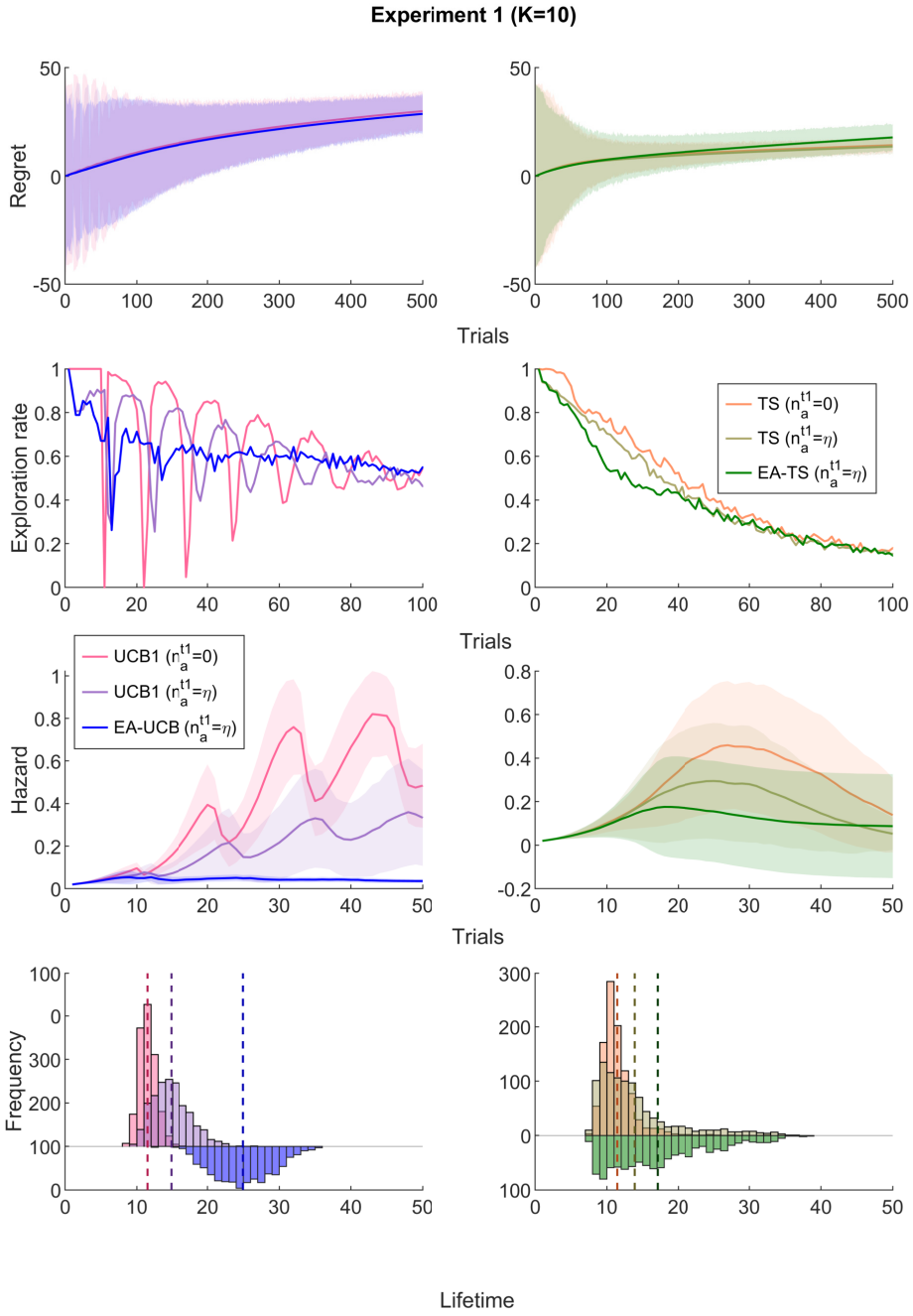

Figure 4.11: Comparison of UCB1 and baseline TS algorithms, considering the presence and absence of the initial pull offset, against their EA counterparts in Experiment 1, utilizing a 10-arm configuration. First row: Regret evolution over time. Second row: Exploration rate dynamics. Third row: Hazard trajectory over time. Fourth row: Lifetime distribution, with the mean lifetime represented by a dashed line.

a broader range of arm rewards, the EA method's advantage diminished, offering only a marginal lifetime increase compared to the baseline model with the arm pull number offset.

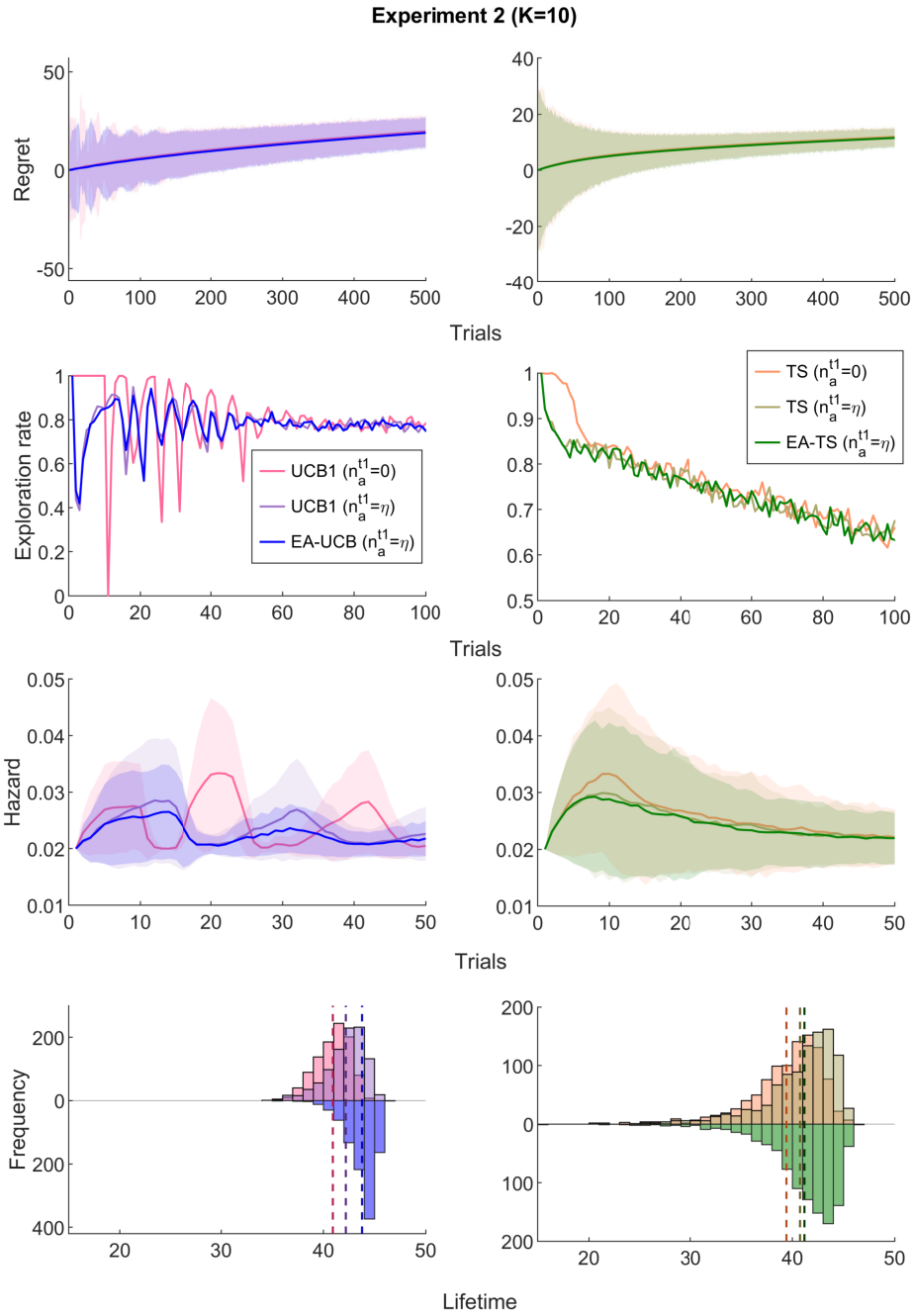

Figure 4.12: Comparison of UCB1 and baseline TS algorithms, considering the presence and absence of the initial pull offset, against their EA counterparts in Experiment 1, utilizing a 10-arm configuration. First row: Regret evolution over time. Second row: Exploration rate dynamics. Third row: Hazard trajectory over time. Fourth row: Lifetime distribution, with the mean lifetime represented by a dashed line.

