Models of aposematism and the role of aversive learning.

Author:
Jan Teichmann

Supervisors:
Prof. Mark Broom
Dr. Eduardo Alonso

Submitted in accordance with the requirements for the degree of Doctor of Philosophy.

City University London

Department of Mathematical Science

June 2015

Contents

List of Figures

Acknowledgements.

This thesis is the result of a three year long journey. As most of the PhDs, I presume, this journey was not always plain sailing and I would like to express my greatest gratitude to my two supervisors Mark Broom and Eduardo Alonso who gave me crucial guidance and support during my doctoral journey.

I am also very grateful to have received a PhD studentship from the Department of Mathematics and for the generous travel bursaries by the Graduate School and the City Future Fund which allowed me to attend inspiring conferences.

Last but not least, I would like to thank my family and friends who have been at the front line when code did not compile, manuscripts were rejected, or ideas hit dead ends.

Declarations.

Parts of this thesis have been published (or submitted for publication) in the following articles which I have co-authored with others:

Jan Teichmann, Mark Broom and Eduardo Alonso. (submitted). When does learning matter: the relationship between Evolution and Reinforcement Learning.

Jan Teichmann, Mark Broom and Eduardo Alonso. The Evolutionarily Dynamics of Aposematism: a Numerical Analysis of Co-Evolution in Finite Populations. *Mathematical Modelling of Natural Phenomena*, **9**, 148–164 (2014).

Jan Teichmann, Mark Broom and Eduardo Alonso. The application of temporal difference learning in optimal diet models. *Journal of Theoretical Biology*, **340**, 11–16 (2014).

My co-authors gave me important feedback in the development of my work and during the editing of my manuscripts. Dr. Mike Speed provided me with input and comments in the revision of the paper "The Evolutionarily Dynamics of Aposematism: a Numerical Analysis of Co-Evolution in Finite Populations.". However, in each case, the scientific work presented is my own.

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

I acknowledge that the copyright of published articles contained within this thesis resides with the copyright holder of those works.

Abstract

The majority of species are under predatory risk in their natural habitat and targeted by predators as part of the food web. Through the process of evolution by natural selection manifold mechanisms have emerged to avoid predation. As Fisher argued, it is the ubiquitous presence of anti-predator adaptations which shows that predation plays a significant role in the ecology and evolution of ecosystems. These ecosystems are intrinsically complex which derives from the high entanglement of organisms interacting in competitive relationships: the prey is part of the predator's environment and vice versa. As a result, the evolution of predator and prey is best described as a co-evolutionary process of predator-prey systems. It is common to classify anti-predator adaptations into 'primary defences' and 'secondary defences'. Primary defences operate before an attack by reducing the frequency of detection or encounter with predators. Secondary defences, which are used after a predator has initiated prey-catching behaviour, commonly involve the expression of toxins or deterrent substances which are not observable by the predator. Hence, the possession of such secondary defence in many prey species comes with a specific signal of that defence. This pairing of a toxic secondary defence and a conspicuous primary defence is known as *aposematism*. Previous models mainly focused on questions of the initial evolution of aposematism in ancestrally cryptic populations. However, the field has a renewed interest in questions beyond the initial evolution of aposematism such as: how conspicuous should a signal be, and how much should be invested into secondary defence? Moreover, which factors influence evolutionary stability of aposematic solutions. Within this context, the role of co-evolution and the mechanisms of aversive learning are at the heart of the current research. On the one hand, to explain stability and persistence of aposematic signals requires a theory of co-evolution of defence and signals. On the other hand, the role of the predator and details of the predator's aversive learning process gained renewed interest of the field. As the selective agent, aversive learning is an important aspect of predator avoidance and of the co-evolution of predator-prey systems. In the first chapter, this thesis will review the literature on aposematism and introduce the different selective pressures acting on aposematic prey. The thesis will then identify open questions of interest around aposematism. In the second chapter the thesis will focus on the perspective of the prey. The introduction of a game theoretical model of co-evolution of defence and signal will be followed by an adaptation of the model for finite populations. In finite populations, investigating the co-evolution of defence and signalling requires an understanding of natural selection as well as an assessment of the effects of drift as an additional force acting on stability. In the third chapter the thesis will adopt the perspective of the predator. It will introduce reinforcement learning as an normative framework of rational decision making in a changing environment. An analysis of the consequences of aposematism in combination with aversive learning on the predator's diet and energy intake will be followed by a lifetime model of optimal foraging behaviour in the presence of aposematic prey in the fourth chapter. In the last chapter I will conclude that the predator's aversive learning process plays a crucial role in the form and stability of aposematism. The introduction of temporal difference learning allows for a better understanding of the specific details of the predator's role in aposematism and presents a way to take the discipline forward.

Chapter 1

Introduction.

This chapter will provide a general introduction to the biology of predator-prey systems: I will start with the evolution of predator and prey and their consequential co-evolution in predator-prey systems. The focus will then move to mechanisms to avoid attack and how it results in a complex signalling system called *aposematism*. I will present the scientific background to the emergence of aposematism and lay out the open questions which motivate the following chapters.

1.1 Predator-prey systems.

There are three fundamental principals which define any biological system. These principals are: reproduction, selection, and mutation.

It has been long recognised that life comes in a profuse variety of shapes, forms, and traits. But as vast as the differences are there are underlying similarities which apparently connect all living things. These similarities are evidence for their descent from common ancestors.

Modern evolutionary theory is the aggregate of many subsidiary ideas from anticipation of nature to genuine interpretation of nature (Osborn, 1896). The notion of joining diversity with similarity in an attempt to explain biological systems goes all the way back to ancient Greek philosophy. In a time when the world was considered to be static Anaximander suggested a dynamic and changing world and is considered as evolution's most ancient proponent. The next great progress towards an inductive theory of evolution based on laws of nature was by Lamarck's *transmutation* theory of 1809. Lamarck proposed that organisms adapt to their local environment through inherited changes over generational time. But Lamarck did not provide a workable mechanism with his theory and is better known today for his flawed principle of inheritance of

traits acquired by the parental generation from use or disuse. The heritability of acquired characteristics was later called *Lamarckism*. Furthermore, Lamarck's transmutation theory was not able to change the prevailing concept of fixed species as it was widely opposed for its lack of empirical evidence.

1.1.1 Darwinism and evolution by natural selection.

The situation changed dramatically with Charles Darwin. Charles Darwin was a naturalist and sailed on board the ship Beagle around the world where he collected and documented flora and fauna. It was his vast collection which made him recognise the differences and similarities of species on an unprecedented scale. He concluded his realisations with the writing of his book 'On the origin of species'. Darwin's main achievement was to present a workable mechanism based on empirical evidence: his theory has the same notion of natural evolution. But in a crucially important difference, he proposed that the observed variation in traits is innate and not acquired. Thus, the variation in traits results in an unequal adaptation to the environment and, consequently, some organisms will survive and reproduce more successfully. It has to be noted that Alfred Russell Wallace formulated independently an almost identical theory.

In summary, the central points of modern evolutionary theory are: (i) organisms having innate variations in their traits and characteristics, (ii) well adapted organisms are more likely to survive and reproduce, which (iii) leads to better adapted organisms.

The focus of current research lays on the search for different factors of the natural law called evolution. Interesting aspects of evolution are for example the effects of sexual selection or the evolution in more complicated environments where we observe the effects of interactions between species and organisms.

1.1.2 Co-evolution.

As a matter of fact natural environments are intrinsically complex. This complexity derives on the one hand from the high entanglement of organisms interacting in competitive relationships with each other. On the other hand, natural environments are also defined by their dynamics of constant change. Thus, evolution in natural environments is defined by the dynamic competitive relationships of organisms. Typically, evolution results in multiple species which successively adapt in response to their adaptations. As an example, predators and prey evolve together as the prey is part of the predator's environment and vice versa: the predators rely on their prey as a food source and evolve necessary traits in order to feed on their prey efficiently. Common traits found in predators are therefore speed, a good sense of sight, hearing, and smell, or

specifically adapted mouthparts. Likewise, the prey evolves means to avoid predation such as speed, crypsis, deterrents, and good senses to detect predators. This phenomena is called *co-evolution* (Janzen, 1980):

> Co-evolution may be usefully defined as an evolutionary change in a trait of the individuals in one population in response to a trait of the individuals of a second population, followed by an evolutionary response by the second population to the change in the first.

1.1.3 Evolutionary game theory.

A modern framework to describe and analyse evolutionary models is in the form of a *game*. In a game the environment is modelled based on individuals, strategies, and payoffs. The payoff an individual receives from taking part in the game depends on the individual's strategy but equally on the strategy of all other individuals taking part in the game. The aim in such a game is to reason about the optimal strategy under the assumption that all individuals taking part in the game are rational and try to maximise their individual payoffs. As the optimal strategy of a single individual in such a game depends reciprocally on all the other individual strategies the task of defining optimality seems almost impossible at first. The modern field of *game theory* holds a wide body of methodologies and frameworks which addresses optimality in games with the Nash equilibrium probably being the most fundamental concept (Nash, 1951). The Nash equilibrium defines a set of optimal strategies in a game of multiple players which do not cooperate and individually maximise their payoffs. Within such a set of strategies no individual can gain any improvement of their own payoff by independently changing their strategy. In the special case of a stable Nash equilibrium changes to an individual's strategy do not impact the optimality of the other individuals strategies either.

The notion of *evolutionary game theory* applies the game theoretical framework to biology where individuals might not reason about their optimal behaviour but show different forms of strategies or adaptations due to their genetic variations. The payoff within evolutionary games is represented by fitness gains or losses with natural selection being the driving force of optimising strategies or adaptations in the evolutionary games. Additionally, the underlying population dynamics of interacting individuals can be ignored as being a separated layer to the evolutionary model. This allows a focus on static games which does not alter the outcome of the analysis in most cases within the biological context: finding the best strategy or adaptation for a specific environment represented by the evolutionary game.

An optimal strategy within a biological context is termed an *evolutionarily*

stable strategy or ESS and is closely related to the stable Nash equilibrium. An ESS is defined as the strategy (S) that no different/ mutant strategy (T) can invade (assuming that a population is playing the ESS uniformly and within the influence of natural selection) (Maynard Smith, 1974) which is the case if and only if one of the following conditions hold:

$$\begin{aligned} E(S,S) &> E(T,S), \text{ or} \\ E(S,S) = E(T,S) &\wedge E(S,T) > E(T,T) \quad \forall\, T \neq S. \end{aligned} \tag{1.1}$$

This has been a simplistic and brief introduction to evolutionary game theory only. The next chapter will discuss models of defended prey and signalling that defence to their predators. Such models use continuous strategies and fall into the category of non-linear games. I refer to the next chapter for a detailed discussion of the ESS in non-linear games with continuous strategies and effects of finite populations using a method called adaptive dynamics.

1.2 How to avoid attack.

Actually, the vast majority of species are under predatory risk in their natural habitat and targeted by predators as part of the food web. Through the process of evolution by natural selection manifold mechanisms have emerged to avoid predation. As Fisher (1930) argued, it is the ubiquitous presence of anti-predator adaptations which shows that predation plays a significant role in the ecology and evolution of ecosystems.

1.2.1 Primary and secondary defences.

It is common to classify anti-predator adaptations into 'primary defences' and 'secondary defences'. By definition, primary defences operate before an attack by reducing the frequency of detection with predators, as in disruptive colouration, countershading, and crypsis, or by reducing the risk of attack given detection, for example by warning colouration, morphological adaptations, chemical defences, mimicry, and aggregation (Robinson, 1969; Edmunds, 1974; Ruxton et al., 2004). Complementarily, secondary defences such as toxins or unpalatable substances reduce the risk of falling prey in an encounter with predators. This general classification, however, is not without limitations as there are interesting grey areas when it comes to warning signals or mimicry, for example. Importantly, anti-predator adaptations are not discrete and independent traits but continuous and interacting.

A common alternative is to categorise mechanisms of anti-predator adaptations into functional groups of

- avoiding detection (which includes avoiding encounters),
- avoiding attack (or falling prey in an attack), and
- deceiving predators.

Given that predation plays a significant role in natural selection, avoiding detection by predators seems like a strategy most obviously favoured by evolution through natural selection. In fact, crypsis is a widely found adaptation to prevent detection by predators. But the broader question of 'Why are species selected for a specific form of defence over another?' is an important starting point for the discussion of conspicuous warning signals in the next chapter. It will be crucial for our understanding of evolutionary dynamics to define defence as involving some kind of 'cost-benefit' trade-off.

Tollrian and Harvell (1999) proposed a framework of five general categories for analysing fitness cost in secondary defences which has been generally adopted for anti-predator adaptations:

Allocation cost or internal cost which arises from allocating limited resources to the erection, maintenance, and operation of a defence.

Environmental cost or external cost which arise from interactions with the environment in relation to the defence.

Opportunity cost or indirect cost which arises from missed alternatives caused by the defence. Adaptations such as crypsis or seasonal behaviour can limit an individuals options in respect to foraging or mating.

Design cost or self-damage cost arises typically in the context of chemical defences to prevent auto-toxicity. But the design is not limited to chemical defences alone.

Plasticity cost relates to inducible defences which allow an individual to respond to changing predatory risk in its environment. The cost arises from the deployment of sensory systems for example.

Concluding, most defences, if not all, incur some kind of fitness cost or other trade-offs. Thus defences require clear benefits which outweigh their costs where the reduced predation is not necessarily the only advantage. I will discuss further benefits in regard to overcoming crypsis in the initial evolution of aposematism.

1.2.2 Aposematism.

On the one hand, the benefits of primary defences to prevent detection are evident, especially, when we assume the risk of predation to be high. On the

Figure 1.1: Yellow-banded poison dart frog, Dendrobates leucomelas, an example of aposematic display. (picture in public domain taken from Wikimedia commons)

other hand, prey may have secondary defences which commonly involve the expression of toxins or deterrent substances. These secondary defences are not directly observable by the predator and the benefits of such defences are not self-explanatory. A predator might attack the defended prey nevertheless for a lack of aversive information.

Hence, many defended prey species use conspicuous signals – either visual, audible or behavioural – in combination with their otherwise non-observable secondary defences to warn predators. This pairing of a toxic secondary defence and a conspicuous warning signal is known as *aposematism* (Poulton, 1890). The most commonly associated warning signal is 'warning colouration' but other signals are known such as conspicuous sounds, behaviours, and odours. An example is the family of poison dart frogs, *Dendrobatidae*, which are native to Central and South America. The species have brightly coloured skin and are at least to some degree toxic (Figure 1.1).

Aposematism is a primary defence and the benefit of avoiding well-defended prey seems to be mutual and obvious. Additionally, signalling is omnipresent and fundamental within biological systems which might make it appear as trivial. But this is far from the truth and aposematism has been the focus of much research by the scientific community in the light of evolutionary theory.

In particular, the initial evolution of these warning signals in ancestrally cryptic populations has been much debated because a novel conspicuous mutant has to overcome the loss of protection of crypsis, which is maintained by its conspecifics. Furthermore, anti-apostatic selection by inexperienced predators results in rare mutants being predated relatively more often (Lindström et al., 2001).

Previous models mainly focused on the questions of the initial evolution of aposematism and there are two established arguments which aid the appearance of conspicuous signals:

1. The Predator's perception and cognitive processes possess specific properties which promote aposematism. For example the usage of aposematic signals as warning flags improves discrimination in educated predators and enhances the learning of unprofitability (Keehn, 1959). Other possible factors could be dietary conservatism (Lee et al., 2010; Thomas et al., 2003) or a shifted peak of the aversive information so that more conspicuous prey individuals are favoured (Leimar et al., 1986; Yachi and Higashi, 1998; Gamberale-Stille and Tullberg, 1996).

2. The opportunity cost of crypsis in combination with the reliability of honest warning signals of well-defended prey drove the evolution of aposematism and consequently the evolution of predator psychology (Sherratt, 2002).

Other more general factors which can further aid the initial evolution of aposematism are spatial aggregation and kin selection. Both factors are widely applied to introduce a more significant number of mutants in theoretical models to overcome the problems of the initial evolution of new traits. I will discuss these factors in more detail in the next chapter where they find application in a model of co-evolution.

A second major issue in the theoretical treatment of aposematism is the problem of dishonesty: it may be beneficial for an undefended individual to use a warning signal too to avoid being attacked (*mimicry*), which in turn can undermine the effectiveness of the signal. When this kind of dishonesty occurs across species, it is known as *Batesian mimicry*; when it happens within species, it is known as *automimicry* (Ruxton et al., 2004). A second form of dishonesty arises when there is continuous variation in toxicity within or between species. In some theoretical treatments, prey with weak secondary defences may choose to signal brightly, in order to compensate for their lack of repellence to predators. Hence key questions of current importance in aposematism theory focus on the questions: how conspicuous should a signal be, and how much should be invested into secondary defences (Speed and Ruxton, 2007; Speed

et al., 2010; Longson and Joss, 2006; Ruxton et al., 2009)? Moreover, which factors influence evolutionary stability of honest signalling and what is the role of mimics in maintaining or destabilising aposematic display (Gamberale-Stille and Guilford, 2004)?

Leimar et al. (1986): Mechanisms of aversive learning in aposematism.

Although many aspects of signalling systems are understood, a key element missing from the current theory is the incorporation of learning such as the role of aversive learning in particular. The importance of this was described in an earlier work by Leimar et al. (1986) which is still one of the most relevant frameworks for the evolution of aposematism.

The key contribution of Leimar et al. are elements of predator psychology contributing to a process of aversive learning in aposematism which could explain the evolution of aposematism assuming that secondary defences involve some cost.

The model defines an inhibitory gradient h which generalises the aversive experience from encounters with n prey individuals of a specific morph (x_i, y_i) to generalised attack probabilities $g(x)$:

$$g(x) = e(x) \prod_i [1 - h(x, x_i, y_i)]^{n_i}, \tag{1.2}$$

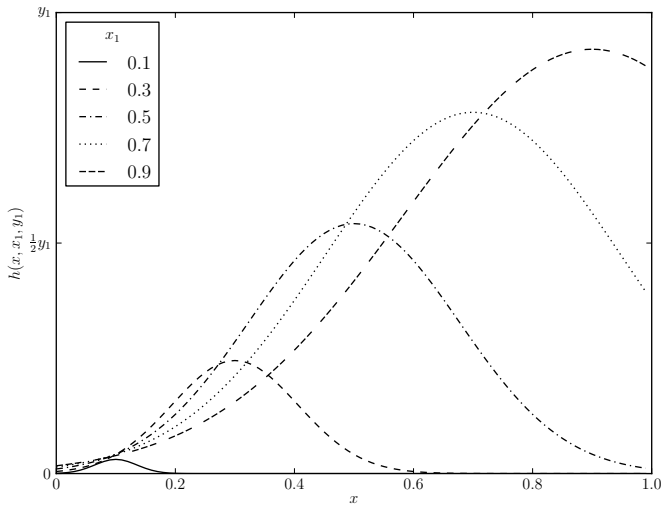
with $e(x)$ being the excitatory gradient of a naive predator, x being the colouration, and y being the degree of unprofitability of prey. The specific nature of the generalisation gradient promotes conspicuous prey through the application of a peak-shift (Figure 1.2 as defined in Leimar et al., 1986). The peak-shift is a psychological phenomenon which results in a bias towards avoidance of more conspicuous prey following an aversive encounter. There is a growing body of empirical evidence which supports the assumption of biased generalisation, and it might play an important role in the stability and initial evolution of aposematism (Yachi and Higashi, 1998).

Other main findings by Leimar et al. (1986) were that aposematism can initially evolve in an otherwise cryptic prey population if there are some elements of kin-selection or a change in environment making crypsis less effective.

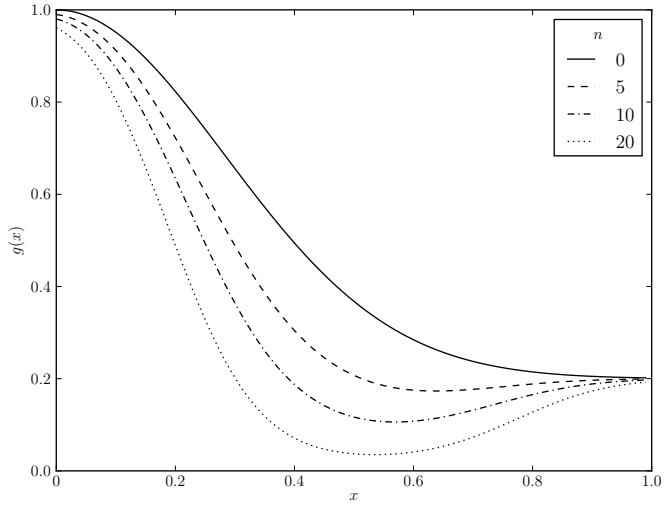
I will build on this ideas in the following chapters of this thesis.

Sherratt (2002): The co-evolution of aposematism.

Even though the peak-shift phenomenon promotes the initial evolution of aposematism it requires a pre-existing and universal psychological response to con-



(a) This chart shows the generalisation of aversive information from an encounter with a specific morph (x_1, y_1) through the inhibitory gradient $h(x, x_1, y_1)$.



(b) This chart shows the attack probability $g(x)$ after n encounters with prey of morph $x_1 = 0.5$ and $y_1 = 0.2$. The graph for $n = 0$ shows the excitatory gradient $e(x)$ of a naive predator. The graph for $n > 0$ show the peak-shift resulting in a bias towards the avoidance of more conspicuous prey.

Figure 1.2: Elements of predator psychology describing a process of aversive learning as defined in Eq. 1.2 following the definition in Leimar et al. (1986) with x being the colouration and y being the degree of unprofitability of prey.

spicuousness. The question of 'Which came first: conspicuous signals or specific properties of predator psychology?' has not been resolved. Furthermore, the generalisation bias itself does not explain why predators show this particular psychological property. Most importantly, aposematism occurs within many different species, over a wide range of taxa, and conspicuous signals manifest themselves over a diverse set of sensory systems. It is unlikely that all this can be explained by a common pre-existing generalisation bias. It is perhaps more feasible to assume that the generalisation bias has arisen from a common selective pressure.

Sherratt (2002) addresses these questions by introducing a model of co-evolution of multiple predators and prey. The difference to previous models is that the foraging behaviour of predators is subject to selection itself. This allows aposematism to evolve through a process of co-evolution of predator and prey.

Sherratt concludes that a novel conspicuous prey item has most likely been encountered by previous naive predators as it is easily detected. For this reason, novel conspicuous prey is likely to be defended to have survived previous encounters with predators.

Likewise, a naive predator which moves into an environment with experienced predators should avoid conspicuous prey as it is easily detected by the other predators and is likely to be defended to have risen to greater numbers.

As soon as the predators have adapted to the correlation of defences and conspicuous signals aposematism arises rapidly, rather than gradually, through runaway co-evolution.

I will build especially on the idea of co-evolution in the following chapter but Sherratt identifies further factors in his model which aid the initial association of defences with conspicuous signals such as the opportunity cost of crypsis and the aggregation of prey.

Speed and Ruxton (2007): The vast variety of aposematic solutions.

We saw that the initial evolution of aposematism itself already presents the scientific community with a wide range of challenges. That is why it might be not surprising that most of the theoretical work is concerned about the factors which drove the initial evolution of aposematism. Nevertheless, recent efforts have tried to explain the evolutionary stability of aposematism in the light of growing empirical studies. The interest in evolutionary stability itself is not new, with Leimar et al. (1986) already discussing evolutionarily stable strategies (ESS) within their framework. The reason for the later neglect of questions regarding ESS in many models was mainly due to the complexity of

the task at hand: empirical studies of aposematic species show a vast variety of aposematic solutions. A broad study by Summers and Clough (2001) showed a positive correlation between conspicuousness and toxicity in the poison dart frog family, *Dendrobatidae*, which supports the ideas of aposematism being a costly handicap signal indicating fitness advantages. But when Darst et al. (2006) revisited three specific dart frog species they actually found the reverse case of negatively correlated toxicity and a established theoretical framework to treat these findings consistently did not exist. The next chapter will present recent work which addresses evolutionary stability within a theoretical framework of co-evolution to fill this gap (Broom et al., 2006; Broom et al., 2008; Teichmann et al., 2014b).

In order to explain the variety of aposematic solutions Speed and Ruxton (2007) introduce a model incorporating marginal costs of both display ψ and secondary defences ζ. The model predicts optimal values of conspicuousness C and defence D of a focal prey population in an environment of multiple predators and other cryptic and undefended non-focal prey populations.

The probability of attack given detection $P(Att)$ is a combination of inherent wariness regarding conspicuousness $W(C)$ (see Section 1.2.2) with repellence R from previous encounters with other defended prey individuals represented by the average toxicity of the focal group D^*:

$$P_i(Att) = Att_{\min}(1 - Att_{\min})W(C_i)R(D^*). \tag{1.3}$$

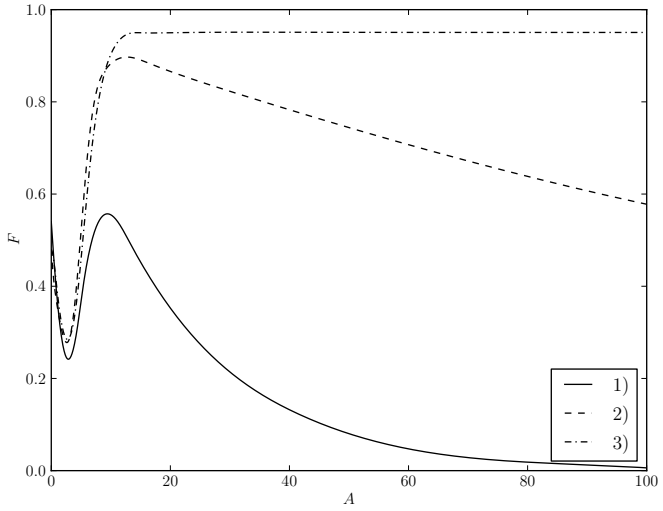
The probability of being killed in an attack $P(Kill)$ is derived from the individual's level of defence D and the conspicuousness of its individual display A:

$$P_i(Kill) = Kill_{\min}(1 - Kill_{\min})K(A_iD_i). \tag{1.4}$$

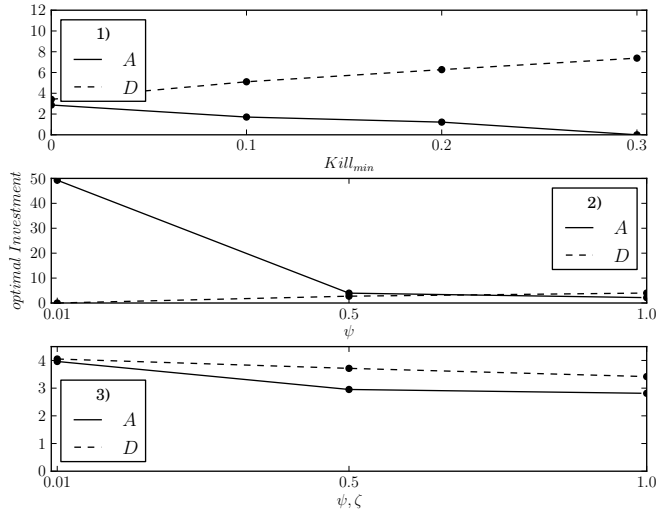
The specific functions used by Speed and Ruxton (2007) are Gompertz functions. After the end of a season the fitness of surviving prey is calculated considering the fecundity costs of secondary defences and display:

$$F_i = exp(-(\psi A_i + \zeta D_i)). \tag{1.5}$$

Figure 1.3 presents the main results of the model: the variety of aposematic solutions can be explained with the variation of the marginal cost of displays for different prey species (Figure 1.3a). Additionally, the model predicts positive as well as negative correlation of aposematic display A and defence D (Figure 1.3b).



(a) This chart shows the variety of aposematic solutions as a result of the marginal cost of display ψ. 1) $\psi = 1$, 2) $\psi = 0.1$, and 3) $\psi = 0$. All other parameters as in Speed and Ruxton (2007).



(b) This chart shows examples of correlation between the optimal investment into defence D and display A. 1) shows negative correlation in dependency of the effectiveness of defence in surviving an attack $Kill_{\min}$, 2) shows negative correlation in dependency of marginal cost of display ψ with constant marginal cost of defence ζ, and 3) shows positive correlation if ψ and ζ vary linearly at the same time. All other parameters as in Speed and Ruxton (2007).

Figure 1.3: The variety of aposematic solutions as described by Speed and Ruxton (2007).

1.3 Problem formulation.

As it turns out, the aspects of aposematism are manifold and far from straightforward. The main interest in the past laid in the initial evolution of aposematism and there is a wide body of theories available today. Nevertheless, the nature of aposematism beyond the initial evolution has still many open questions and remains a challenging research area. I have presented seminal models which have made significant progress on these exciting and open questions around aposematism and have influenced the field greatly by laying the foundations of much research which followed. In particular, the importance of the role of aversive learning was described in the earlier work by Leimar et al. (1986), but was omitted from later modelling developments in order to simplify analytical tractability. However, interesting questions arise as to how a predator incorporates the information gained from an encounter with prey into a generalised approach to predation and defence. In summary, the main aspects of aposematism the field has a great interest in are:

- the properties of aposematic solutions and their stability,
- the role of co-evolution,
- the influence of predator psychology in particular aversive learning, and
- the consequences of aposematism for predator and prey populations in regards to their fitness and foraging behaviour.

1.4 Thesis Aims.

The aim of this project is to develop models of the signalling of invisible defences beyond the initial evolution of aposematism.

In chapter 2 I will focus on aposematism from the prey perspective based upon the model by Broom et al. (2006), which presents the first explicit mathematical model of a relationship between the conspicuousness of aposematic signals and the strength of the defence that they advertise. My analysis will address current research questions of the field around co-evolution of aposematism and stability in finite populations in the presence of drift. The chapter will point out the importance of the predator's role within aposematism.

In chapter 3 I will develop theories of aposematism from the predator perspective. The motivating questions revolve around how the predator incorporates aversive encounters with aposematic prey into generalised foraging behaviour. Recent studies by Alonso and colleagues (Alonso and Schmajuk, 2012; Alonso and Mondragón, 2006; Alonso et al., 2001) have looked at the concept of

learning in a more general setting, investigating learning algorithms that allow for the transfer of acquired knowledge between stimuli that share certain features (e.g., intensity or modality). The idea is to introduce stimulus-action-outcome associations in the form of aversive learning so that predators are able to generalise their experience from encounters with aposematic prey to the aversiveness of aposematic prey populations. Building on the current theories of generalisation and discrimination in the form of aversive learning I will discuss the effects of aposematic prey on the fitness and energy intake of a predator.