4.6 Discussion and Conclusion

In this study, we introduced a learning framework based on the MAB problem, factoring in agent energy and lifespan constraints. Energy-adaptive variants of the ϵ-greedy, UCB1, and TS algorithms were introduced, aiming for a harmonized exploration-exploitation trade-off predicated on energy availability. The energy adaptive algorithms, the'optimistic' UCB-based decision-making, and Bayesian-based TS approaches achieved prolonged lifetimes while maintaining comparable regret to traditional algorithms.

The observed extension in lifetime aligns with prior research on animal foraging, implying that the degree of energetic deprivation likely dictates the transition between exploration and exploitation strategies (Lea et al. 2012, Katz & Naug 2015, Lin et al. 2019). Intriguingly, the algorithm having the longest lifetime doesn't necessarily correlate with the lowest regret. For instance, while the TS algorithm consistently registers the least regret across diverse environments, agents utilizing TS seldom achieve the longest lifetime compared to counterparts using other algorithms. This underscores the idea that the decision-making mechanisms in the brain are modulated by a combination of both economic and evolutionary considerations, and may not always lead to optimal decisions (Pearson et al. 2014). Our result could indicate that the intrinsic decision-making system within animal brains may emphasize survival over food acquisition.

Both the EA-UCB and EA-TS algorithms successfully extend agent lifetimes across various testing scenarios, and they offer satisfactory regret values in a robust manner. Given the established efficacy of UCB algorithms in replicating decision-making patterns across various species and foraging contexts (Srivastava et al. 2013, Morimoto 2019, Naito et al. 2022), so as the Bayesian approaches (J. Valone 2006, McNamara et al. 2006, Arganda et al. 2012, Morimoto 2019), these outcomes are in line with expectations.

From the experimental perspective, the efficacy of the model can be assessed through a comparison between the model's predictions and empirical data. For example, the precision of the algorithms in this study can be evaluated using the empirical dataset that examines the foraging behavior of fruit fly larvae across five regions with varying

concentrations of food, as documented by Morimoto et al. (2018). Beyond leveraging the existing dataset, further experiments on Drosophila could adjust the starvation period before learning, and concentrate on gathering data related to the frequency of altering the odor selection and the time of accessing the odor associated with the highest reward.

In our results, EA-UCB consistently outlives the ϵ-greedy and TS-based algorithms under various testing environments. This hints to the UCB's potential to mirror the innate learning mechanisms in organisms during foraging tasks. Given that the UCB approach exhibits an "optimistic" behavior (Lattimore & Szepesvári 2020)—overestimating rewards for less-explored arms—it is conceivable that animals may innately lean towards optimism when presented with positive food reinforcement.

We examined the robustness of the energy-regulated exploration/exploitation behaviors in environments with diverse reward structures, including a single high-reward arm and environments where arm rewards symmetrically deviate from the foraging cost. Both the energy-adaptive variants of the UCB and TS methods showed consistent lifetime extensions across different environments, irrespective of the number of arms. The proposed EA-UCB and EA-TS algorithms, maintain a more stable hazard trajectory by reducing exploration during energy shortages, leading to an extended lifespan. The underlying rationale can be understood through survival analysis, where the predicted lifespan is influenced by the characteristics of hazard fluctuations. When the average fluctuating hazard surpasses the constant hazard, the anticipated lifetime is likely to diminish. Conversely, if the fluctuating hazard, on average, remains below the constant hazard, the predicted lifetime could potentially increase (Clark et al. 2003). Our findings emphasize that an optimal lifetime strategy, for robust performance, requires not only a reduced hazard, but also a steady hazard trajectory.

In the ϵ-greedy methods, the energy adaptive variant doesn't noticeably enhance lifetime or reduce regret. However, it does yield a bimodal lifespan distribution, contrasting the unimodal distribution observed in UCB and TS algorithms. This discrepancy likely stems from the distinct exploration rate. Specifically, while UCB and TS algorithms generally adopt a pattern of intense exploration initially followed by increased exploitation, the ϵ-greedy approach preserves a consistent exploration

rate throughout. This bimodal distribution in the ϵ-greedy methods results in a pronounced variance in lifespan. Notably, despite the variation in the testing environments, certain fortunate agents persistently approach the optimal predicted lifetime, l^*, suggesting their early identification and persistent exploitation of the optimal arm for a majority of the trials. The presence of such "fortunate agents" could elevate the average lifetime, especially in more intricate scenarios, such as when the number of arms increases.

In the initialization phase, our EA algorithms mitigate over-exploration in initial trials to conserve energy. While traditional MAB approaches, unconstrained by energy and lifespan, often sample each arm once during initialization Auer et al. (2002), Slivkins et al. (2019), Lattimore & Szepesvári (2020), incorporating energy constraints can lead agents to riskily explore unsampled arms with dwindling energy reserves. Our solution introduces a minor offset to the arm selection count. Future work can explore more refined solutions to this challenge.

In the late learning phase, for both UCB and TS, the baseline model and their EA variations converge on a similar value of high exploitation rate. This convergence indicates that the energy-adaptive modifications here do not significantly alter the long-term exploration-exploitation balance, focusing instead on optimizing survival in earlier stages.

In this study, the EA method extended lifetime primarily by reducing exploration activity when energy levels dropped, typically by halting the exploration of low-reward arms. Comparing this with a model that randomly reduces exploration time over time, we found that the random reduction can also extend lifetime in certain cases, particularly when applied to the ϵ-greedy approach. However, for UCB and TS, this random reduction only extended lifetime when the decay rate of exploration was set to an optimal value. In contrast, the EA method automatically adjusted exploration reduction to achieve optimal lifetime extension.

In testing Experiment 2, where arms have varying mean rewards, energy adaptive approaches including EA-UCB and EA-TS demonstrate regrets marginally below their corresponding baseline algorithms while consistently excelling in lifespan performance. This implies that curtailing exploration during low-energy phases can

potentially enhance not only agent longevity but also long-term reward accumulation in specific scenarios. When an energy adaptive strategy boosts both lifespan and reward collection (i.e. has less regret), it signifies the model's proficiency in eliminating unnecessary exploration—beneficial even for immortal agents.

In the study of Reinforcement Learning, there's been growing interest in incorporating constraints into the MAB framework. One standout approach is the Bandits with Knapsacks (BwK), which introduces constraints into the MAB paradigm and has found applications in diverse domains. For instance, in online advertising, budget constraints determine the amount of ads displayed to users (Agarwal et al. 2014, Agrawal & Devanur 2014). Similarly, in clinical trials, resources including medical infrastructure and medications can affect the treatment (Badanidiyuru et al. 2018). Our model progressively halts as the resource depletes, offering a learning process for the budget-constrained MAB problems, and showing potential in solving BwK problems.

In conclusion, this work highlights the complexity of applying reinforcement learning strategies to biological agents, emphasizing the value of context-specific strategies and stable hazard trajectories. Our findings illustrate that unconventional methods can be potent under distinct conditions or group objectives, stressing the need for versatile approaches. The novel energy adaptive variations based on ϵ-greedy UCB and TS algorithm, and specialized metrics provide a roadmap for future nuanced exploration in theoretical understanding and practical applications.

Chapter 5

Conclusion and Future Work

This thesis explores the interplay between brain learning and energy, focusing on associative conditioning in Drosophila. We introduced variations of reinforcement learning within the multi-armed bandit framework, inspired by the energy dynamics of fruit flies. These variations address several aspects from the synaptic level to the behavior level, including the influence of metabolic energy on memory pathways, its function as a contextual factor in decision-making, and its role in adjusting the balance between exploration and exploitation.