Chapter 4 will develop a predator lifetime model which incorporates life history traits which have been abstracted away in the previous chapter. I will compare results of an individual based foraging simulator driven by reward motivated objectives with a generalisation of behavioural repertoires driven by fitness. The model will address questions concerning what optimal behaviour is and why there might be a discrepancy between maximising rewards and maximising fitness.

I will summarise my findings in chapter 5 and draw conclusions of how the necessity of learning to avoid certain defended prey affects the characteristics of aposematic solutions.

1.5 Thesis layout.

This thesis is a multidisciplinary excursion into the methodologies of different fields. Usually, theses have a clear cut-off point which distinguishes between previous work by others and the new contributions. However, in this thesis I decided to include previous work by others within the flow of the discussion as I explore the different methodologies, such as evolutionary game theory, optimal foraging theory, reinforcement learning, and the psychology of rewards amongst others.

To allow a clear identification of my contributions to the current state of the field of aposematism, I provide a list of the thesis layout, as follows:

- Chapter 1: A literature review and introduction to the field of aposematism.
- Section 2.1: The review of a game theoretical model of coevolution by Broom et al. (2006).
- Section 2.2: An extension of the previous model describing the evolutionary dynamics of aposematism: a numerical analysis of co-evolution in finite populations.

- Section 3.1: An introduction to the field and methodology of Reinforcement Learning.
- Section 3.2: A new model for the application of Q-learning in optimal diet models with Section 3.2.1 giving a general introduction to the field of optimal foraging theory.
- Section 3.3: When does learning matter? A new model investigating the relationship between Evolution and Reinforcement Learning.
- Section 4.1: A general introduction and review of the psychological elements of learning and rewards.
- Section 4.2: The definition of a new predator lifetime model.
- Section 4.3: The introduction of a new learning based foraging simulator which builds on methodologies which are introduced and reviewed in Sections 4.3.1, 4.3.2, and 4.3.3.
- Section 4.3.4: The application of Reinforcement learning in the new predator lifetime model and the presentation of the results in Section 4.3.5.
- Section 4.4: The interpretation of the new lifetime model with regard to Darwinian fitness and the discussion in Section 4.5.

Chapter 2

Aposematism from the preys' perspective.

In this chapter I will focus on questions around aposematism which relate to the prey. As I laid out in the introduction the main focus of previous models has been the emergence of aposematism. Mechanisms which were identified in aiding the emergence of aposematism are the opportunity cost of crypsis, the improved discrimination of prey in educated predators through warning flags, dietary conservatism, and peak shifted aversiveness functions. The following model will expand the theoretical frameworks incorporating co-evolution of defence and signalling of that defence. Parts of this chapter have been published in Teichmann et al. (2014b).

2.1 A game theoretical model of co-evolution by Broom et al.

Broom et al. (2006) introduced a game theoretic model of prey-predator interaction to describe the co-evolution of secondary defence and signalling. The model investigates the general mechanisms of aposematism rather than specific species or environments, assuming general function shapes. Building on that, the model was further developed in Broom et al. (2008) using exemplary and plausible functions (Table 2.1) to demonstrate the solutions predicted from Broom et al. (2006). (See the following Section 2.2 for a modified version of the framework using specific functions.) The model from Broom et al. (2008) considers a single population of prey individuals where individuals i are described by two parameters (r_i, t_i).

The parameter t reflects the individual investment into secondary defence.

Symbol	Meaning
r	the conspicuousness of an aposematic signal
t	the level of toxicity of secondary defences
$F(t)$	the fertility of an individual of toxicity t
$K(t)$	the probability that an individual of toxicity t is killed in an attack
$H(t)$	the aversiveness of an individual of toxicity t
$S(x)$	the similarity function of individuals differing in appearance by x with $x = \|r_1 - r_2\|$
$I(r)$	the level of aversive information of an individual
$D(r)$	the rate at which individuals of conspicuousness r are detected
$Q(I)$	the probability that a predator will attack an individual associated with a level of aversive information I
a	the fraction of mutants in the population
t_c	the level of toxicity which becomes aversive, hence for which $H(t_c) = 0$

Table 2.1: Exemplary functions as introduced in (Broom et al., 2006; Broom et al., 2008).

This secondary defence is not observable by the predator and could be unpalatable toxins, for example. The expression of secondary defence comes with a cost of decreasing fecundity $F(t)$. On the other hand, secondary defence is advantageous in surviving an attack, reducing the chance of being killed $K(t)$.

The parameter r describes the conspicuousness of an aposematic signal, with $r = 0$ referring to maximal crypsis. The quality of signalling is associated with an unfavourable higher rate at which individuals are detected by predators $D(r)$. In contrast, conspicuousness of an aposematic signal is beneficial in combination with secondary defences by increasing the predator's level of aversive information regarding an individual $I(r)$.

Unlike the model of (Leimar et al., 1986) where naïve predators without any experience from encounters with prey individuals learn to avoid unpalatable prey types, the model of (Broom et al., 2006) is assumed to be in equilibrium: population sizes are constant and predators are experienced and have full knowledge of the population's aversiveness and signalling strategies. This knowledge is expressed in a level of aversive information I about individuals which depends upon the appearance of the individual in question, and the properties of the population of prey individuals.

A prey individual contributes to the level of aversiveness about others through a combination of three factors. Firstly the likelihood of it being encountered, which is proportional to the detection probability $D(r)$ described above. Secondly the aversive effect of its consumption, which depends upon its value of t through the function $H(t)$, which is increasing with t. There is a critical value of defence t_c for which $H(t_c) = 0$; this corresponds to levels of defence above

this being aversive, and levels below this being actually beneficial, encouraging the consumption of more prey of this type.

Finally, this individual will affect the response of the predator to others only if it is sufficiently similar to them, which is indicated by a general similarity function $S(x)$, where $x = |r - r_j|$ is the difference between the r value (r_j) of the above individual and that of any given targeted individual. If x is large, then the two are very different and the similarity function takes a very small value. For $x = 0$ they are identical, and the similarity function is set to equal 1. The peaked form of the similarity function (where the function is differentiable w.r.t r everywhere but at the peak, where the left-derivative and right-derivative are different) is characteristic of this model and responsible for the resulting broad range of alternate ESS.

The information from a single population individual from the population with parameters (r_j, t_j) about our focal individual with parameters (r_i, t_i) is thus proportional to $D(r_j)H(t_j)S(|r_i - r_j|)$. The total aversive information from the population is the sum of this over all individuals (in the original work this was multiplied by the ratio of prey individuals N to predators n, but this is simply a scaling factor), giving

$$I_i = \sum_{j=1, j\neq i}^{N} D(r_j)H(t_j)S(|r_i - r_j|). \tag{2.1}$$

The predator then uses the information about individual i to decide whether to attack it if it is encountered, choosing the attack probability $Q(I)$, which is decreasing in I, based on the amount of aversive information I_i.

The payoff to an individual, which is simply the ratio of its fecundity $F(t_i)$ and its rate of being killed, is described as follows:

$$Z_i = \frac{F(t_i)}{D(r_i)Q(I_i)K(t_i) + \lambda} \ . \tag{2.2}$$

The original model adds a constant λ to the denominator to represent death due to events other than predation, but in some of the analysis in that model (and in my model to follow) this term is set to zero.

The evolution of secondary defences is promoted by an increased inclusive fitness through both the greater chances of escape of individuals, and the reduction of the likelihood that educated predators re-attack prey individuals or their relatives in the future. Modelling the evolution of signalling and defence using kin grouping was introduced in (Leimar et al., 1986), destabilising crypsis in favour of aposematism. The parameter a in the model of (Broom et al., 2006) describes the initial protection of mutants from predation: mutants occur in

groups with local proportion a either as a consequence of first appearing in a self-contained locality or through invasion.

The main findings of the underlying model of (Broom et al., 2006) were:

- Crypsis can be destabilised without the assumption of a naïve tendency of avoiding suspicious prey individuals or the usage of a peak-shifted aversiveness towards more suspicious individuals by co-evolution.
- The strength of aposematic displays is a reliable indicator of the strength of defence, that is, the correlation between the two is positive.
- If the conditions support the emergence of aposematism there are multiple stable solutions laying on an increasing line $t_{\mathrm{opt}}(r)$. Hence, the diversity of different solutions for secondary defence and warning displays is a consequence of the underlying co-evolution.
- Aposematism is not a necessary condition for optimality of highly defended prey populations and stable cryptic populations can possess aversive levels of defence.
- There are conditions which interfere with anti-apostatic selection and allow diversity of appearance in poorly defended prey individuals.

2.1.1 Optimal toxicity.

As for the question of optimal investment into costly defence, the level of secondary defence is not observable by the predator and has no effect on the generalisation of aversiveness. The optimal toxicity of this model t_{opt}, where the derivative of Z w.r.t t is 0 at $t = t_{opt}$ in a population which plays t_{opt}, is given by

$$g_1 = \frac{F'}{F} - \frac{K'}{K} - aI_1\frac{Q'}{Q}\frac{H'}{H} = 0. \tag{2.3}$$

I will refer to the resident population with subscript 1 and to the mutants with subscript 2. Following the definition of function I in (Broom et al., 2008), the level of aversive information for the resident population is given as follows:

$$I_1 = ((1-a)D(r_1)H(t_1) + aD(r_2)H(t_2)S(x)), \tag{2.4}$$

where F' represents the derivative of the fecundity function (and similarly for other functions), and each function is evaluated at $t = t_{opt}$ and a specifc value of r (for a given r-value, t_{opt} will be different).

If the product I_1Q'/Q is decreasing with r_1, Equation (2.3) increases with r_1. This increase is compensated for by a variation of the optimal toxicity. If the

absolute gradient $|d/dt(K'/K)|$ is greater than $|d/dt(F'/F)|$ the benefits of secondary defence outweigh the cost on the fitness and the optimal toxicity will be an increasing function of r_1.

2.1.2 Aposematic signals.

Regarding aposematic signals, the predator generalises aversive information between the resident population and the mutants based on their similarity in appearance described by the similarity function $S(x)$. Mutant groups can potentially invade a resident population from two directions. Using the terminology of Broom et al. (Broom et al., 2006), the conditions for resisting mutant invasion are a composition of the effects of recollection on the amount of aversive information

$$g_2 = D'/D + aI_1Q'D'/QD, \tag{2.5a}$$

and the effects of generalisation

$$g_3 = (1-a)I_1Q'S'(0)/Q. \tag{2.5b}$$

Due to the peaked shape of the similarity function $S(x)$, these conditions are different for mutants with higher and lower values of r_2 than the population. For $r_2 < r_1$ they depend on the left derivative of Z

$$\frac{\partial Z_l(r_1, r_2)}{\partial r} = -g_2 + g_3 \quad \text{for } r_2 < r_1 \tag{2.6a}$$

which must be positive for stability, and for $r_2 > r_1$ the right derivative of Z

$$\frac{\partial Z_r(r_1, r_2)}{\partial r} = -g_2 - g_3 \quad \text{for } r_2 > r_1 \tag{2.6b}$$

which must be negative for stability.

This makes it easier for mutants with weaker signals to invade a population. Therefore, it is sufficient for stability of aposematic signals in *infinite* populations to show that (2.5a) is positive for mutants with weaker signals $r_2 < r_1$ with the value of r_1 for which the condition holds being R. As a consequence, all signals with conspicuousness $r_1 > R$ are also stable leading to an infinite number of stable solutions. With regard to *cryptic solutions* ($r_1 = 0$), a population can only be invaded by more conspicuous mutants ($r_2 > r_1$). The cryptic solution is stable if $-g_2 - g_3 < 0$. From this it follows:

(i) There is always a cryptic stable solution if there is no investment into secondary defence which occurs if $g_1(r_1 = 0, t_1 = 0) < 0$.

(ii) In the case of an investment into secondary defence a cryptic solution is

only stable if this investment is aversive causing $I_1 > 0$, a is sufficiently small (below some critical value) and the predator is able to sufficiently distinguish between the resident population and mutants.

2.2 The Evolutionary Dynamics of Aposematism: a Numerical Analysis of Co-Evolution in Finite Populations.

For my analysis I build on the model of co-evolution of secondary defence and signalling which provides a framework utilising aversive learning to predict negative as well as positive correlation. I use the model as introduced previously in Section 2.1 (Broom et al., 2006; Broom et al., 2008). The underlying model assumed a population equilibrium with constant population sizes and experienced predators with full knowledge of the prey population's aversiveness and signalling strategies, and that signals and toxins could impose costs on prey. Key predictions of this model were the destabilisation of crypsis and the diversity of aposematic solutions as a consequence of the underlying co-evolution. This analysis will extend the analytical considerations of this earlier work with numerical analysis. A key point is the introduction of the effects of finite populations on the evolution of aposematism. Through the introduction of drift I gain new insights into the inter-population diversification of aposematic displays in coherence with intra-population anti-apostatic selection. Especially in small populations, introducing drift as an additional process cannot be ignored and, as I show, it influences evolutionary stability. Drift is always a factor in real population systems (Willi et al., 2012) but is usually neglected for two main reasons: firstly it complicates analysis, and secondly it is generally assumed that it will not make too much of a difference if population size is sufficiently large. However, there is a growing body of evidence stressing the importance of drift as a force which can challenge natural selection (Barton and Charlesworth, 1984; Gillespie, 2001; Ellegren, 2009).

Additionally, the exploration of numerical methods and finite populations will substantially improve the accessibility of the previous models (Broom et al., 2006; Broom et al., 2008; Leimar et al., 1986) with regard to testability and validation of predictions. This will make these analytical models of the co-evolution of signalling and defence more understandable.

The following results of this model are based on specific functions governed by single parameters (Table 2.2) which are motivated by the discussion in (Broom et al., 2006; Broom et al., 2008) and were reintroduced in detail in Section 2.1: (i) the parameter f_0 describes the extent of the adverse impact

Symbol	Definition
$F(t)$	$\exp(-f_0 t)$
$K(t)$	$1/(1+k_0 t)$
$H(t)$	$t - t_c$
$D(r)$	$1 - \exp(-d_0(r+0.1))$
$S(x)$	$\max(0, 1 - v_0 x)$ with $x = \lvert r_i - r_j \rvert$
I_i	$\sum_j^N D(r_j)H(t_j)S(r_i, r_j)/\epsilon$ for $i \neq j$
$Q(I)$	$\min(1, \exp(-q_0 I) + q_{\min})$
Z_i	$F(t_i)/(D(r_i)Q(I_i)K(t_i))$
t	the level of toxicity of secondary defences
r	the conspicuousness of an aposematic signal
t_c	the level of toxicity which becomes aversive, hence for which $H(t_c) = 0$
ϵ	general encounter rate with prey individuals

Table 2.2: The specific functions used by the model of co-evolution in finite populations (Section 2.2) based on the exemplary functions as introduced in (Broom et al., 2006; Broom et al., 2008) and summarised in Section 2.1.

of the investment t on the fecundity $F(t)$. (ii) The parameter k_0 describes the significance of an investment t on decreasing the likelihood of being killed in an attack $K(t)$. (iii) The parameter d_0 describes the predator's ability to discover prey individuals $D(r)$. (iv) The parameter v_0 describes the predator's ability to differentiate between prey individuals $S(x)$ and (v) the parameter q_0 describes the predator's sensitivity towards the aversive information I in relation to the attack probability $Q(I)$.

Next I introduce necessary modifications to the original model (Broom et al., 2006; Broom et al., 2008): The previous model used a constant λ to represent secondary causes of death. For aposematism to be effective, predation needs to be a prominent selective pressure. To segregate the mechanisms of aposematism, the main risk of death in the adopted model is assumed to be due to predation only. I will introduce a minimal attack probability to the model instead which can reflect other possibilities of death as I will discuss later. It seems a valid assumption e.g. for short living arthropods that predation dominates other risks of death. The fitness function is given as follows:

$$Z_i = \frac{F(t_i)}{D(r_i)Q(I_i)K(t_i)} \ . \tag{2.7}$$

The proposed functions $D(r)$, $Q(I)$, and $K(t)$ (Broom et al., 2008) came with the disadvantage of potentially biological meaningless solutions when one of the functions approaches zero for some values. In particular, the original formulation of Q had to be adapted: as I assume $\lambda = 0$ in (2.2) from the previous model (Broom et al., 2006) it requires the introduction of a minimal

attack probability to avoid unrealistic immortality. When the rate of death is close to zero in this way, this issue is also important in regard to precision of floating point numbers in the numerical simulations: beyond the well-behaved range of $D(r)$, $Q(I)$, and $K(t)$ the functions introduced as part of the new methodology result in numerical instability following round-off errors. Lastly, I include a new parameter ϵ in the aversive information function I as a scaling factor of the contribution of each individual to the aversive information I which can be interpreted as a general encounter rate of a predator with the prey population. The resulting model incorporates selection depending upon both the strategy of the individual but also of the population, as a consequence of the generalisation of aversive information, and is difficult to evaluate analytically. The introduction of numerical analysis of finite populations will allow me to draw conclusions using evolutionary dynamics by looking into aspects of selective pressure and drift respectively as follows.

2.2.1 Visualisation of the Fitness Landscape: a Numerical simulation.

The fitness of an individual is described by the payoff function Z_i (Table 2.2, (2.7)) and depends on its strategy (r_i, t_i) and on the composition of the population due to the generalisation of aversive information based on similarity. I consider invasion of a mutant group into a monomorphic resident population, i.e. all members of the resident population play an identical strategy. As previously indicated, I will refer to the resident population with subscript 1 and to the mutants with subscript 2. Mutants can differ from residents in either of the two strategy components, and I consider the selective pressure in two different directions, one in each component. In each component I consider the payoff of mutants of different types x_2 against the resident population x_1, indicated by the payoff function $Z(x_2, x_1)$. The selective pressure is defined by the derivative of the payoff function Z with respect to x_2 (strictly this derivative does not exist in the direction of r but the expression below is still meaningful as I discuss in section 2.2.3), which is visualised as a gradient ∇Z over a grid of points using a numerical 5 point stencil approximation of each partial derivative separately with $h = 1 \times 10^{-5}$ as follows:

$$\frac{\partial Z}{\partial x_2} = \frac{-Z(x_1 + 2h, x_1) + 8Z(x_1 + h, x_1) - 8Z(x_1 - h, x_1) + Z(x_1 - 2h, x_1)}{12h} . \tag{2.8}$$

The population composition is defined by the values of the population size (N) and the mutant group fraction (a). The final visualisation shows the gradient of the payoff function ∇Z as a vector field (Fig. 2.1a) representing the

selective pressure, the fitness value of an individual in a homogeneous population in the background as a heat map (Fig. 2.1c), and the behaviour of $D(r)Q(I)$ as a contour plot (Fig. 2.1b). The product DQ can in principle increase, decrease or remain constant according to the choice of functions. It has been discussed that biological meaningful functions assuming aposematism to be a handicap signal indicating fitness advantage should result in an increasing or only slowly decreasing product DQ with increasing signal strength r in homogeneous resident populations (Broom et al., 2006). Therefore, we require $dDQ/dD > 0$ with $DQ = D(exp(-AD) + q_{\min})$ where $A = Nq_0H(t)/\epsilon$ as the definition of I in $Q(I)$ simplifies to a product of D and the scaling factor A in the case of a homogeneous prey population. Finally, from

$$\frac{d}{dD}DQ = \exp(-AD) - AD\exp(-AD) + q_{\min} \tag{2.9}$$

we have $A < (1 + q_{\min}\exp(AD))/D$.

Corresponding to this derivation the contour plot in Figure 2.1b shows the values of $A - (1 + q_{\min}\exp(AD))/D$ with values ≤ 0 indicating the parameter range where the product DQ is increasing with signal strength r.

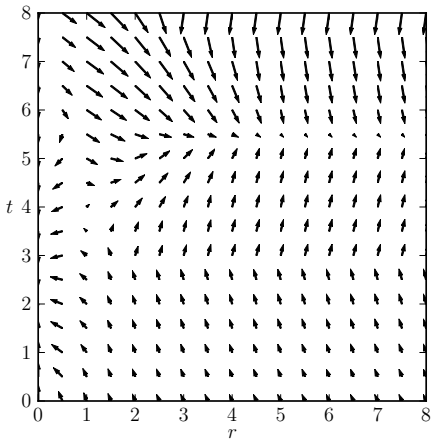
2.2.2 The Moran process and drift.

In finite, and especially in small, populations, random sampling can result in a change of allele frequency (Masel, 2011). Consequently, natural selection is not the only force acting on populations and neutral mutations (or in rare cases even unfavourable mutations) can take over entire populations. Therefore, the extent of drift needs to be considered in the evaluation of stability. The probability x_2 of a group of mutants ($a_n = aN$) invading a population is termed the fixation probability. The evolution of the population is modelled by a version of the Moran process (Moran, 1962), which is the classical way to model the evolution of finite populations. The Moran process is a Markov process where the state of the population (in this model denoted by the number of mutants) changes according to a transition matrix, and each change represents the replacement of an individual by one of another type.

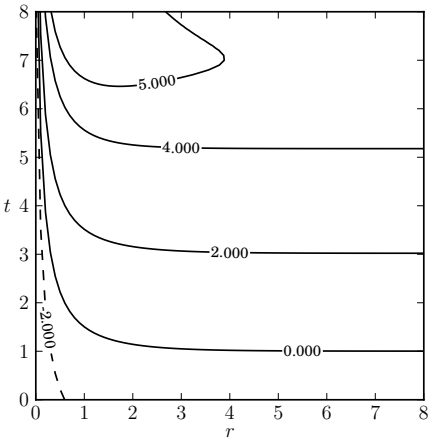
The determinant transition probabilities $p_{i\to i+1}$ and $p_{i\to i-1}$ that form the matrix are a combination of the chance of random selection and relative fitness $w_F = F_2/F_1$. As my model incorporates secondary defence, the transition probabilities are extended by the corresponding mortality w_K as defined as follows:

$$w_K = D(r)Q(I)K(t). \tag{2.10}$$

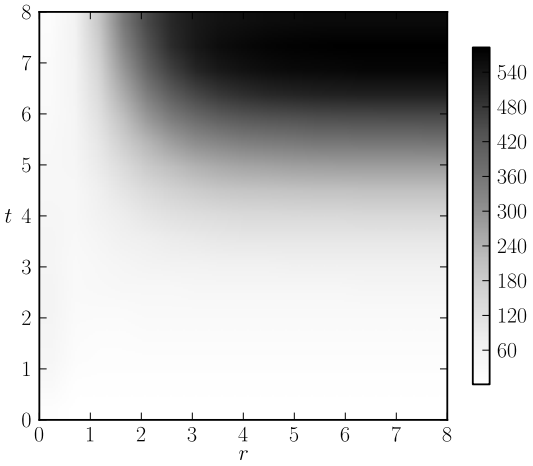
The final transition probabilities are a combination of random death and



(a) An exemplary vector field ∇Z as defined in Equation (2.8) representing the selective pressure on an individual as part of a finite population with the occurrence of mutants respectively.



(b) The qualitative behaviour of $D(r)Q(I)$ in homogeneous resident populations as described in the derivation from (3.2). The contour lines represent $\frac{\partial DQ}{\partial D}$ with negative values indicating an increasing product with signal strength r.



(c) The individual fitness Z_i (Equation (2.7)) in a homogeneous population without the occurrence of mutants.

Figure 2.1: The different elements of evolutionary dynamics in the co-evolution of aposematism in finite populations. All parameters as in Figure 2.4a.

fitness related birth:

$$\begin{aligned} p_{i\to i-1} &= \frac{N-i}{iw_F+N-i}\ \frac{i}{N}w_{K_2}, \\ p_{i\to i+1} &= \underbrace{\frac{iw_F}{iw_F+N-i}}_{\text{birth}}\ \underbrace{\frac{N-i}{N}w_{K_1}}_{\text{death}}. \end{aligned} \tag{2.11}$$

Thus for the mutant to increase in number ($p_{i\to i+1}$) a resident individual has to encounter a predator with probability $((N-i)/N)$ and has to be killed with probability (w_{K_1}) first. Secondly, it has to be replaced by a mutant according to the mutants relative fitness in the population with probability $((w_F i)/(w_F i+N-i))$.

For a group of mutants the fixation probability is given by the closed form of the transition matrix (2.12) (Weibull, 1997; Nowak, 2006):

$$x_2 = \frac{1+\sum_{j=1}^{a_n-1}\prod_{k=1}^{j}\gamma_k}{1+\sum_{j=1}^{N-1}\prod_{k=1}^{j}\gamma_k} \quad \text{with} \quad \gamma_i = \frac{p_{i\to i-1}}{p_{i\to i+1}}\ . \tag{2.12}$$

To reflect the non-constant selection of my model, the relative fitness w_F and the mortalities w_K are recalculated on each step based on the changed population composition. The strategy of the mutants is chosen along the derivative of the payoff function (2.13), that is, in the direction of strongest selection, to represent the toughest opponent possible with the highest fixation probability as follows:

$$\begin{pmatrix} r_2 \\ t_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ t_1 \end{pmatrix} + h\frac{\nabla Z}{\|\nabla Z\|}\ , \tag{2.13}$$

with $h = 1\times 10^{-5}$. The final visualisations in Figure 2.4 show the diversion from neutral drift as a log score $\log_{10}(Nx_2/a_n)/h$, with a score of zero occurring when $x_2 = a_n/N$, indicating equal fitness between mutants and residents and so full dominance of neutral drift over selective pressure. The fixation probability x_2 is only approximately linear in h for a small range of values because of the trade-off between detection probability and predator generalisation. Note that this means that the drift score too is only independent of h over small ranges.

For infinite populations a resident strategy which is fitter than any mutant strategy within a population composed of a mixture of the resident and mutant strategies, where the frequency of the mutants in the population is sufficiently small, can resist invasion from all such mutants, and is termed an *evolutionarily stable strategy* (ESS). The spatial structure of the population can be considered by allowing the fraction of mutants *in the local area* to be a significant proportion of the population, even when their overal proportion is small, and this approach was taken in the previous models (Broom et al., 2006; Broom et al., 2008) as well

as in my model. However, as the effect of mutants is expressed through a level of aversive information I which persists with the fully experienced predators the details of a spatial structure can be ignored in this framework. The definition of an ESS in a finite population requires an extra condition: in addition to the equivalent of the above condition (that the fitness of a single mutant within a population of residents is strictly less than that of the residents), the fixation probability of a single mutant needs to be less than $1/N$ (Taylor et al., 2004; Nowak, 2006). For my model I consider an invading group of size $a_n = aN$, and I thus adapt this definition accordingly: in order to consider the fitness of mutant and resident individuals within such a population mixture, I compare the fixation probability of the mutant group with the corresponding neutral probability of $x_2 = a_n/N = a$.

I note that in my model the two conditions above actually reduce to one (as is often the case), since the aversive information function increases with the increasing number of mutants, so if a mutant is fitter than the resident population when introduced as a small proportion of the population it will still be fitter when in a larger proportion, so the fixation probability condition is always satisfied when the fitness condition is.

2.2.3 Results.

As natural selection acts on the individual level, in *infinite* populations an evolutionarily stable solution (ESS) cannot be beaten by invaders (Section 2.1) (Christiansen, 1991). However, reducing population size quickly increases the effect of drift (Whitlock, 2000) which can destabilise populations. Using the functions presented in Table 2.2, the strategy dynamics were analysed regarding the co-evolution of r and t and the stability of strategies. Recall from Section 2.2.2 the conditions for a strategy to be an ESS in a finite population, and in particular that for my model I only need to consider a comparison of the fitnesses of resident and mutant strategies. Figure 2.4 shows two representative simulations. In general, the proportion of mutants needs to be relatively high for aposematism to evolve in small populations and the results will be discussed for each parameter respectively as follows.

Optimal Toxicity.

The original model (Broom et al., 2006; Broom et al., 2008) predicted that the optimal toxicity t_{opt} will be an increasing function of r (Section 2.1.1) and an exemplary simulation is presented in Figure 2.4a.

Additionally, Figure 2.4b shows the possibility of negative correlation between signal strength and level of defence. Firstly, I note that increasing t in

Equation (2.3) in Section 2.1.1 also increases the product I_1Q'/Q so that scenarios with strong aversiveness $H(t)$ are possible where the optimal toxicity is a decreasing function of r instead. That the aversiveness and its influence on the extent of learning can reverse the correlation between signal strength and level of defence represents a new insight for this model. Secondly, if the sizes of the gradients discussed in Section 2.1.1 are reversed, the optimal toxicity will be a decreasing function of r as the cost on fitness restricts investments into secondary defence.

This reflects two concepts: on the one hand, the signal correlates with the amount of secondary defence as the disadvantage of more conspicuous signals needs to be compensated for by better secondary defence. On the other hand, a clearer signal leads to more efficient aversive learning so that the investment into secondary defence can be lowered. The original claim (Broom et al., 2006) that the ESS value of t_1, t_{opt}, is increasing with that of r_1, r_{opt}, if $I_1 > 0$, $V(I_1) = -I_1Q'/Q$ is increasing with positive I_1 (2.4) at r_{opt}, and that $D(r)Q(I_1)$ is an increasing function of r_1, can be violated by having a steep aversiveness function $H(t)$. This allows a new type of solution for steep aversive information functions I_1 which depends on the scaling factor ϵ and the aversiveness $H(t)$. Additionally, increasing the predator's sensitivity towards aversive information via the parameter q_0 will have the same effect.

Figure 2.4 also shows that if I_1 (2.4) is not increasing sufficiently with large values of r_1, $lim_{r_1\to\infty} I_1Q'/Q = C$, the optimal toxicity levels out and is the solution to Equation (2.14):

$$-f_0 + \frac{k_0}{1 + k_0 t_1} - \frac{aC}{t_1 - t_c} = 0, \tag{2.14}$$

all constant terms are positive, except C which is negative. For $t_1 > t_c$ the expression on the right decreases with t_1. For t_1 just bigger than t_c it is clearly positive, after which it is decreasing, and in the limit as t_1 tends to infinity it is negative. There is thus exactly one root in (t_c, ∞) which is t_{opt}.

Aposematic signals.

As discussed previously in Section 2.1.2 (Broom et al., 2006) it is easier for mutants with weaker signals to invade a population and in *infinite* populations it is therefore sufficient to show that the left sided derivative of Z (2.7) is positive for mutants with weaker signals $r_2 < r_1$. Through the introduction of drift, this argumentation is no longer valid, as I discuss below. A sufficient condition for the *existence of aposematic signals* in finite populations remains that the left sided derivative of Z (2.7) is positive for mutants with weaker signals. The

group size of mutants a describes the tradeoff between the influence of D'/D and $S'(0)$. In the case of predators which are highly able to distinguish between individuals ($S'(0) \ll -1$), the evolution of aposematism from crypsis requires a to be large, as mutants do not benefit from a predator's generalisation between residents and mutants (2.5b).