5.1 Learning Regulated by Energy

In this study, Chapter 2 and Chapter 3 develop two-armed bandit models analogous to the brain structure of fruit flies, suggesting energy-regulated learning mechanisms in biological systems. Subsequently, Chapter 4 transitions to an algorithmic framework, developing multi-armed bandit algorithms inspired by the energy-dependent learning behaviors evident in animals, which applies the RL algorithms as a means to deepen our understanding of learning processes within the brain.

Chapter 2 introduces a computational model detailing how energy influences learning and memory in these systems. This model, centered on fruit flies, demonstrates notable energy savings and adaptability in response to negative stimuli, which contributes to a longer lifespan compared to models with only one memory pathway. Such regulation improves survival in hazardous environments, reflecting findings from Plaçais & Preat (2013) that fruit flies deactivate expensive memory processes to

survive during food scarcity. This observation implies that energy-regulated learning might play a role in optimizing survival during starvation. Furthermore, the modulation of learning by energy may vary according to the reinforcement type. Specifically, starvation has been shown to impede LTM formation in contexts of aversive conditioning (Plaçais & Preat 2013), whereas it seems to facilitate LTM development in situations involving appetitive reinforcement (Krashes & Waddell 2008). This distinction points a direction for future research to explore how energy regulation of LTM formation in appetitive settings impacts survival outcomes.

Inspired by experimental evidence indicating that starvation leads flies to approach food odors (Lin et al. 2019), in Chapter 3, metabolic energy was introduced as a contextual factor influencing decision-making, and the model demonstrates an increased likelihood of an approach action under starvation conditions. This provides a framework for considering the interaction of the energy signals in decision-making, and opens up possibilities for empirical validation and model refinement.

In Chapter 4, we design a learning setup that simulates foraging behavior using an MAB framework. This model integrates factors such as energy expenditure, energy intake, and an estimation of the agent's lifespan based on metabolic energy levels observed throughout the learning process. Additionally, we introduced an energy-regulated exploration/exploitation parameter to various well-known algorithms. Our results indicate that the energy-adaptive UCB and TS methods not only prolong the agent's lifespan but also achieve a level of regret comparable to established baseline algorithms. The results imply the possibility of energy-regulated exploration and exploitation mechanisms within the animal brain. Furthermore, these algorithms demonstrate potential for addressing machine learning challenges with constrained resources.

5.2 Estimation of the Learning Energy

Estimating the metabolic costs involved in learning is a complex task due to the intricate interplay and dynamic nature of animal bodily functions. Additionally, the variability in the costs across different brain regions, affected by aspects like the degrees of connectivity (Tomasi et al. 2013) and the number of neurons (Herculano-

Houzel 2011), which further complicates the estimation. Moreover, our limited understanding of neuron behavior and interaction during the learning process is a significant barrier to fully grasping how neural activities influence metabolic costs at various stages of learning.

This thesis addresses this complexity by focusing on the Drosophila brain, known for its simplicity and genetic manipulability. Our approach to estimating energy costs is twofold: Chapter 2 and Chapter 3 delve into the synaptic level, assessing the energy required for synaptic weight changes, while Chapter 4 extends the analysis to the behavioral level, and evaluate the comprehensive energy expenditure involved in learning, decision-making, and reward acquisition during each experimental trial.

On the synaptic level, this thesis explores the energy expenditure associated with synaptic plasticity by proposing a model where the metabolic cost is directly proportional to the extent of synaptic weight modification. The past research from Li & Van Rossum (2020) applied a similar approach, since protein synthesis plays a key role in memory consolidation, and this process can be energetically costly (Hernandez & Abel 2008). However, while the energy cost is relatively fixed for each amino acid addition to the polypeptide chain during the protein synthesis (Bier et al. 1999), it remains unclear whether the metabolic cost of synaptic modification linearly correlates with the amount of synaptic weight changes. This suggests future research in quantifying protein synthesis and the associated energy expenditure during memory formation.

On the behavioral level, we examine the trade-off in a foraging context, balancing the risk and energy expenditure associated with exploring new options against the known energy intake from familiar sources. In Chapter 4, we postulate a constant foraging energy cost for each learning trial. Whereas in a real case scenario, this may not fully reflect the complexities of an animal's brain, where energy expenditure could vary under different conditions. For example, the study from Huang et al. (2012) has shown a decrease in metabolic power during motor learning, attributed to more efficient muscle coactivation and movement stabilization. Additionally, animals are believed to possess molecular and physiological mechanisms that reduce metabolic costs in response to starvation (Plaçais & Preat 2013, McCue et al. 2017). These findings suggest the potential of future research incorporating dynamic metabolic

rates throughout the learning process.

Moreover, Chapter 3 evaluates learning efficiency by examining the performance enhancement relative to synaptic weight modifications for both potentiation and depression. It is noted that the energy expenditures for synaptic potentiation and depression might differ, with synaptic depression particularly leading to an increase in inactive neurons, suggesting a potential for energy conservation (Harris et al. 2012). Potentiation, on the other hand, may require integrating additional receptors into the postsynaptic membrane, thereby increasing synaptic energy demands (Wieraszko 1982). Future research may distinctly assess the energy costs associated with synaptic potentiation and depression based on their difference in physiological mechanism.

5.3 Energy-Adaptive Learning for Survival

Through survival analysis, this study establishes a connection between energy dynamics and expected lifespan, examining how energy-controlled learning impacts an organism's longevity. In Chapter 2, we model a basic aversive learning task with binary choices of avoidance and approach. The outcomes reveal that energy-modulated memory pathways, characterized by differing memory retention, enhance the longevity of fruit flies. This extension of lifespan is noted in comparison to scenarios where only a single memory pathway is utilized. Chapter 4 explores a foraging scenario with multiple choices, where energy regulates the balance between exploration and exploitation, results demonstrate that energy-regulated learning can extend the lifespan compared to non-regulated methods. The findings of these two chapters suggest that an energy-regulated learning mechanism in the brain exists for evolutionary reasons and aids animal survival in low-energy environments.

This study highlights the varying impact of energy on the lifespan, contingent on several factors, such as the learning phase and the stability of energy changes. As detailed in Chapter 3, effective energy regulation during the initial phase of learning is crucial for prolonging the lifespan of fruit flies. This regulation remains influential in the later stages of learning, though its effect is markedly less pronounced compared to the initial phase. Additionally, the research indicates that when the av-

erage energy level is constant, a stable energy profile is more conducive to longevity than fluctuating energy changes. This suggests that the brain is more inclined to favor learning processes with steady energy modifications, presumably as a survival mechanism.

Beyond extending lifetime, future research could also explore the relationship between energy-adaptive learning mechanisms and survival strategies, for instance, in the context of evolutionary algorithms, which are optimization techniques inspired by natural selection, foucousing on the populations of solutions evolving selection, mutation, and recombination (Whitley et al. 1996). Investigating the impact of energy constraints on these processes could yield valuable insights into optimizing both survival and performance in dynamic environments.

5.4 The Multifaceted Role of Dopamine Neurons

This research has studied the multifaceted role of dopamine neurons in the MB of fruit flies, emphasizing their significance in energy-adaptive learning. As detailed in Section 1.1, DANs are instrumental not only in encoding reinforcement signals, which interact with synaptic plasticity (Waddell 2013), but also in processing environmental contextual information, including energy signals (Lin et al. 2019, Zolin et al. 2021). This dual functionality enables fruit flies to adapt their behavior in dynamic environments. Building on these empirical findings, our models, as outlined in Chapter 2 and Chapter 4, leverage DANs to encode reinforcement signals and regulate synaptic weight changes. Moreover, this study extends the role of DANs in influencing memory retention and the bias induced by contextual information, suggesting a broader scope of DAN involvement in energy-adaptive learning.

In Chapter 2, we propose a DAN-modulated threshold model for LTM formation, positing an energy-adaptive learning mechanism governed by dopaminergic reward prediction error signals. This model suggests that stronger stimuli generate larger error signals, thereby enhancing LTM utilization. Compared to models with a fixed consolidation threshold, our approach increases LTM longevity and decreases unnecessary LTM formation, highlighting the adaptability of biological memory systems to dopaminergic signal intensity. This aligns with existing research on dopamine's

critical role in memory persistence in mammals (Lisman et al. 2011, O'Carroll et al. 2006, Bethus et al. 2010).