With the focus on *stability* in finite populations, I must consider invasion by mutants with both higher and lower values of r_2. As I discussed in Section 2.2.2, in my model the mutant that is the fittest when introduced at small frequency also has the highest fixation probability. Thus the key question is whether mutants with higher or lower values of r_2 are the fitter, and thus even when the left and right derivatives are different, the expression in Equation (2.8) shows the pressure in the population due to drift. We would obtain a simple point solution where Equation (2.8) is equal to zero, provided that there is a compatible t_{opt}. The condition that the right sided derivative of Z (2.6b) is negative for mutants with stronger signals is generally difficult to satisfy: as a group of mutants usually benefits in an aposematic population from the generalisation of aversive information, most solutions will be unstable in r. Therefore, there is theoretically no stable level of signalling. However, the dominance of drift can create a pseudo-stability which has similar characteristics to the infinite amount of stable solutions with $r_1 > R$ as predicted for infinite populations in Section 2.1.2. Here there is a range of different values of r_1 and t_1 which gives an area of stability rather than just a line: the extremely flat fitness landscape leads to a region where drift is approximately neutral to 3 decimal places (Figure 2.4).

With regard to the cryptic solution, it is always stable if the predator is not deterred by weak warning flags. This is the case in the model presented as a consequence of the functional form of Q in Table 2.2. The aversive information I has to reach a threshold of $I > \ln(-1/(q_{\min} - 1))/q_0$ to have any deterrent effect on the predator's attack probability Q resulting in $Q = 1$ and $Q' = 0$ if the threshold is not reached in the case of too weak warning flags.

Minimal attack probability.

As discussed earlier, the introduction of a minimal attack probability $q_{\min}$ in Equation Q in Table 2.2 avoids the problem of immortality in the model. There exist two options to introduce a minimal attack probability:

(i) On the one hand, the attack probability function Q can be shifted by $q_{\min}$ resulting in $Q(I) = \min(1, \exp(-q_0 I) + q_{\min})$.

(ii) A second possibility would be for the minimal attack probability to be realised using a simple cut-off value: $Q(I) = \text{mid}(1, q_{\min}, \exp(-q_0 I))$. The

cut-off would result in $Q' = 0$ for sufficiently low values of Q, which requires stronger optimal secondary defence considering Equation (2.14).

Option (i), using a shifted function Q, is the choice of my model as I think results are more realistic (the results of option (ii) are not presented) and allows the interpretation as a source of secondary cause of death as I discuss as follows: for high values of (r_1, t_1) the fitness function is extremely flat and changes in the parameters (r, t) have barely any effect on the payoff which is clearly dominated by almost neutral drift. This results in Q being effectively independent of (r, t) and $q_{\min}$ being the dominant influence. This means that the minimal attack probability behaves similar to the introduction of a secondary cause of death λ. Even though this model is simplified assuming $\lambda = 0$ (2.2) it is able to reproduce the effects of secondary causes of death.

2.2.4 Discussion.

This model introduces a more flexible methodology for assessing evolutionary stability in previous game theoretical models of the co-evolution of aposematic signalling and secondary defence (Broom et al., 2006; Broom et al., 2008). My main conclusions are:

(i) the number of mutants needs to be relatively high (e.g. within a locality of the population) for aposematism to evolve in small populations,

(ii) drift is an important force acting on real population systems and increases inter-population diversification of aposematic solutions,

(iii) in terms of evolutionary stability drift results in a region comparable to an evolutionarily stable set (ESSet), where strategies can change within the region due to approximately neutral drift, but resist invasion from outside,

(iv) anti-apostatic selection (selection against rare prey types) prevents intra-population dishonesty of the aposematic display (automimicry as well as continuous variation in toxicity), and

(v) enhanced predator aversion learning reduces the level of aversive defence investment through the compensation of more conspicuous displays.

In previous models of the co-evolution of aposematic signalling and defence (Broom et al., 2006; Broom et al., 2008) the conditions for the evolutionary stability of a signal and associated level of defence were found for a model assuming a large (effectively infinite) population. A feature of the model was that there was an infinite set of evolutionarily stable strategies (ESS) in many cases, where for a given level of signal there was a unique level of defence,

but as long as the whole population displayed the same signal it was stable against invasion by any other strategy. In particular this infinite set consisted of a continuum (a line) of ESSs, so for any ESS a small mutation could be to another ESS as a potential invader. Thus it was of particular interest to consider finite populations, and consequently the effect of drift, and whether the strategies which were ESSs in the original model could still be considered stable.

Hence, I shall elaborate on how the introduction of drift as an additional evolutionary force in finite populations acts on the resident population. Drift cannot be neglected when the population size is relatively small and the payoff function is relatively flat, as in the original models. As a side effect of the flat gradient of the payoff function almost neutral drift dominates for values of high secondary defence and strong signal strength. Even though a distinct aposematic solution may exist, populations hardly converge towards them as drift dominates selective pressure. Notably in the case of small population size, the diversity of aposematic solutions extends from the previously predicted line to a wider plane-like parameter range (Figure 2.4). With regard to stability this result can be interpreted as a finite population version of the idea of an evolutionarily stable set, where strategies can change within the region due to drift, but resist invasion from outside. The widespread variation of secondary defence and aposematic displays has been discussed in Speed et al. (2010) as a consequence of frequency-dependent intra-population cheating in an ecological model in which prey acquires its anti-predatory defences from the environment, e.g. from a food source. In my co-evolutionary model the inter-population diversification is a consequence of drift which is an interesting result: intra-population cheating (or automimicry) is problematic in the context of evolutionary stability as it undermines the effectiveness of the signal since cheating appears to be at a selective advantage (Jones et al., 2013). Drift instead allows a wide diversity of inter-population aposematic solutions without the introduction of destabilising cheating or automimicry on the intra-population level.

On the other hand, the stability of aposematism is tightly bound to anti-apostatic selection: even though the diversity of stable inter-population aposematic solutions is high, a stable aposematic population needs to look alike or the level of aversive information suffers and aposematism loses its advantage (2.5b). The required degree of uniformity depends upon the predator's ability to distinguish different prey individuals (v_0). The condition of close resemblance holds if mutations are mostly silent without effects on the phenotype and the mutation rate is reasonably low.

In addition to the solutions predicted from the earlier models of Broom et al. (2006) and Broom et al. (2008), I observed a new set of possible solutions

with negative correlation between secondary defence and signal strength (Figure 2.4b). For this solution to appear, the aversiveness of secondary defence needs to be rapidly increasing ($H(t)$) or the predator needs to be very sensitive towards aversive information (via the parameter q_0). In this thesis I focused on examples for the later case of increased sensitivity towards aversive information via the learning related parameter q_0. Under these circumstances, increasing conspicuousness improves prey distinction and stimulates aversive learning to such a degree that necessary investments into secondary defence can be lowered. This result is contrary to the decreasing aposematic display with increasing toxicity as consequence of aposematic display and investment into toxicity competing for a common resource as in Blount et al. (2009) and Lee et al. (2011). In my model the accelerated learning process of strong aversion is the reason of the negative correlation and allows lower levels of toxicity with increasing conspicuousness. See Section 1.2.2 which discussed Speed and Ruxton (2007) for another example of this phenomenon as a result of specific marginal costs using Gompertz functions rather than aversive learning.

Generally, the frequency of mutants has to be relatively high for aposematism to evolve in small populations. This may be seen as requiring a component of kin selection or the invasion of a rival population into a locality.

Furthermore, the predatory risk of death needs to dominate other risks of death for aposematic solutions to emerge. As an unbounded payoff function has the side effect of unrealistic immortality the attack probability needs to have a minimal value of $q_{\min}$. This makes additional risks of death redundant as they can be considered as part of $q_{\min}$ as discussed earlier.

Finally, the cryptic solution is always stable if the predator is not deterred by small diversions from full crypsis. This seems like a reasonable conclusion and for questions related to the possibility of overcoming the stability of crypsis I refer to models of the initial emergence of aposematism describing mechanisms such as dietary conservatism, a shifted peak of the aversive information function, the opportunity cost of crypsis, and specific properties of the predator's perception or cognitive processes (Mappes et al., 2005; Speed and Ruxton, 2005; Marples et al., 2005; Lee et al., 2011).

It is evident from the limitations of my model that future work requires a more sophisticated description of aversive learning moving from the equilibrium of educated predators to the original considerations of uneducated predators. Reinforcement learning will be used in the following chapter to look into conditions of special cases such as mimicry, which cannot be explained by bare co-evolution of warning displays and secondary defences.

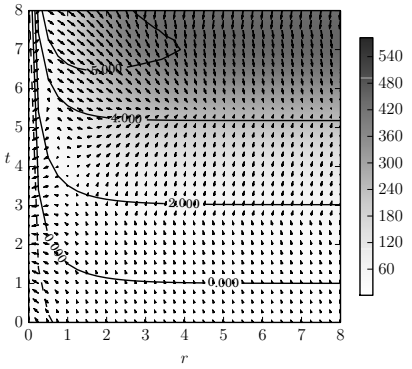
```
Def gradient_Z(r, t):
  h = 1e-5
  population(x, y) = (1-a) * N * [r, t] + a * N * [x, y]
  partial_t = (-Z(r, t+2h, population(r,t+2h) )+ 8Z(r,t+h,
    population(r,t+h) -8Z(r,t-h, population(r,t-h)) + Z(r,t-2h,
    population(r,t-2h))) / 12h
  partial_r = (-Z(r+2h, t, population(r+2h, t)) + 8Z(r+h,t,
    population(r+h, t)) -8Z(r-h, t, population(r-h, t)) +
    Z(r-2h, t, population(r-2h, t))) / 12h

Def Z(r, t, population):
  z = Fitness(t) / (Discover(r) * Attack(r, population) *
    Killed(t))
```

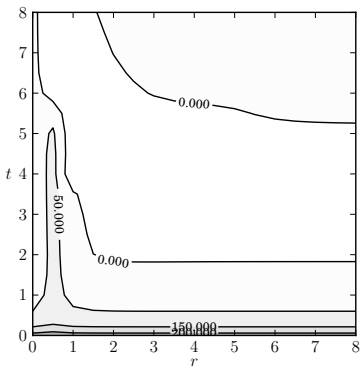
Figure 2.2: Pseudo-code of the 5 point stencil approximation of selective pressure defining the fitness landscape.

```
for r in range(R):
  for t in range(T):
    fit_vector = gradient_Z(r,t)
    mutants = (h * fit_vector / length(fit_vector)) +
      array([r, t])
    residents = array([r,t])
    for i in range(N):
      population = residents + i * mutants
      fit_residents = F(t)
      fit_mutants = F(mutants[1])
      w_F   = fit_mutants / fit_residents
      w_K_1 = Discovered(r)*Attacked(residents)*Killed(t)
      w_K_2 = Discovered(mutants[0])*Attacked(mutants)*
        Killed(mutants[1])
      p_minus = ((N - i)/(w_F*i + N-i)) * (i/N)*w_K_2
      p_plus = ((w_F * i)/(w_F*i + N-i)) * ((N-i) / N)*w_K_1
      gamma[i] = p_minus / p_plus
    x_a = (1.0 + np.sum( map(np.prod, [[gamma[i] for i in
      range(j)] for j in range(a-1)]) )) /
      (1.0 + sum( map(prod,
      [[gamma[i] for i in range(j)] for j in range(N-1)]) ))
    x_a = log10( N*x_a/a ) / h
```

Figure 2.3: Pseudo-code of drift estimation.

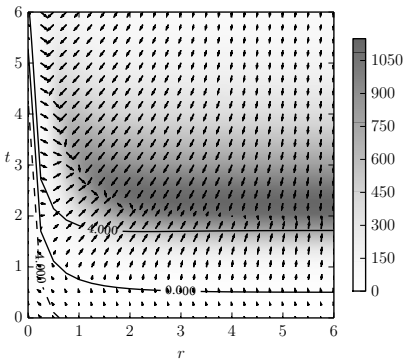


(a.1) Visualisation of the fitness landscape indicating the selective pressure on the population as described in Section 2.2.1 and detailed in Figure 2.1.

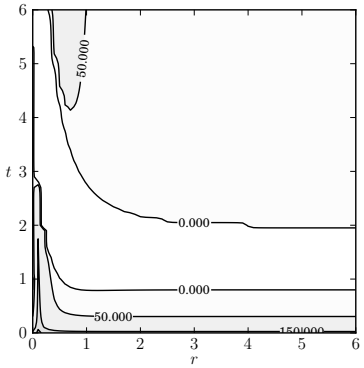


(a.2) Visualisation of drift as introduced in Section 2.2.2. The plot utilizes 3 significant figures leading to a wide area of neutral drift within the 0.000 boundaries.

(a) Exemplary populations dynamics of positive correlation for a simulation with the parameters: $n = 500$, $\epsilon = 500$, $a_n = 200$, $v_0 = 1$, $d_0 = 1$, $t_c = 0$, $q_0 = 1$, $q_{\min} = 1 \times 10^{-3}$, $k_0 = 5$, $f_0 = 0.5$.



(b.1) Visualisation of the fitness landscape indicating the selective pressure on the population as described in Section 2.2.1 and detailed in Figure 2.1.



(b.2) Visualisation of drift as introduced in Section 2.2.2. The plot utilizes 3 significant figures leading to a wide area of neutral drift within the 0.000 boundaries.

(b) Exemplary populations dynamics of negative correlation for a simulation with the parameters: $n = 500$, $\epsilon = 500$, $a_n = 250$, $v_0 = 10$, $d_0 = 1$, $t_c = 0$, $q_0 = 2$, $q_{\min} = 1 \times 10^{-3}$, $k_0 = 5$, $f_0 = 1$. The decisive parameter for negative correlation to occur is q_0 which affects the predator's sensitivity towards aversive information. The other parameters were changed for scaling purposes.

Figure 2.4: Results of the co-evolution of aposematism in finite populations including the influence of drift. The decisive parameter for the qualitative behaviour of the correlation of r and t is the learning related q_0. The other parameters are modified for scaling purposes.

Chapter 3

Aposematism from the predators' perspective.

In the previous chapters I presented a wide body of theory which addresses the emergence and evolution of aposematism and the effects of aposematism on the evolution of prey. In this chapter I will move the focus onto the predator. As the selective agent, the field has a renewed interest in the role of the predator's aversive learning process. Within the context of aposematism an interesting but also fundamental question remains: how does a predator incorporate the information gained from an encounter with prey into a generalised approach to predation and defence? I will discuss what motivates behaviour in predators and especially, which currency might drive foraging behaviour in aposematic predator-prey systems. In particular, this chapter will investigate the effects of aposematic prey on the fitness and energy intake of predators to better understand the selective pressure arising from aversive learning.

Parts of Section 3.2 have been published in Teichmann et al. (2014a) and parts of Section 3.3 have been submitted for publication.

3.1 Reinforcement learning.

Learning is a widely present and successful strategy of adaptation. The importance of studying animal behaviour with the purpose of researching learning arose from empirical and evolutionary paradigms within psychology. The growing importance of empirical methodologies and the emergence of evolutionary theories pointing out the relation of man and animals have paved the way for the study of animal behaviour (Shettleworth, 1999).

Ensuing studies into the aspects of predictive learning have been mostly con-

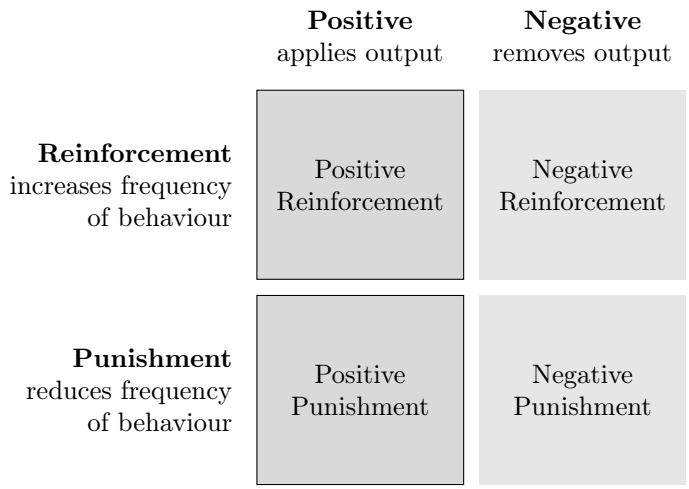


Figure 3.1: The different procedures in operant conditioning (or associative learning of actions and consequences) (Bouton, 2007).

ducted based on the principles of conditioning. In conditioning two events or so called stimuli are repeatedly paired which results in the formation of a link between them. This link will allow an individual to predict the occurrence of one event upon the presentation of the other event (Mackintosh, 1994; Pearce and Bouton, 2001; Hall, 2002; Alonso and Schmajuk, 2012). In classical conditioning such associations are formed between stimuli themselves where a neutral stimulus becomes conditioned when repeatedly paired with an unconditioned stimulus or reinforcer. Examples of classical conditioning are forms of fear response or taste aversion. In the case of operant or instrumental conditioning the associations are formed between stimuli and voluntary responses. Examples are associations between behaviour and reward or punishment as a consequence for that behaviour. Operant conditioning is defined by two dimensions (Figure 3.1): on the one hand, it is characterised by the frequency change of some behaviour. We speak of rewards, or *reinforcement*, if an increase in the frequency of some behaviour can be observed. On the contrary, *punishment* is defined as to reduce the frequency of some behaviour. On the other hand, operant conditioning is classified by the nature of the response to some behaviour: we speak of *positive* reinforcement or punishment if the response to some behaviour is in the application of an output and of *negative* if the response to some behaviour is in the removal of an output. For the remainder of this thesis we will focus on the case of *positive reinforcement* and *positive punishment*.

Either way, with each pairing, the prediction error – the discrepancy between the predicted outcome and the actual outcome – is reduced through learning. With increasing associative strength between the events one stimulus may fully

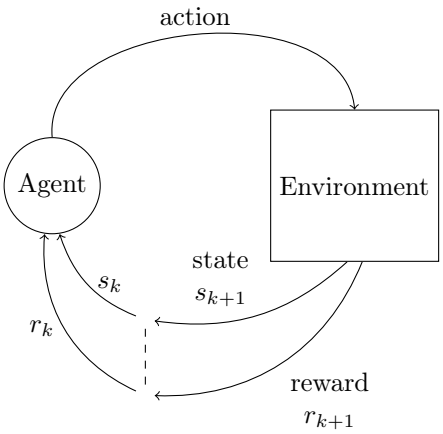


Figure 3.2: The reinforcement learning model with the two entities agent and environment. The agent's actions at iteration k have subsequent effects on the environment's state and rewards at iteration $k+1$. The dashed line indicates that the agent experiences the consequences of its actions with a delay.

predict the other, at which point no further learning occurs. Thus, during early phases of conditioning large prediction errors produce great increases in associative strength. But as learning progresses these changes decrease in size. Finally, with the growing ability of one stimulus to predict the other the associative strength approaches an asymptotic level. This simple idea is pervasive in psychology as it accounts for many learning phenomena which are at the basis of complex cognitive processes, and in that its predictions have been observed across species. Additionally, it is also the core of a great number of clinical models (Haselgrove and Hogarth, 2013; Schachtman and Reilly, 2011).

A modern theory of operant conditioning in particular can be found within the computational field of *reinforcement learning* (RL) which provides a normative framework for optimal behaviour in order to maximise rewards and avoid punishment. The definition of reinforcement learning is vague as it is the description of a learning problem rather than the characteristics of a particular methodology. The learning problem is to find a mapping of situations to actions which maximise a reward signal. This mapping has to be obtained through exploration, trial and error, or goal-directed learning from interactions, which distinguishes reinforcement learning from supervised learning from an external supervisory signal, learning from examples. A second characteristic of reinforcement learning is delayed rewards and actions having subsequent consequences on future rewards. These two characteristics define the learning problem (Figure 3.2) and any method solving it is considered to be a computational reinforcement learning method.

This definition of the learning problem might sound abstract but reinforcement learning is actually very intuitive as it describes the learning problem of most situations where individuals interact with their environment using trial and error to solve a task. The reinforcement learning problem emphasis that in more realistic environments it is usually impossible to define examples of desired behaviour which are representative for all states of an individual's environment. But it is in these unknown situations when learning is most beneficial and an individual has to rely on its own experience. This is where a classical trade-off arises between exploration and exploitation: to maximise a reward an individual should perform actions that it knows to be rewarding from previous experience. But to find such actions in the first place an individual had to explore actions with unknown outcome. The dilema between exploting existing experience to obtain rewards and exploring new actions in order to improve decisions in the future is characteristic of reinforcement learning.

3.1.1 The elements of reinforcement learning.

Besides the two distinct main components of the RL model (Figure 3.2) – agent and environment – a RL method has three additional common elements: a *policy*, a *reward function*, and a *value function*. A model of the environment itself is optional and not necessary for solving the RL problem. A model might be useful in planning an individual's actions which can improve learning. In the following models of this chapter I will focus on model-free RL methods. I refer to Chapter 4 where I will describe two model-based RL methods.

The policy, denoted by $\pi(x, u)$, describes the agent's mapping of perceived states (x) of the environment to actions (u). According to the previous introduction of the conditioning paradigm, the policy describes the stimulus-response association. Thus, the policy is the core of each RL method as it sufficiently determines behaviour of the agent in its environment. There are many different implementations of policies which differ greatly in their complexity and application. As pointed out previously a policy has to address the exploitation-exploration trade-off. Within this context lays an import distinction between *on*-policy and *off*-policy learning methods. In on-policy learning the individual improves the policy which it uses to determine behaviour. Consequently, such policies have to be stochastic, or *soft*, to be suitable for the RL problem which results in $\pi(x, u) > 0$ for all states x in the state space and all actions u in the action space. Examples of such a policy are ϵ-greedy policies or Gibb's soft-max policy. An example of a simple on-policy learning algorithm is $SARSA$ which is an acronym for 'state-action-reward-state-action' (Rummery and Niranjan, 1994). It describes the basic elements of the algorithms iterative update rule

and is the on-policy version of the later introduced Q-learning algorithm in Section 3.1.3.

In off-policy learning methods the individual uses two distinct policies: the first policy determines behaviour, called the *behaviour policy*, and the second policy is the one which is improved, called the *estimation policy*. This distinction allows the agent to use a deterministic behaviour policy as long as the estimation policy is stochastic without contradicting the RL requirement of exploration. Consequently, the usage of an off-policy algorithm allows the separation of the exploration trade-off from the optimal control problem, as I will elaborate on in more detail in the introduction of Q-learning in Section 3.1.3. Thus, off-policy algorithms are well suited for the application in mathematical models for their analytical tractability as illustrated in Section 3.2.

The reward function in RL describes the immediate feedback from the environment as a single numerical value. The rewards are strictly state dependent and the RL objective is to maximise the reward an agent receives in the long term. The agent has no influence on the reward function itself but it can alter its policy to affect the state of the environment and the rewards received accordingly.

The reward function itself is only an indicator of the immediate quality of the environment. Complementarily, the value function (3.1),

$$V^{\pi}(x) = E_{\pi}\left\{R_k | x_k = x\right\}, \tag{3.1}$$

is the indicator of the long term desirability of a realised state x_k at iteration k of the environment.

In RL the policy is based on the predictions of the value function. The value function acts as a predictor of future rewards: it describes the total reward R an agent can expect of a state and future states of the environment following the agent's policy. In RL the aim of an agent is to select actions leading to states with high values not necessarily states with high rewards. This allows an agent utilising a RL method to optimise long term reward accumulation even though an action might yield only suboptimal immediate rewards. Whereas the rewards are a primary quantity of the environment the value of a state has to be estimated by the agent. This estimation of the value function is at the core of the RL problem describing the future rewards depending on the long term effects of an agent's actions influencing the future states of the environment.

Strictly speaking, the RL problem can also be solved without estimating the value function. There are examples of search methods which solve the RL problem successfully. Such methods search the policy space instead of the value space and are usually applied if the individual cannot perceive the state

of the environment or the learning problem cannot be described in terms of a Markov decision process. There are two main categories for searching the policy space: *evolutionary* methods, e.g. genetic algorithms and simulated annealing, and *direct policy search* methods, e.g. policy gradient ascent, policy search by dynamic programming (Bagnell et al., 2003; Peters and Schaal, 2008). I will explore a policy gradient method in chapter 4. Within this chapter I will discuss the relation of learning and evolution in more detail in Section 3.3.

In summary, solving the RL problem with its different elements is composed of two sub-problems called (i) the *prediction* problem and (ii) the *control* problem. The prediction problem is closely related to the policy evaluation with the objective to estimate the value function under the current policy V^{π}. In addition, the control problem is about finding the optimal policy π^{*}.

3.1.2 Temporal difference learning.

As previously discussed, in RL an individual learns from experience of interacting with its environment. Each time it performs an action in some state the individual receives a real-valued reward that indicates the immediate value of this state-action transition. However, for a complete RL method it requires a strategy of how to use the gathered experience. The two traditional methods of learning the value function in RL are *dynamic programming* and *Monte Carlo methods*. In Monte Carlo methods the individual learns the value function directly from the final return of a state and its actions without requiring a model of the environment. In dynamic programming the individual uses a model of the environment's dynamics to learn from previous experience without having to wait for a final outcome of its actions. *Temporal difference learning* is a method which resides between dynamic programming and Monte Carlo methods (Barto et al., 1989).

The last decade has seen a proliferation of research on the neural and psychological mechanisms of reinforcement learning (Dayan and Daw, 2008; Doya, 2007; Maia, 2009; Niv, 2009; Rangel et al., 2008; Schultz, 2002; 2007). In turn, reinforcement learning has been postulated as a general model of human economic decision making and neuroeconomics (Glimcher et al., 2008; Platt and Huettel, 2008; Rangel et al., 2008; Schultz, 2008). We know from studies of neural correlates in behaving animals that reinforcement signals in the brain represent the reward prediction error rather than a direct reward-reinforcement relation (Berns et al., 2001; Niv, 2009; Schultz et al., 1997; Montague et al., 1996). Temporal difference (TD) learning is a RL methodology which reflects these insights by representing states and actions in terms of predictions about future rewards. On the one hand, like dynamic programming, TD learning uses

previous estimates of the value function to learn continuously without having to wait for the final return of a state and an individual's actions. On the other hand, like Monte Carlo methods, TD learning is model-free. The environment is represented by moving targets rather than by a model and the learning objective is to iteratively update the targets towards its true values based on experience from interactions with the environment. Furthermore, the computational theories are increasingly supported by experimental data describing the activity of dopaminergic neurons, mediate reward processing and reward dependent learning (Schultz et al., 1997; Montague et al., 2004; Daw and Doya, 2006; Dayan and Niv, 2008).

I will discuss the details of TD learning by considering Q-learning, an exemplary TD learning method as follows.

3.1.3 Q-learning.

In the models of aversive learning to come I will make use of Q-learning extensively. As outlined previously, the learning individual will have no predefined model of the environment. Rather the learning individual has to draw on experience from trial-and-error interactions with its environment to learn optimal behaviour, in particular, it will use Q-learning.

Q-learning is a simple algorithmic implementation of reinforcement learning. In particular, it is a model free method which allows learning about sequential decision tasks in a Markovian environment from experienced rewards without the necessity of building representations of the environment. Instead, the algorithm uses moving target values as I will explain below. With regard to the different elements of RL, Q-learning is an off-policy TD control method:

The learning process takes place in the value space and consists of a reward prediction R termed the *action-value function* (3.2) of taking action u in state x at iteration k:

$$Q(x, u) = E\{R_k | x_k = x, u_k = u\}. \tag{3.2}$$

The condition for the application of the action-value function and Q-learning is a Markovian decision process:

$$P\{x_{k+1} = x',\, r_{k+1} = r \,|\, x_k,\, u_k\}. \tag{3.3}$$

The individual learns from iterative interactions with its environment. At each iteration k the learning individual finds itself in state x_k of its environment.

The actual learning process targets the individual's value prediction following action u_k in state x_k as described by the action-value function (3.2). This action-value function is an approximation of the actual function $Q^*(x, u)$. Con-

sequently, the aim of the learning process is to find $Q^{\pi}(x_k, u_k) \approx Q^*(x, u)$.

To obtain the current Q values it involves an iterative update process (Q-learning) which is typically formulated in an algorithmic representation because of its origin in computing, as follows:

$$Q'(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha \underbrace{\left(\overbrace{r_{k+1} + \gamma \max_{u_{k+1}} Q(x_{k+1}, u_{k+1})}^{\text{target}} - Q(x_k, u_k) \right)}_{\text{prediction error}}, \quad (3.4)$$

with α being the learning rate. Figure 3.3 shows the Q-learning algorithm in pseudo-code. Importantly, the individual not only takes immediate rewards into account but also the sum of discounted future rewards with γ being the discount factor. This combines an ubiquitous interest in rewards with the uncertainty of future events namely:

$$\begin{aligned} R_k &= \sum_{i=0}^{K} \gamma^i r_{k+i+1} \\ &= r_{k+1} + \sum_{i=1}^{K} \gamma^i r_{k+i+1} \\ &= r_{k+1} + \gamma \sum_{i=0}^{K} \gamma^i r_{k+i+2} \\ &= r_{k+1} + \gamma R_{k+1}. \end{aligned} \quad (3.5)$$

Finally, the learning individual bases its decision process on $Q(x_k, u_k)$ following a Gibb's soft-max policy:

$$\pi(x, u) = P\left(u_k = u | x_k = x, Q(x_k, u_k)\right) = \frac{\exp\left(Q(x, u)\right)}{\sum_u \exp\left(Q(x, u)\right)} . \quad (3.6)$$

Effectively knowing all of the current Q values gives the probability that the individual chooses a specific option for the next interaction with the environment. As Q-learning is an off-policy learning method, the individual derives an optimal estimation policy π^* from approximating Q^* through a greedy (deterministic) selection $\max_{u_{k+1}} Q(x_{k+1}, u_{k+1})$. This makes Q-learning the preferred method of this thesis as the deterministic behaviour policy in Q-learning allows an analytical solution of the learning problem in Section 3.2. However, as discussed previously in Section 3.1.1 Q-learning has to address the exploration-exploitation trade-off which is the reason for choosing Gibb's soft-max policy as a stochastic estimation policy.