Chapter 4 delves into the impact of energy signals on decision-making processes in DANs prior to learning. We introduced a framework that integrates hunger signals as a contextual influence in decision-making, drawing upon empirical evidence that demonstrates fruit flies exhibit a heightened tendency to approach odors under conditions of hunger (Inagaki et al. 2014, Lin et al. 2019). This approach offers a novel method for studying DANs in various contexts, including signals related to the internal states and external environment. Future work could involve fitting experimental data into this model for further validation, and applying the framework to additional contextual factors beyond energy.

5.5 Reinforcement Learning in the Brain

The hypothesis that the brain employs RL mechanisms is supported by evidence of dopamine signaling mirroring temporal difference (TD) algorithms (Schultz et al. 1997, Schultz 2002). Empirical studies, identifying TD error correlates in dopamine-centric regions like the ventral and dorsal striatum, further corroborate this (Schönberg et al. 2007, Niv et al. 2012). Despite these advancements, the detailed workings of RL in the brain remain partially understood. Here, Chapter 2 and Chapter 3 investigate the learning dynamics within the MB and its relationship with energy states by mapping MB's neuronal learning onto two-armed bandit frameworks. Chapter 4 employs MAB algorithms to deepen the understanding of energy-regulated exploration and exploitation mechanisms. This research underscores the utility of RL modeling as a tool for exploring the complexities of reward learning and decision-making processes.

By comparing biological learning mechanisms with artificial RL models, the study in this thesis can enrich our understanding of learning and decision-making. Chapter 2 delves into memory retention variability, drawing parallels to reward discounting in RL. We observe that low-retention memory weakens over time, influencing decisions based on recent memories, akin to the effect of a low discount factor in RL prioritizing immediate rewards. Prior studies also link dopamine response to reward

prediction errors in RL, suggesting that memory decay can boost learning motivation (Kato & Morita 2016). In Chapter 3, the application of contextual bandit approaches, commonly used in RL-based recommendation systems, aids in exploring Drosophila's brain learning mechanisms in relation to metabolic energy. Inspired by the empirical findings (Inagaki et al. 2014, Lin et al. 2019), the model expresses an increased tendency to approach an odor under starvation. Furthermore, in Chapter 4, an energy-adaptive variation of the UCB algorithm demonstrated significant lifetime extension compared to its baseline, suggesting that energy-efficient learning strategies have the potential to reduce unnecessary exploration in the animal brain.

Also, this study underscores the divergence between animal behavior and artificial RL models. Although the RL learning rules can be elegant mathematically and exhibit remarkable learning performance, past research argued it's challenging for the brain to apply these rules directly due to various reasons, such as its computational complexity (Krebs et al. 1978). The empirical study on mice trained in tasks with changing reward probabilities revealed deviations from the performance of an ideal RL agent. Instead, the mice adopted a near-optimal strategy that can be characterized by a collection of equivalent models (Beron et al. 2022). As discussed in Chapter 2 and Chapter 4, when applying the energy-efficient learning methods, animals may sacrifice learning performance for reduced metabolic costs, which is different from traditional RL models that prioritize reward maximization. Specifically, observation in Chapter 4 suggests that animal the brain's innate RL system employs evaluation metrics rooted in evolutionary and economic considerations, such as survival, rather than solely optimizing for reward collection.

Appendix A

Neuroeconomic Trade-off for learning in Chapter 2

The optimal memory strategy maximizes the lifespan. The animal has to decide whether to invest energy in LTM of the CS-US associate. The situation can be compared to the human dilemma of whether to spend money on education: investment in education will on average pay off financially in the long run, but only if the life expectancy is long enough and bankruptcy can be avoided. In general, the optimal strategy will depend on the unknown future, which might include extra energy rewards or starvation, yet one can hope to develop a robust heuristic strategy.

We derive an expression for the change in lifetime given a small weight update, which in turn leads to a small change in the hazards $\delta h(t)$. Under this assumption, the change in expected lifetime between LTM and no learning (NL) can be expanded as

$$l - l^{NL} \approx -\sum_t \left[e^{-\sum_{t'} h^{NL}(t')} \sum_{t'} \delta h(t') \right]$$

Given ARM learning with a small weight change Δw, the temporary reduction in stimulus hazard is

$$\delta h_s(t) = |\Delta w| h_s^0 \frac{\partial P_-(w_-, w_+, \mu, t)}{\partial w_-} \exp(-t \log \gamma)$$

The calculation of the approach probability $P_-(w_-, w_+, \mu, t)$ can be found in Equation (2.2). For LTM learning, the expression is similar but the decay term is absent.

LTM learning at the same time increases starvation hazard as

$$\delta h_M = c_{LTM}|\Delta w| h_M^{NL}$$

The difference in expected lifetime between ARM and LTM learning is in first order of $|\Delta w|$,

$$l^{ARM} - l^{LTM} \approx |\Delta w| \sum_t \left[e^{-\sum_{t'} h^{NL}(t')} \sum_{t'} \left\{ h_s^0 \frac{\partial P_-(w_-, w_+, \mu, t')}{\partial w_-} (1 - e^{-t' \log \gamma}) + c_{LTM} h_M^{NL} \right\} \right] \tag{A.1}$$

Where it should be noted that because learning decreases the probability of encountering the stimulus ($\partial P/\partial w < 0$), the first term in the curly brackets is negative, while the second term is strictly positive. h_s^0 denotes the stimulus hazard if it is approached.

When the lifetime difference is larger than zero, ARM learning should be chosen over LTM learning. While complex, the expression gives insight in when ARM memory is preferable to LTM. It happens when: 1) The stimulus hazard h_s^0 is small, 2) when the impact of the learning on the choice probability $\partial P/\partial w$ is small, e.g. late in the learning process, 3) the ARM decay γ is slow, and 4) the energy cost of LTM, c_{LTM} is high. Finally, the first r.h.s term attenuates the benefit of long-lasting memory, so that ARM is generally preferable when the expected lifetime is short.

Nevertheless, it would appear challenging for a fly to estimate the expected lifetime based on this expression to decide whether to use ARM or LTM memory, looking for approximate heuristic algorithms that only rely on observables directly accessible by the organism and are close to optimal under various conditions.

Appendix B

The Analysis of the Single-trial Learning in Chapter 3

The probability of taking the desired action P_+, with respect to the initial weight values, the weight modification Δw^{Σ}, and the synaptic adjustment ratio α indicating the fraction of synaptic modifications directed towards the desired behavior, for the four synaptic adjustments strategies detailed in Section 3.3.1 are:

- Potentiate both w^+ and w^- (top to right edge):

$$P_+(\alpha, \Delta w^{\Sigma}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\mu\left((2\alpha-1)\Delta w^{\Sigma} - w_0^- - w_0^+\right)}{\sqrt{2}\sigma\sqrt{((1-\alpha)\Delta w^{\Sigma} + w_0^-)^2 + (\alpha\Delta w^{\Sigma} + w_0^+)^2}}\right)\right]$$

- Potentiate w^+ and depress w^- (bottom to right edge):

$$P_+(\alpha, \Delta w^{\Sigma}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\mu\left(\Delta w^{\Sigma} - w_0^- + w_0^+\right)}{\sqrt{2}\sigma\sqrt{((\alpha-1)\Delta w^{\Sigma} + w_0^-)^2 + (\alpha\Delta w^{\Sigma} + w_0^+)^2}}\right)\right]$$

- Depress both w^+ and w^- (left to bottom edge):

$$P_+(\alpha, \Delta w^{\Sigma}) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{\mu\left((2\alpha-1)\Delta w^{\Sigma} + w_0^- - w_0^+\right)}{\sqrt{2}\sigma\sqrt{((\alpha-1)\Delta w^{\Sigma} + w_0^-)^2 + (w_0^+ - \alpha\Delta w^{\Sigma})^2}}\right)\right]$$

- Depress w^+ and potentiate w^- (left to top edge):

$$P_+(\alpha, \Delta w^{\Sigma}) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{\mu\left((\Delta w^{\Sigma} + w_0^- - w_0^+\right)}{\sqrt{2}\sigma\sqrt{((1-\alpha)\Delta w^{\Sigma} + w_0^-)^2 + (w_0^+ - \alpha\Delta w^{\Sigma})^2}}\right)\right]$$

When there's no bias, the value of both w_0^+ and w_0^- are 0.5, the derivative of the P_+ with respect to the synaptic adjustment ratio α, under the assumptions that α, Δw^{Σ}, w^+, and w^- are bounded within [0, 1], and that μ and σ are strictly positive, the resulting expression when potentiating both w^+ and w^- (the top to right edge in Figure 3.2 (a)) is as:

$$\frac{\partial P_+}{\partial \alpha} = \frac{0.2\Delta w^{\Sigma}(1+\Delta w^{\Sigma})^2 e^{-\left(\frac{((-1+2\alpha)\Delta w^{\Sigma})^2\mu^2}{2((0.5+\Delta w^{\Sigma}-\alpha\Delta w)^2+(0.5+\alpha\Delta w^{\Sigma})^2)\sigma^2}\right)\mu}}{(0.25+0.5(1-\alpha)\Delta w+0.5\alpha\Delta w+(0.5-\alpha+\alpha^2)(\Delta w^{\Sigma})^2)\sqrt{(0.5+\Delta w^{\Sigma}-\alpha\Delta w^{\Sigma})^2+(0.5+\alpha\Delta w^{\Sigma})^2}\sigma} \geq 0 \quad \text{(B.1)}$$

In the analysis of the derivative denoted in Equation (B.1), a component-wise examination reveals its inherent non-negativity across the defined ranges of α and Δw^{Σ}. This conclusion is supported by several factors: the exponential function e^{-x}, which is always positive for any real x, reinforcing the derivative's non-negativity; the numerator's composition of non-negative terms, including a constant multiplier, Δw^{Σ} and its square, alongside the positive parameter μ; and the denominator's structure, which combines a square root of the sum of squares and the positive parameter σ, alongside polynomial components dependent on α and Δw^{Σ}, all ensuring the overall non-negativity of $\frac{\partial P_+}{\partial \alpha}$. Employing a consistent methodology enables the determination of the range for $\frac{\partial P_+}{\partial \alpha}$ across the remaining three synaptic adjustment strategies:

- Potentiate w^+ and depress w^- (bottom to right edge):

$$\frac{\partial P_+}{\partial \alpha} = -\frac{0.4(\Delta w^{\Sigma})^2(1+(-1+2\alpha)\Delta w^{\Sigma})\, e^{-\left(\frac{(\Delta w^{\Sigma})^2\mu^2}{2\left(0.5+2(0.5(-1+\alpha)+0.5\alpha)\Delta w^{\Sigma}+(1-2\alpha+2\alpha^2)(\Delta w^{\Sigma})^2\right)\sigma^2}\right)\mu}}{\left((0.5+(-1+\alpha)\Delta w^{\Sigma})^2+(0.5+\alpha\Delta w^{\Sigma})^2\right)^{3/2}\sigma} \leq 0$$

- Depress both w^+ and w^- (left to bottom edge):

$$\frac{\partial P_+}{\partial \alpha} = -\frac{0.4(\Delta w^{\Sigma})^2(1+(1-2\alpha)\Delta w^{\Sigma}) e^{-\left(\frac{(\Delta w^{\Sigma})^2\mu^2}{2\left((0.5-\alpha\Delta w^{\Sigma})^2+(0.5+\Delta w^{\Sigma}-\alpha\Delta w^{\Sigma})^2\right)\sigma^2}\right)\mu}}{\left((0.5-\alpha\Delta w^{\Sigma})^2+(0.5+\Delta w^{\Sigma}-\alpha\Delta w^{\Sigma})^2\right)^{3/2}\sigma} \leq 0$$

- Depress w^+ and potentiate w^- (left to top edge):

$$\frac{\partial P_+}{\partial \alpha} = -\frac{0.2(-1+\Delta w^{\Sigma})^2 \Delta w^{\Sigma} e^{-\left(\frac{((-1+2\alpha)\Delta w^{\Sigma})^2 \mu^2}{2\left((0.5+(-1+\alpha)\Delta w^{\Sigma})^2+(0.5-\alpha\Delta w^{\Sigma})^2\right)\sigma^2}\right)\mu}}{\left(0.25-0.5\Delta w^{\Sigma}+(0.5-\alpha+\alpha^2)(\Delta w^{\Sigma})^2\right)\sqrt{(0.5+(-1+\alpha)\Delta w^{\Sigma})^2+(0.5-\alpha\Delta w^{\Sigma})^2}\sigma} \leq 0$$

For all these four synaptic adjustment strategies under consideration, the sign of $\frac{\partial P_+}{\partial \alpha}$ remains constant across the defined ranges of α and Δw^{Σ}, indicating that, as shown in Figure 3.2 (a), the extrema consistently occur at the corners. This implies that in the absence of bias before this single-trial learning, the most effective learning strategy invariably utilizes the entirety of the weight change to either potentiate or depress a single synapse.

Bibliography

Abdelrahman, N. (2023), Modeling olfactory processing and insights on optimal learning in constrained neural networks: learning from the anatomy of the Drosophila mushroom body, PhD thesis, University of Sheffield.

Adams, G. K., Watson, K. K., Pearson, J. & Platt, M. L. (2012), 'Neuroethology of decision-making', *Current opinion in neurobiology* **22**(6), 982–989.

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L. & Platt, M. L. (2017), 'A primer on foraging and the explore/exploit trade-off for psychiatry research', *Neuropsychopharmacology* **42**(10), 1931–1939.

Agarwal, D., Long, B., Traupman, J., Xin, D. & Zhang, L. (2014), Laser: A scalable response prediction platform for online advertising, *in* 'Proceedings of the 7th ACM international conference on Web search and data mining', pp. 173–182.

Agrawal, R., Hegde, M., Teneketzis, D. et al. (1990), 'Multi-armed bandit problems with multiple plays and switching cost', *Stochastics and Stochastic reports* **29**(4), 437–459.

Agrawal, S. & Devanur, N. R. (2014), Bandits with concave rewards and convex knapsacks, *in* 'Proceedings of the fifteenth ACM conference on Economics and computation', pp. 989–1006.

Amin, H. & Lin, A. C. (2019), 'Neuronal mechanisms underlying innate and learned olfactory processing in drosophila', *Current Opinion in Insect Science* **36**, 9–17.

Arganda, S., Pérez-Escudero, A. & de Polavieja, G. G. (2012), 'A common rule for decision making in animal collectives across species', *Proceedings of the National Academy of Sciences* **109**(50), 20508–20513.

Aso, Y., Herb, A., Ogueta, M., Siwanowicz, I., Templier, T., Friedrich, A. B., Ito, K., Scholz, H. & Tanimoto, H. (2012), 'Three dopamine pathways induce aversive odor memories with different stability', *PLoS genetics* **8**(7), e1002768.

Aso, Y., Siwanowicz, I., Bräcker, L., Ito, K., Kitamoto, T. & Tanimoto, H. (2010), 'Specific dopaminergic neurons for the formation of labile aversive memory', *Current biology* **20**(16), 1445–1451.

Aso, Y. et al. (2014*a*), 'Mushroom body output neurons encode valence and guide memory-based action selection in drosophila', *elife* **3**, e04580.

Aso, Y. et al. (2014*b*), 'The neuronal architecture of the mushroom body provides a logic for associative learning', *elife* **3**, e04577.

Attwell, D. & Laughlin, S. B. (2001), 'An energy budget for signaling in the grey matter of the brain', *Journal of Cerebral Blood Flow & Metabolism* **21**(10), 1133–1145.

Auer, P., Cesa-Bianchi, N. & Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine learning* **47**, 235–256.