Now, the iterative Q-learning algorithm expands as follows: at iteration k,

the learning individual interacts with the environment of state x_k which is a realisation from the state space X. Following the decision policy π, the learning individual takes action u_k out of the action space U. As a result of this interaction at iteration k, the individual experiences an immediate reward r_{k+1}. The terminology refers to the experienced reward at the subsequent iteration $k+1$ which emphasises that the reward is a consequence of the individual's action. Next, the learning individual forms a target value which is a composition of the experienced reward r_{k+1} and discounted future rewards. Thereby, future rewards are unobserved and a prevailing estimate $Q(x_{k+1}, u_{k+1})$. This allows the agent to learn continuously by filling in future rewards with moving averages which is known as *bootstrapping*. The difference between the target value and the estimate at iteration k gives the prediction error. Finally, the Q-learning algorithm updates the estimate $Q(x_k, u_k)$ to $Q'(x_k, u_k)$ towards the formed target value, subsequently reducing the prediction error. As the Q-learning algorithm uses bootstrapping, these targets are moving ones.

The recursive component of the Q-learning algorithm is a fundamental property used across RL described by the *Bellman equation*. The Bellman equation allows the optimisation problem to be broken up into simpler sub-problems by describing the value of a decision in terms of the value of an initial choice and the value of the remaining decision problem. The optimal value function (3.1) has to be a result of choosing optimal actions in each state. Consequently, the optimal value function can be written as a function of the action-value function (3.2) as follows

$$V^*(x) = \max_u Q^*(x, u). \tag{3.7}$$

Using the Markov properties of the learning problem the Bellman optimality equation (Sutton and Barto, 1998) allows a redefinition of Q^* in terms of a recursive optimisation of each decision based on state transition and reward emission probabilities as follows:

$$\begin{aligned} Q^*(x,u) &= E\big\{r_{k+1} + \gamma \max_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) | x_k = x, u_k = u\big\} \\ &= \sum_{x_{k+1}} P(x_{k+1}|x_k = x, u_k = u)\Big[P(r_{k+1}|x_k = x, u_k = u, x_{k+1}) \\ &\quad + \gamma \max_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\Big] \end{aligned} \tag{3.8}$$

and from (3.7) we get

$$\begin{aligned}\max_u Q^*(x,u) &= V^*(x) \\ &= \max_u \sum_{x_{k+1}} P(x_{k+1}|x_k = x, u_k = u)\Big[P(r_{k+1}|x_k = x, u_k = u, x_{k+1}) \\ &\quad + \gamma V^*(x_{k+1})\Big].\end{aligned} \tag{3.9}$$

Equations (3.8) and (3.9) are Bellman optimality equations for Q^* and V^* respectively which have an unique solution in the case of finite Markov decision problems independent of the specific policy (Sutton and Barto, 1998). Q^* represents a one-step-ahead search and provides the optimal expected long-term payoff. Therefore, any greedy policy on Q^* is optimal. However, in the application of Q-learning to biological problems of predator-prey interactions I presume that a constantly changing environment requires on-going exploration and learning as Q^* changes with time, e.g. through changes in the availability and yield of food sources. Therefore, exploration is a precondition and persistent cost of learning in the following models with the choice of Gibb's soft-max policy as the estimation policy in the models of this chapter.

```
Q ← 0
s_k ← s_0
WHILE learning DO
    a_k ← π(s_k, Q)
    s_(k + 1) ← f(s_k, a_k)
    Q(s_k, a_k) ← Q(s_k, a_k) + α (r_(k + 1) +
        γ max_a Q(s_(k + 1), a) − Q(s_k, a_k) )
    s_k ← s_(k + 1)
```

Figure 3.3: Q-learning algorithm in pseudo-code.

3.2 An application of Q learning in optimal diet models.

In this model I will apply Q-learning to investigate foraging behaviour in uncertain environments. I will analyse a predator's diet choice and energy intake in the light of defended prey and the presence of Batesian mimics within a context of aversive learning. In particular, I will introduce a pre-condition of exploration for successful aversion formation and show how it predicts foraging behaviour in the presence of conflicting rewards which is conditionally suboptimal in a fixed environment but allows better adaptation in changing environments.

3.2.1 Introduction to optimal foraging theory.

Optimal foraging theory (OFT) is an ecological theory which makes predictions about foraging behaviour. Foraging is a main component of animal behaviour and observing animals in the wild usually shows them either searching for food or feeding. The motivation of OFT lays in the fact that survival is dependent on a sufficient energy intake. Episodes of starvation can negatively affect an individual's fitness and prolonged starvation can lead to death. Therefore, an individual's survival and consequent fitness is a function of its foraging success. Assuming that there is individual variance in the foraging strategies which determine foraging success, natural selection will act on these foraging strategies and evolution will take its course. But not only the survival of an individual is dependent on its foraging success. Even in an environment where food is abundant an individual must have a sufficient energy intake for reproduction. Thus, OFT predicts that natural selection will act on the efficiency of energy acquisition and storage in relation to an individuals survival and reproduction by maximising the energy gain per unit time:

$$\frac{E}{h+s}\ , \tag{3.10}$$

where E is the average energy per prey item, h is the average handling time of a prey item, and s is the average time taken to acquire the prey item. Prey profitability as defined by (3.10) is the most commonly assumed currency in OFT and allows for some interesting predictions:

- If the handling time is shorter than the time to acquire a prey item ($h < s$) predators will be generalists, meaning they will forage on every prey item they encounter as prey is sparse.
- In the opposite case of prey with longer handling times ($h > s$) predators will be specialists and ignore mediocre prey items.

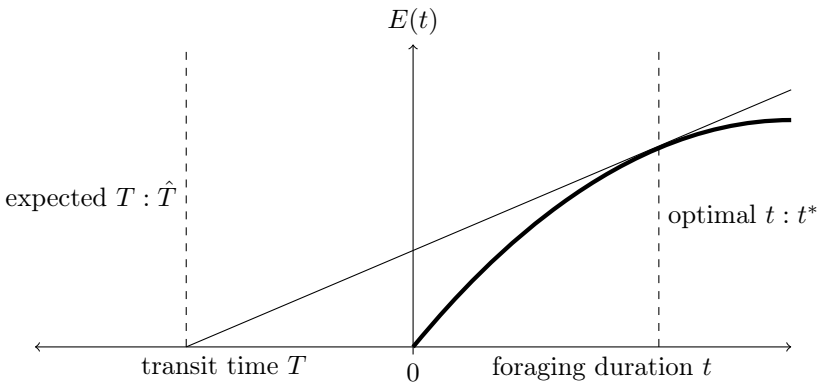


Figure 3.4: Marginal Value Theorem. The optimal foraging time in a patch t^* is given by the tangent of the cumulative energy gain from foraging in a patch $E(t)$.

These predictions are made under the assumption that prey items are distributed randomly in the environment. However, prey can be found in relatively discrete patches. In an environment with prey distributed in patches the predator has to take the additional travelling time between patches into account. The predator is now facing a choice of where, when, and for how long to forage on a specific patch of prey items. The complex problem of foraging ecology in a patchy environment has a simple theoretical solution called the *marginal value theorem* (Charnov, 1976). The marginal value theorem defines the optimal time a predator should spend foraging in a single patch. The main assumption is that the resources of a patch are finite and deplete through the exploitation of a patch. The curve describing the overall energy gain as a function of time is therefore gradually levelling off. The optimisation needs to take the travel time between patches into account and the net energy gain $R(t)$ is therefore defined as follows:

$$R(t) = \frac{E(t)}{t+T} \ , \tag{3.11}$$

where E is the energy gain from foraging in a patch, T is the travel time between patches, and t is the time spent foraging in a patch. Thus the derivative is given by

$$R'(t) = \frac{(T+t)E'(t) - E(t)}{(T+t)^2} \ . \tag{3.12}$$

The derivative is zero when

$$E'(t) = \frac{E(t)}{T+t} = R(t), \tag{3.13}$$

which is a maximum as the second derivative

$$R''(t) = \frac{E''(t)}{T+t} \tag{3.14}$$

has the same sign as $E''(t)$ which is negative by definition of the patch model. Figure 3.4 shows the graphical solution of the marginal value theorem. Let the optimal foraging time be t^* then the slope of the tangent is given by $E(t^*)/(T+t^*)$ which is equal to $E'(t^*)$.

The marginal value theorem has found wide application in ecological models of animal behaviour for its simplicity and flexibility. Its successful application depends on the choice of currency and the definition of the net energy gain from foraging in patches. Examples range from mating choice to territory sizes (Parker, 1978; Krebs, 1980).

3.2.2 Optimal foraging theory and learning.

Both theories, OFT and MVT, proved successful in a wide range of applications (Stephens, 1986) but a later review also showed that foragers stayed significantly longer in patches than predicted (see Nonacs (2001) for a review of 26 quantitative studies testing quantitative predictions of the MVT).

Predators face the challenge of securing a sufficient energy intake in the face of changing and uncertain environments. Through the evolution of predator-prey interactions manifold mechanisms have emerged to avoid predation as discussed previously. Of particular interest to this thesis is aposematism as an anti-predator adaptation where secondary defences, commonly involving the possession of toxins or deterrent substances, are combined with conspicuous signals as warning flags.

There is a wide body of theory which addresses the emergence and evolution of aposematism (Ruxton et al., 2004; Yachi and Higashi, 1998; Broom et al., 2006; Leimar et al., 1986; Lee et al., 2011; Marples et al., 2005). However, the field of aposematism has a renewed interest in the role of the predator and details of the predator's aversive learning process. In particular, the role of aposematism in memory formation has been widely studied (Speed, 2000; Svádová et al., 2009; Skelhorn and Rowe, 2006; Johnston and Burne, 2008; Speed and Ruxton, 2005). As the selective agent, aversive learning is an important aspect of predator avoidance. It has been shown that predation of defended prey is rather a state dependent decision and predators can increase their attack rates on defended prey e.g. when particularly hungry (Barnett et al., 2007; Sherratt, 2003). There have been suggestions of an interaction of appetitive learning with aversive learning to explain predator behaviour of ingesting toxins in these

situations (Hagen et al., 2009).

An interesting perspective is to look at the predator and the consequences of aposematism in combination with aversive learning on the predator's diet and energy intake. In particular, the role of mimics in the evolution of aposematism and their effect on foraging is not very well understood (Gamberale-Stille and Tullberg, 2001; Lev-Yadun and Gould, 2007; Svádová et al., 2009; Holen, 2013). A predator may utilise sampling to distinguish between the toxic model and the mimic (Gamberale-Stille and Tullberg, 2001; Darst, 2006; Holen, 2013).

The traditional way of analysing and predicting foraging behaviour is the application of optimal foraging theory (OFT) which maximises the predator's net fitness per unit time (MacArthur and Pianka, 1966; Stephens and Krebs, 1987; Sih and Christensen, 2001). However, OFT has well known limitations: OFT usually fails to correctly predict foraging behaviour on mobile prey in complex environments (Sih and Christensen, 2001; Pyke, 1984; Perry and Pianka, 1997). It can be argued that OFT was never intended for predictions in the case of mobile prey and that the simple optimisation per unit time omits the uncertainty of more complex environments. There are models which address optimal foraging under the constraints of risk and uncertainty and previously extended OFT with learning (McNamara and Houston, 1985). The two main approaches to optimal behaviour in dynamic decision making are dynamic programming (DP) and stochastic optimal control methods (e.g. Bayesian decision theory) (Houston and McNamara, 1982; Stephens and Charnov, 1982; McNamara and Houston, 1985; Mangel and Clark, 1986; McNamara et al., 2006). Dynamic programming in particular has found broad application in behavioural ecology and has been used in models of dynamic decision making to identify optimal behaviour numerically (Clark and Mangel, 2000). A common factor of all these models is that they are *model based*: they depend on a representation of the environment in the form of a model developed from expert knowledge and the learning objective is to find the parameters which optimise the representational model.

Alternatively, reinforcement learning (RL) is a normative framework of rational decision making in a changing and complex environment, as introduced earlier in Section 3.1. RL combines the computational task of maximising rewards and the algorithmic implementation of learning without an explicit supervisory control signal (Mitchell, 1997; Sutton and Barto, 1998).

Neural correlates of behaving animals show that reinforcement signals in the brain represent the reward prediction error rather than a direct reward-reinforcement relation. Temporal difference (TD) learning reflects these insights by representing states and actions in terms of predictions about future rewards (Niv, 2009; Berns et al., 2001). Additionally, TD learning is *model-free*: the

environment is represented by moving targets rather than by a model and the learning objective is to iteratively update the targets towards its true values based on experience from interactions with the environment. TD learning has been widely used in artificial systems to choose appropriate actions in complex non-stationary environments. Furthermore, the computational theories of RL are increasingly supported by experimental data describing the activity of dopaminergic neurons, mediate reward-processing and reward-dependent learning (Schultz et al., 1997; Montague et al., 2004; Daw and Doya, 2006; Dayan and Niv, 2008).

The further discussion of my model is structured as follows: In the next two sections I apply a TD learning algorithm in a model of predator interactions with conspicuous prey to gain insights on how aversive learning influences foraging in uncertain environments, and present the results. Next I discuss the main findings and discuss similarities and differences to the optimisation approach of traditional OFT. In particular, I will compare TD learning with McNamara et al. (McNamara and Houston, 1985) and Sherratt (Sherratt, 2003).

3.2.3 Model definition.

In my model the predator interacts with its environment to find an optimal foraging strategy to optimise its rewards. The predator's environment offers a stable background of alternative food sources. Additionally, the predator has the choice to include a conspicuous looking type of prey into its diet. However, the conspicuous prey population may consist of an aposematic model species and a Batesian mimic species.

The predator is not able to distinguish models and mimics based on their appearance and utilises experience to learn the optimal foraging behaviour. The model is presented in Figure 3.5.

I term the action of falling back on the alternative background food sources as $u = 0$ and the action of attacking conspicuous prey as $u = 1$.

I assume the population of conspicuous prey consists of a fraction p of Bateysian mimics and a fraction $1 - p$ of defended models. The reward signal for the alternative stable background food source is $r_{k+1} = \{1 \,|\, u = 0\}$. The reward signal for ingesting a mimic individual is $r_{k+1} = \{2 \,|\, u = 1,\, i = \text{mimic}\}$ and $r_{k+1} = \{1 - t^2 \,|\, u = 1, i = \text{model}\}$ for ingesting a model individual with toxicity t. These reward signals do not have to necessarily represent fitness related entities (Pyke, 1984). In this model I simply assume mimics to be rewarding and that toxicity has a non-linear effect on the reward, which seems like a reasonable assumption.