Averbeck, B. B. (2015), 'Theory of choice in bandit, information sampling and foraging tasks', *PLoS computational biology* **11**(3), e1004164.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Filho, W. J., Lent, R. & Herculano-Houzel, S. (2009), 'Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain', *Journal of Comparative Neurology* **513**(5), 532–541.

Badanidiyuru, A., Kleinberg, R. & Slivkins, A. (2018), 'Bandits with knapsacks', *Journal of the ACM (JACM)* **65**(3), 1–55.

Balasubramanian, V. (2021), 'Brain power', *Proceedings of the National Academy of Sciences* **118**(32), e2107022118.

Bélanger, M., Allaman, I. & Magistretti, P. J. (2011), 'Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation', *Cell metabolism* **14**(6), 724–738.

Bell, W. J. (2012), *Searching behaviour: the behavioural ecology of finding resources*, Springer Science & Business Media.

Bell, W. J., Cathy, T., Roggero, R. J., Kipp, L. R. & Tobin, T. R. (1985), 'Sucrose-stimulated searching behaviour of drosophila melanogaster in a uniform habitat: modulation by period of deprivation', *Animal Behaviour* **33**(2), 436–448.

Bennett, J. E., Philippides, A. & Nowotny, T. (2021), 'Learning with reinforcement prediction errors in a model of the drosophila mushroom body', *Nature communications* **12**(1), 1–14.

Beron, C. C., Neufeld, S. Q., Linderman, S. W. & Sabatini, B. L. (2022), 'Mice exhibit stochastic and efficient action switching during probabilistic decision making', *Proceedings of the National Academy of Sciences* **119**(15), e2113961119.

Bethus, I., Tse, D. & Morris, R. G. (2010), 'Dopamine and memory: modulation of the persistence of memory for novel hippocampal nmda receptor-dependent paired associates', *Journal of Neuroscience* **30**(5), 1610–1618.

Bier, D. M. et al. (1999), 'The energy costs of protein metabolism: lean and mean on uncle sam's team', *The role of protein and amino acids in sustaining and enhancing performance* pp. 109–119.

Boehm, A. C., Friedrich, A. B., Hunt, S., Bandow, P., Siju, K., De Backer, J. F., Claussen, J., Link, M. H., Hofmann, T. F., Dawid, C. et al. (2022), 'A dopamine-gated learning circuit underpins reproductive state-dependent odor preference in drosophila females', *Elife* **11**, e77643.

Brea, J., Urbanczik, R. & Senn, W. (2014), 'A normative theory of forgetting: lessons from the fruit fly', *PLoS computational biology* **10**(6), e1003640.

Brown, M. R., Crim, J. W., Arata, R. C., Cai, H. N., Chun, C. & Shen, P. (1999), 'Identification of a drosophila brain-gut peptide related to the neuropeptide y family', *Peptides* **20**(9), 1035–1042.

Bruckmaier, M., Tachtsidis, I., Phan, P. & Lavie, N. (2020), 'Attention and capacity limits in perception: A cellular metabolism account', *Journal of Neuroscience* **40**(35), 6801–6811.

Burke, C. J., Huetteroth, W., Owald, D., Perisse, E., Krashes, M. J., Das, G., Gohl, D., Silies, M., Certel, S. & Waddell, S. (2012), 'Layered reward signalling through octopamine and dopamine in drosophila', *Nature* **492**(7429), 433–437.

Busto, G. U., Cervantes-Sandoval, I. & Davis, R. L. (2010), 'Olfactory learning in drosophila', *Physiology* **25**(6), 338–346.

Campbell, R. A., Honegger, K. S., Qin, H., Li, W., Demir, E. & Turner, G. C. (2013), 'Imaging a population code for odor identity in the drosophila mushroom body', *Journal of Neuroscience* **33**(25), 10568–10581.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R. & Stoltz, G. (2013), 'Kullback-leibler upper confidence bounds for optimal sequential allocation', *The Annals of Statistics* pp. 1516–1541.

Carter, E. C. & Redish, A. D. (2016), 'Rats value time differently on equivalent foraging and delay-discounting tasks.', *Journal of Experimental Psychology: General* **145**(9), 1093.

Chapelle, O. & Li, L. (2011), 'An empirical evaluation of thompson sampling', *Advances in neural information processing systems* **24**.

Charnov, E. L. (1976), 'Optimal foraging, the marginal value theorem', *Theoretical population biology* **9**(2), 129–136.

Chittka, L. & Niven, J. (2009), 'Are bigger brains better?', *Current biology* **19**(21), R995–R1008.

Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. (2003), 'Survival analysis part i: basic concepts and first analyses', *British journal of cancer* **89**(2), 232–238.

Cognigni, P., Felsenberg, J. & Waddell, S. (2018), 'Do the right thing: neural network mechanisms of memory formation, expression and update in drosophila', *Current opinion in neurobiology* **49**, 51–58.

Cohen, J. D., McClure, S. M. & Yu, A. J. (2007), 'Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration', *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**(1481), 933–942.

Cohn, R., Morantte, I. & Ruta, V. (2015), 'Coordinated and compartmentalized neuromodulation shapes sensory processing in drosophila', *Cell* **163**(7), 1742–1755.

Colomb, J., Kaiser, L., Chabaud, M.-A. & Preat, T. (2009), 'Parametric and genetic analysis of drosophila appetitive long-term memory and sugar motivation', *Genes, Brain and Behavior* **8**(4), 407–415.

Davis, R. L. (2004), 'Olfactory learning', *Neuron* **44**(1), 31–48.

Davis, R. L. (2005), 'Olfactory memory formation in drosophila: from molecular to systems neuroscience', *Annu. Rev. Neurosci.* **28**, 275–302.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. (2006), 'Cortical substrates for exploratory decisions in humans', *Nature* **441**(7095), 876–879.

Fanson, B. G., Weldon, C. W., Pérez-Staples, D., Simpson, S. J. & Taylor, P. W. (2009), 'Nutrients, not caloric restriction, extend lifespan in queensland fruit flies (bactrocera tryoni)', *Aging cell* **8**(5), 514–523.

Fonseca-Azevedo, K. & Herculano-Houzel, S. (2012), 'Metabolic constraint imposes tradeoff between body size and number of brain neurons in human evolution', *Proceedings of the National Academy of Sciences* **109**(45), 18571–18576.

Gagliano, M., Vyazovskiy, V. V., Borbély, A. A., Grimonprez, M. & Depczynski, M. (2016), 'Learning by association in plants', *Scientific reports* **6**(1), 38427.

Gerstner, W., Kistler, W. M., Naud, R. & Paninski, L. (2014), *Neuronal dynamics: From single neurons to networks and models of cognition*, Cambridge University Press.

Girard, M., Jiang, J. & van Rossum, M. C. (2023), 'Estimating the energy requirements for long term memory formation', *arxiv* p. 2301.09565.

Gittins, J., Glazebrook, K. & Weber, R. (2011), *Multi-armed bandit allocation indices*, John Wiley & Sons.

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A. & Dally, W. J. (2016), 'Eie: Efficient inference engine on compressed deep neural network', *ACM SIGARCH Computer Architecture News* **44**(3), 243–254.

Harris, J. J., Jolivet, R. & Attwell, D. (2012), 'Synaptic energy use and supply', *Neuron* **75**(5), 762–777.

Hattori, D., Aso, Y., Swartz, K. J., Rubin, G. M., Abbott, L. & Axel, R. (2017), 'Representations of novelty and familiarity in a mushroom body compartment', *Cell* **169**(5), 956–969.

Herculano-Houzel, S. (2011), 'Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution', *PloS one* **6**(3), e17514.

Hernandez, P. J. & Abel, T. (2008), 'The role of protein synthesis in memory consolidation: progress amid decades of debate', *Neurobiology of learning and memory* **89**(3), 293–311.

Hige, T., Aso, Y., Modi, M. N., Rubin, G. M. & Turner, G. C. (2015), 'Heterosynaptic plasticity underlies aversive olfactory learning in drosophila', *Neuron* **88**(5), 985–998.

Hoeffding, W. (1994), 'Probability inequalities for sums of bounded random variables', *The collected works of Wassily Hoeffding* pp. 409–426.

Holliday, M., Potter, D., Jarrah, A. & Bearg, S. (1967), 'The relation of metabolic rate to body weight and organ size', *Pediatric research* **1**(3), 185–195.

Howarth, C., Peppiatt-Wildman, C. M. & Attwell, D. (2010), 'The energy use associated with neural computation in the cerebellum', *Journal of cerebral blood flow & metabolism* **30**(2), 403–414.

Huang, H. J., Kram, R. & Ahmed, A. A. (2012), 'Reduction of metabolic cost during motor learning of arm reaching dynamics', *Journal of Neuroscience* **32**(6), 2182–2190.

Huetteroth, W., Perisse, E., Lin, S., Klappenbach, M., Burke, C. & Waddell, S. (2015), 'Sweet taste and nutrient value subdivide rewarding dopaminergic neurons in drosophila', *Current biology* **25**(6), 751–758.

Inagaki, H. K., Panse, K. M. & Anderson, D. J. (2014), 'Independent, reciprocal neuromodulatory control of sweet and bitter taste sensitivity during starvation in drosophila', *Neuron* **84**(4), 806–820.

Isabel, G., Pascual, A. & Preat, T. (2004), 'Exclusive consolidated memory phases in drosophila', *Science* **304**(5673), 1024–1027.

J. Valone, T. (2006), 'Are animals capable of bayesian updating? an empirical review', *Oikos* **112**(2), 252–259.

Karbowski, J. (2019), 'Metabolic constraints on synaptic learning and memory', *Journal of neurophysiology* **122**(4), 1473–1490.

Kato, A. & Morita, K. (2016), 'Forgetting in reinforcement learning links sustained dopamine signals to motivation', *PLoS computational biology* **12**(10), e1005145.

Katz, K. & Naug, D. (2015), 'Energetic state regulates the exploration–exploitation trade-off in honeybees', *Behavioral Ecology* **26**(4), 1045–1050.

Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L. & Chawla, S. (2015), 'Efficient thompson sampling for online matrix-factorization recommendation', *Advances in neural information processing systems* **28**.

Keasar, T., Rashkovich, E., Cohen, D. & Shmida, A. (2002), 'Bees in two-armed bandit situations: foraging choices and possible decision mechanisms', *Behavioral Ecology* **13**(6), 757–765.

Kim, Y.-C., Lee, H.-G. & Han, K.-A. (2007), 'D1 dopamine receptor dda1 is required in the mushroom body neurons for aversive and appetitive learning in drosophila', *Journal of Neuroscience* **27**(29), 7640–7647.

Krashes, M. J., DasGupta, S., Vreede, A., White, B., Armstrong, J. D. & Waddell, S. (2009), 'A neural circuit mechanism integrating motivational state with memory expression in drosophila', *Cell* **139**(2), 416–427.

Krashes, M. J., Keene, A. C., Leung, B., Armstrong, J. D. & Waddell, S. (2007), 'Sequential use of mushroom body neuron subsets during drosophila odor memory processing', *Neuron* **53**(1), 103–115.

Krashes, M. J. & Waddell, S. (2008), 'Rapid consolidation to a radish and protein synthesis-dependent long-term memory after single-session appetitive olfactory conditioning in drosophila', *Journal of Neuroscience* **28**(12), 3103–3113.

Krebs, J. R., Kacelnik, A. & Taylor, P. (1978), 'Test of optimal sampling by foraging great tits', *Nature* **275**(5675), 27–31.

Krittika, S. & Yadav, P. (2019), 'An overview of two decades of diet restriction studies using drosophila', *Biogerontology* **20**(6), 723–740.

Lattimore, T. & Szepesvári, C. (2020), *Bandit algorithms*, Cambridge University Press.

Laughlin, S. B., de Ruyter van Steveninck, R. R. & Anderson, J. C. (1998), 'The metabolic cost of neural information', *Nature neuroscience* **1**(1), 36–41.

Lea, S. E., McLaren, I. P., Dow, S. M. & Graft, D. A. (2012), 'The cognitive mechanisms of optimal sampling', *Behavioural processes* **89**(2), 77–85.

Li, H. L. & Van Rossum, M. C. (2020), 'Energy efficient synaptic plasticity', *Elife* **9**, e50804.

Lin, S. (2023), 'The making of the drosophila mushroom body', *Frontiers in Physiology* **14**, 5.

Lin, S., Senapati, B. & Tsao, C.-H. (2019), 'Neural basis of hunger-driven behaviour in drosophila', *Open biology* **9**(3), 180259.

Linford, N. J., Bilgir, C., Ro, J. & Pletcher, S. D. (2013), 'Measurement of lifespan in drosophila melanogaster', *JoVE (Journal of Visualized Experiments)* (71), e50068.

Lisman, J., Grace, A. A. & Duzel, E. (2011), 'A neohebbian framework for episodic memory; role of dopamine-dependent late ltp', *Trends in neurosciences* **34**(10), 536–547.

Loewenstein, Y. (2008), 'Robustness of learning that is based on covariance-driven synaptic plasticity', *PLoS Computational Biology* **4**(3), e1000007.

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M. et al. (2012), 'The drosophila melanogaster genetic reference panel', *Nature* **482**(7384), 173–178.

Mair, W., Piper, M. D. W. & Partridge, L. (2005), 'Calories do not explain extension of life span by dietary restriction in drosophila', *PLoS biology* **3**(7), e223.

Mao, Z. & Davis, R. L. (2009), 'Eight different types of dopaminergic neurons innervate the drosophila mushroom body neuropil: anatomical and physiological heterogeneity', *Frontiers in neural circuits* **3**, 5.

Margulies, C., Tully, T. & Dubnau, J. (2005), 'Deconstructing memory in drosophila', *Current biology* **15**(17), R700–R713.

McCue, M. D., Terblanche, J. S. & Benoit, J. B. (2017), 'Learning to starve: impacts of food limitation beyond the stress period', *Journal of Experimental Biology* **220**(23), 4330–4338.

McNamara, J. M., Green, R. F. & Olsson, O. (2006), 'Bayes' theorem and its applications in animal behaviour', *Oikos* **112**(2), 243–251.

McNamara, J. M. & Houston, A. I. (1985), 'Optimal foraging and learning', *Journal of Theoretical biology* **117**(2), 231–249.

McNay, E. C. & Gold, P. E. (2002), 'Food for thought: fluctuations in brain extracellular glucose provide insight into the mechanisms of memory modulation', *Behavioral and cognitive neuroscience reviews* **1**(4), 264–280.

Mery, F. & Kawecki, T. J. (2005), 'A cost of long-term memory in drosophila', *Science* **308**(5725), 1148–1148.

Min, K.-J., Flatt, T., Kulaots, I. & Tatar, M. (2007), 'Counting calories in drosophila diet restriction', *Experimental gerontology* **42**(3), 247–251.

Morand-Ferron, J. (2017), 'Why learn? the adaptive value of associative learning in wild populations', *Current opinion in behavioral sciences* **16**, 73–79.

Morimoto, J. (2019), 'Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data', *Journal of theoretical biology* **467**, 48–56.

Morimoto, J., Nguyen, B., Tabrizi, S. T., Ponton, F. & Taylor, P. (2018), 'Social and nutritional factors shape larval aggregation, foraging, and body mass in a polyphagous fly', *Scientific reports* **8**(1), 14750.

Müller, U. (2013), Memory phases and signaling cascades in honeybees, *in* 'Handbook of behavioral neuroscience', Vol. 22, Elsevier, pp. 433–441.

Murphy, K. P. (2007), 'Conjugate bayesian analysis of the gaussian distribution', *def* **1**($2\sigma 2$), 16.

Musso, P.-Y., Tchenio, P. & Preat, T. (2015), 'Delayed dopamine signaling of energy level builds appetitive long-term memory in drosophila', *Cell reports* **10**(7), 1023–1031.

Naito, A., Katahira, K. & Kameda, T. (2022), 'Insights about the common generative rule underlying an information foraging task can be facilitated via collective search', *Scientific Reports* **12**(1), 8047.

Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. (2012), 'Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain', *Journal of Neuroscience* **32**(2), 551–562.

Niven, J. E. & Laughlin, S. B. (2008), 'Energy limitation as a selective pressure on the evolution of sensory systems', *Journal of Experimental Biology* **211**(11), 1792–1804.

Olshausen, B. A. & Field, D. J. (2004), 'Sparse coding of sensory inputs', *Current opinion in neurobiology* **14**(4), 481–487.

Owald, D., Felsenberg, J., Talbot, C. B., Das, G., Perisse, E., Huetteroth, W. & Waddell, S. (2015), 'Activity of defined mushroom body output neurons underlies learned olfactory behavior in drosophila', *Neuron* **86**(2), 417–427.

Owald, D. & Waddell, S. (2015), 'Olfactory learning skews mushroom body output pathways to steer behavioral choice in drosophila', *Current opinion in neurobiology* **35**, 178–184.

O'Carroll, C. M., Martin, S. J., Sandin, J., Frenguelli, B. & Morris, R. G. (2006), 'Dopaminergic modulation of the persistence of one-trial hippocampus-dependent memory', *Learning & memory* **13**(6), 760–769.

Pearson, J. M., Watson, K. K. & Platt, M. L. (2014), 'Decision making: the neuroethological turn', *Neuron* **82**(5), 950–965.

Perisse, E., Owald, D., Barnstedt, O., Talbot, C. B., Huetteroth, W. & Waddell, S. (2016), 'Aversive learning and appetitive motivation toggle feed-forward inhibition in the drosophila mushroom body', *Neuron* **90**(5), 1086–1099.

Placais, P.-Y., de Tredern, É., Scheunemann, L., Trannoy, S., Goguel, V., Han, K.-A., Isabel, G. & Preat, T. (2017), 'Upregulated energy metabolism in the drosophila mushroom body is the trigger for long-term memory', *Nature Communications* **8**(1), 1–14.

Plaçais, P.-Y. & Preat, T. (2013), 'To favor survival under food shortage, the brain disables costly memory', *Science* **339**(6118), 440–442.

Potter, W. B., O'Riordan, K. J., Barnett, D., Osting, S. M., Wagoner, M., Burger, C. & Roopra, A. (2010), 'Metabolic regulation of neuronal plasticity by the energy sensor ampk', *PloS one* **5**(2), e8996.

Quinn, W. G., Harris, W. A. & Benzer, S. (1974), 'Conditioned behavior in drosophila melanogaster', *Proceedings of the National Academy of Sciences* **71**(3), 708–712.

Raiff, B. R. & Yoon, J. (2010), 'From bench to bedside: A review of impulsivity: The behavioral and neurological science of discounting', *Behavioural Processes* **2**(84), 632–633.

Raiffa, H., Schlaifer, R. et al. (1961), 'Applied statistical decision theory'.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z. et al. (2018), 'A tutorial on thompson sampling', *Foundations and Trends® in Machine Learning* **11**(1), 1–96.

Schönberg, T., Daw, N. D., Joel, D. & O'Doherty, J. P. (2007), 'Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making', *Journal of Neuroscience* **27**(47), 12860–12867.

Schultz, W. (2002), 'Getting formal with dopamine and reward', *Neuron* **36**(2), 241–263.

Schultz, W., Dayan, P. & Montague, P. R. (1997), 'A neural substrate of prediction and reward', *Science* **275**(5306), 1593–1599.

Slivkins, A. et al. (2019), 'Introduction to multi-armed bandits', *Foundations and Trends® in Machine Learning* **12**(1-2), 1–286.

Smid, H. M. & Vet, L. E. (2016), 'The complexity of learning, memory and neural processes in an evolutionary ecological context', *Current Opinion in Insect Science* **15**, 61–69.

Srivastava, V., Reverdy, P. & Leonard, N. E. (2013), On optimal foraging and multi-armed bandits, *in* '2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)', IEEE, pp. 494–499.

Staddon, J. E. R. (2016), *Adaptive behavior and learning*, Cambridge University Press.

Striedter, G. F. (2006), 'Précis of principles of brain evolution', *Behavioral and Brain Sciences* **29**(1), 1–12.

Sutton, R. S. & Barto, A. G. (2018), *Reinforcement learning: An introduction*, MIT press.

Suzuki, A., Stern, S. A., Bozdagi, O., Huntley, G. W., Walker, R. H., Magistretti, P. J. & Alberini, C. M. (2011), 'Astrocyte-neuron lactate transport is required for long-term memory formation', *Cell* **144**(5), 810–823.

Tempel, B. L., Bonini, N., Dawson, D. R. & Quinn, W. G. (1983), 'Reward learning in normal and mutant drosophila', *Proceedings of the National Academy of Sciences* **80**(5), 1482–1486.

Thompson, W. R. (1933), 'On the likelihood that one unknown probability exceeds another in view of the evidence of two samples', *Biometrika* **25**(3-4), 285–294.

Tomasi, D., Wang, G.-J. & Volkow, N. D. (2013), 'Energetic cost of brain functional connectivity', *Proceedings of the National Academy of Sciences* **110**(33), 13642–13647.

Trannoy, S., Redt-Clouet, C., Dura, J.-M. & Preat, T. (2011), 'Parallel processing of appetitive short-and long-term memories in drosophila', *Current Biology* **21**(19), 1647–1653.

Tsao, C.-H., Chen, C.-C., Lin, C.-H., Yang, H.-Y. & Lin, S. (2018), 'Drosophila mushroom bodies integrate hunger and satiety signals to control innate food-seeking behavior', *Elife* **7**, e35264.

Tully, T., Preat, T., Boynton, S. & Del Vecchio, M. (1994), 'Genetic dissection of consolidated memory in drosophila', *Cell* **79**(1), 35–47.

Tully, T. & Quinn, W. G. (1985), 'Classical conditioning and retention in normal and mutant drosophila melanogaster', *Journal of Comparative Physiology A* **157**(2), 263–277.

Turner, G. C., Bazhenov, M. & Laurent, G. (2008), 'Olfactory representations by drosophila mushroom body neurons', *Journal of neurophysiology* **99**(2), 734–746.

Waddell, S. (2013), 'Reinforcement signalling in drosophila; dopamine does it all after all', *Current opinion in neurobiology* **23**(3), 324–329.

Wang, Z., Wei, X.-X., Stocker, A. A. & Lee, D. D. (2016), Efficient neural codes under metabolic constraints., *in* 'NIPS', pp. 4619–4627.

Whitley, D., Rana, S., Dzubera, J. & Mathias, K. E. (1996), 'Evaluating evolutionary algorithms', *Artificial intelligence* **85**(1-2), 245–276.

Wieraszko, A. (1982), 'Changes in the hippocampal slices energy metabolism following stimulation and long-term potentiation of schaffer collaterals-pyramidal cell synapses tested with the 2-deoxyglucose technique', *Brain research* **237**(2), 449–457.

Wolff, G. H. & Strausfeld, N. J. (2015), 'Genealogical correspondence of mushroom bodies across invertebrate phyla', *Current Biology* **25**(1), 38–44.

Wright, G. A. (2011), 'Appetitive learning: memories need calories', *Current Biology* **21**(9), R301–R302.

Yamagata, N., Ichinose, T., Aso, Y., Plaçais, P.-Y., Friedrich, A. B., Sima, R. J., Preat, T., Rubin, G. M. & Tanimoto, H. (2015), 'Distinct dopamine neurons mediate reward signals for short-and long-term memories', *Proceedings of the National Academy of Sciences* **112**(2), 578–583.

Yu, Y., Karbowski, J., Sachdev, R. N. & Feng, J. (2014), 'Effect of temperature and glia in brain size enlargement and origin of allometric body-brain size scaling in vertebrates', *BMC evolutionary biology* **14**(1), 1–14.

Zolin, A., Cohn, R., Pang, R., Siliciano, A. F., Fairhall, A. L. & Ruta, V. (2021), 'Context-dependent representations of movement in drosophila dopaminergic reinforcement pathways', *Nature neuroscience* **24**(11), 1555–1566.