I consider two different cases (Figure 3.5):

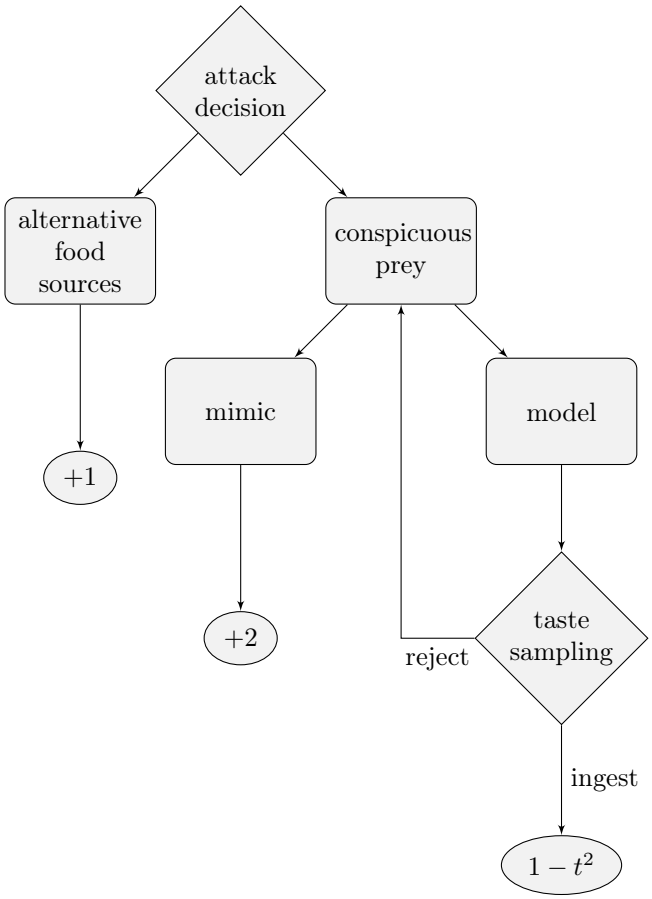


Figure 3.5: The predator's interaction with its environment and possible reward signals. The predator has the ability to recognise toxic models by taste-sampling. t stands for the toxicity of defended models.

1. The predator has the ability to use taste-sampling to distinguish models from mimics assuming that the model's toxicity t operates as a clue to the predator. This foraging strategy is also called *go-slow behaviour* (Guilford, 1994). The probability of rejecting a model based on taste-sampling is given as follows:
$$d(t) = 1 - \frac{1}{1 + d_0 * t}. \tag{3.15}$$

2. The predator has no ability to distinguish mimics and models and the encounter is solely frequency dependent i.e. $d_0 = 0$ in Equation (3.15).

Based on the growing understanding of learning at the computational and neural level I use Temporal Difference (TD) learning to implement the predator's aversive learning: in particular, I use Q-learning as introduced in Section 3.1.3 (Watkins, 1989). The predator utilises experience to infer the optimal foraging behaviour derived from the action-value function (3.2) of previous encounters with the conspicuous prey. At each iteration k the learning individual finds itself in state x_k of its environment, accordingly, x_k is the current composition of the conspicuous prey population in this model. The actual learning process targets the individual's value prediction following action u_k (respectively, choosing the alternative food source or the conspicuous prey for foraging) in state x_k as described by the action-value function (3.2).

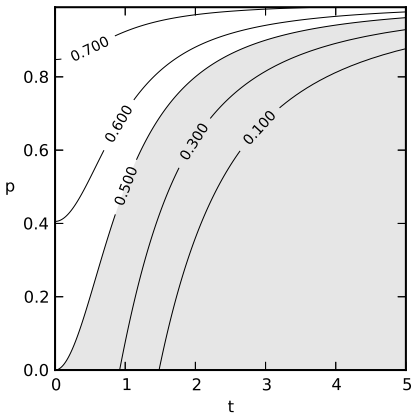
I assume the environment to be uncertain with non-stationary parameters t and p over a predator's lifespan and, therefore, it requires a precondition of continuous exploration of the environment. Thus, this model uses Gibb's soft-max policy which is the stochastic policy of taking action u in state x as defined previously in Equation (3.6).

3.2.4 Results.

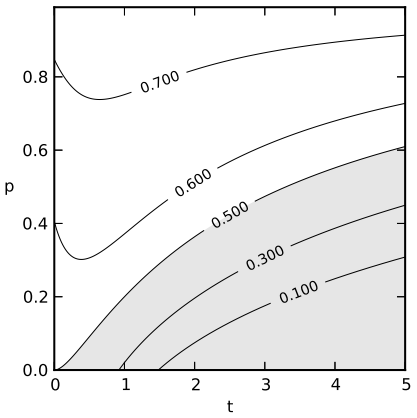
In the case of the predator being unable to distinguish models from mimics ($d_0 = 0$) the average reward signal is soley frequency dependent and given as

$$R = \begin{cases} 1 & \text{if } u = 0 \\ 2p + (1 - t^2)(1 - p) & \text{if } u = 1. \end{cases} \tag{3.16}$$

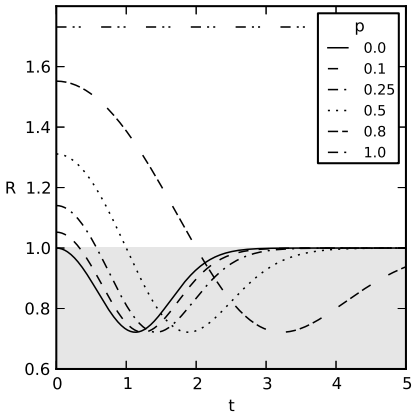
If the predator utilises taste-sampling it can distinguish models from mimics based on the model's toxicity and will not ingest the toxic model with probability $d(t)$ given in (3.15). After the predator rejects a conspicuous prey individual it will stay in the locality and forage for another conspicuous prey individual. The average reward signal incorporating taste sampling derives from the geometric



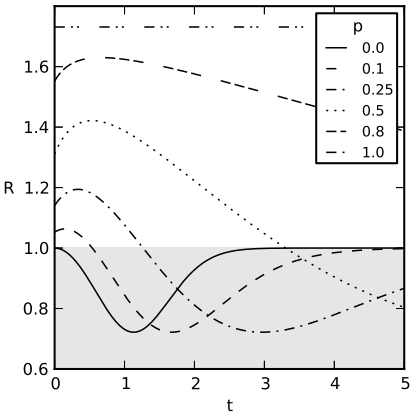
(a) Predator attack probability (π) of conspicuous prey without taste-sampling ($d_0 = 0$) following Gibb's soft-max policy (3.6). The shaded area indicates aversive toxicity.



(b) Predator attack probability (π) of conspicuous prey utilising taste-sampling ($d_0 = 3$) (3.15) following Gibb's soft-max policy (3.6). The shaded area indicates aversive toxicity.



(c) The predator's average reward (R) from interacting with its environment without taste-sampling ($d_0 = 0$). The shaded area indicates suboptimal rewards due to foraging on aversive prey.



(d) The predator's average reward from interacting with its environment utilising taste-sampling ($d_0 = 3$). The shaded area indicates suboptimal rewards due to foraging on aversive prey.

Figure 3.6: The results of a predator foraging in an environment offering an aposematic food source which uses Q-learning to derive the optimal foraging strategy. All results use a discount rate $\gamma = 0.5$. t stands for the toxicity of models and p for the fraction of mimics in the conspicuous population.

series and is given as follows:

$$R = \begin{cases} 1 & \text{if } u = 0 \\ 2p\frac{1}{1-(1-p)d(t)} + (1-t^2)(1-p)\frac{(1-d(t))}{1-(1-p)d(t)} & \text{if } u = 1. \end{cases} \tag{3.17}$$

To obtain the optimal diet we find the correct, discounted action-value function Q^* by solving the TD learning problem:

$$0 = R + \gamma \max_{u_{k+1}} Q(x_{k+1}, u_{k+1}) - Q(x_k, u_k), \tag{3.18}$$

which is in this model a system of two equations ($u = \{0, 1\}$) with two unknowns ($Q(x, u = 0), Q(x, u = 1)$) for each possible state x of the environment. The choice of Q-learning as an off-policy control method with a greedy behaviour policy defines $\max_{u_{k+1}} Q(x_{k+1}, u_{k+1})$ and allows an analytical solution.

Figures 3.6a and 3.6b show the probability of an experienced predator attacking conspicuous prey based on the frequency of mimics (p) and the model's toxicity (t). We define aversiveness as $\pi(u = 1) < 0.5$ with the threshold toxicity (t^*) given in (3.7) for which conspicuous prey becomes aversive and $R(u = 0, t^*) = R(u = 1, t^*)$ holds, as follows:

$$t^* = \begin{cases} \sqrt{\frac{p}{1-p}} & \text{if } d_0 = 0 \\ \frac{\sqrt{p^2 d_0^2 - 4p^2 + 4p} + pd_0}{2(1-p)} & \text{otherwise.} \end{cases} \tag{3.19}$$

We see that taste-sampling lowers the aversiveness of defended conspicuous prey when mimics are present.

Figures 3.6c and 3.6d show the average reward (R) of an experienced predator. Mimics increase the average reward of the predator through increased foraging on non-aversive conspicuous prey. Conversely, increasing toxicity of the models reduces the average reward for the predator until the increasing toxicity intake from mistakenly ingested models becomes aversive.

3.2.5 Discussion

I apply Q-learning to the problem of optimal foraging behaviour of an experienced predator in an uncertain environment. My motivation lays in the recognised importance of aversive learning in aposematism and the difficulties of the classical OFT approach to predict foraging behaviour on mobile prey (Sih and Christensen, 2001). In the case of mobile prey additional factors of prey handling and uncertainty need to be considered, making the OFT model increasingly complex (Holen, 2013). Instead, reinforcement learning offers a normative framework of rational decision making in a changing and complex environment

with growing evidence of neural correlates.

The TD learning based approach puts the emphasis on experience including discounted future rewards and requires exploration of the action space. This is fundamentally different to the OFT models of net fitness maximisation per unit time. It has been long argued that a learning animal cannot be foraging optimally and vice versa due to the exploration exploitation trade-off (Ollason, 1980).

I hypothesise that a non-stationary environment introduces great uncertainty on the prey-population's parameters t and p which selects for learning in evolving predators to adapt quicker to their changing environment. Evidence for this claim has to come from an evolutionary model and is subject to future work. To coincide widely with the original OFT methodology, I assume that the learning process is sufficiently faster than the frequency of change of the environment to concentrate solely on the experienced predator. At the core of Q-learning is the approximation of the action-value function (3.2) using an iterative update rule based on moving targets (3.4). However, this model allows the analytical solution of the action-value function and excludes the iterative learning phase of Q-learning. Furthermore, I assume that the conspicuous prey inhabit a distinct locality. These assumptions allow me to solve the TD learning problem directly (3.18) and I present the policy a predator adopts through Q-learning under a constraint of continuous exploration.

In the context of previous foraging models which incorporated learning, the presented learning methodology is model-free. Relevant models, among others, are from McNamara and Houston (1985) and Sherratt (2003). McNamara's learning rule describes a Monte Carlo method using past events to learn the maximum possible long-term rate as defined by the marginal value theorem (Charnov, 1976). It uses discounted experience from past interactions with the environment to optimize a current parameter estimation. The corresponding concept in TD learning is termed *eligibility trace* and is bridging TD learning and Monte Carlo methods. Eligibility traces can make TD learning more efficient but as I exclude the iterative learning phase it has no application in this model. Nevertheless, TD learning is conceptually different as its learning objective is based on bootstrapping future rewards rather than optimising the current estimate of a parameter from past events.

Sherratt's model (Sherratt, 2003) uses Bayesian learning based on dynamic programming. The learning objective is to infer the Bayesian posterior mean estimate of the fraction of defended prey in an unknown population from past experience. The model uses Beta distributions in the Bayesian inference to represent an assumed underlying binomial distribution of defence in a group of prey. The main assumption for the application of dynamic programming

is the existence of a finite time horizon where the predator ceases attacking completely. Sherratt's model provides an optimal sampling strategy for novel prey populations with constant values for cost and benefit of an attack. However, the model cannot provide optimal foraging policies in changing populations or when defence is not just binomially distributed.

I conclude that TD learning is a new approach to optimal foraging in dynamic environments where cost-benefit values of attacking prey do not necessarily follow simple distributions. The model-free objective of TD learning makes it an ideal method for learning in complex and dynamic environments where parameters are subject to constant change.

My model confirms expected results such as that mimics in general lower the aversiveness of the conspicuous prey population and undermine aposematism. Nevertheless, highly toxic models can sustain aversion even for high frequencies of mimics especially in predators not utilising taste sampling. Importantly, mimicry requires mixing with model prey in the case of a learning predator as an experienced predator could utilise spatial information about the prey populations to discriminate between models and mimics. However, it requires exploration for a predator to gain insights about its environment and to form aversive memory. Therefore, even an aversive prey population experiences some level of predation. My model predicts that a taste-sampling predator increases its attack rate on mixed conspicuous prey populations in the case of moderately defended models and rewarding mimics. The taste-sampling predator gains increased rewards from moderately defended models as it allows for better discrimination of models and mimics. (In more strongly defended prey the increasing cost of mistakes will outweigh the benefits of improved discrimination of prey.) This is a contrary finding to (Holen, 2013) in which mimics benefit from moderately defended models. This difference is founded on the representation of toxins as recovery time in the OFT maximisation approach and the lack of occasional exploration and consequent exposure to models in order to maintain aversion for highly toxic models.

An interesting paradox is the foraging behaviour on aversive prey which reduces the reward for the predator further before recovering through increasingly falling back on alternative background food sources, (the adopted attack policy for certain parameters results in an average reward R which lays in the shaded area in Figures 3.6c and 3.6d, and is suboptimal). This is a result of the conflicting reward signals of mimics and models and the necessity of exploration of the action space in the face of uncertainty for successful aversion formation. Additionally, an increasing frequency of mimics slows the switching to alternative food sources through further extended uncertainty. Similar results have been observed in counter conditioning and operant conflict situations (Williams

and Barry, 1966; Blaisdell et al., 2000; Mazur and Ratti, 1991; Matsushima et al., 2008). My model predicts a fixed amount of average long term toxicity intake which a predator tolerates motivated either by the higher reward signal of ingested mimics or as a consequence of uncertainty. (Although the toxicity of immediate rewards which induce switching to alternative food sources depends on the amount of mimics and the specific rewards, see Equation (3.19) and Figures 3.6a and 3.6b, the average reward function described in Equations (3.16) and (3.17) has a fixed minimum as presented in Figures 3.6c and 3.6d). This foraging behaviour on aversive prey for a specific parameter space is conditionally suboptimal in a stationary environment (even if only during an individual's lifetime) but I note that a) it reflects what real animals do, and b) it is a good policy precisely because environments are inherently uncertain. Finally, the switching behaviour between food sources shows so called contrast effects and depends on the initial toxicity of the model population and not solely on the absolute change of toxicity. With respect to the same absolute toxicity change, the predator switches faster to the alternative food source if the environment becomes aversive than it would incorporate the conspicuous food source if the environment becomes rewarding. Summarising, the main conclusions are as follows:

- TD learning is a suitable approach to optimal foraging in changing environments.
- Even aversive prey experience some level of predation as part of the predator's aversive memory formation.
- Taste-sampling lowers the effective aversiveness of conspicuous prey if mimics are present.
- Intermediate toxicity of aposematic models increases the predator's foraging on conspicuous prey through increased discrimination from taste-sampling and higher average rewards when mimics are rewarding.
- The conflicting reward signals from mimics and models cause uncertainty and conditionally suboptimal foraging behaviour on aversive prey.
- The uncertainty is linked to a fixed amount of average toxicity intake which predators tolerate in order to forage on rewarding mimics before switching to mediocre background food sources.
- Taste-sampling extends the range of parameters where suboptimal foraging occurs.

- Switching between food sources shows contrast effects and depends on the initial toxicity of models and not solely on the absolute change of toxicity in the environment.

3.3 When does learning matter?

Through evolution, animals are generally very well-adapted to the environment in which they find themselves. Phenotypic plasticity, the ability to adapt the phenotype in response to different conditions of the environment, allows for suitable adaptations even in the face of changing environments (Pigliucci, 2001). Thus both physical abilities and behaviours of animals are generally appropriate to their environment. Nevertheless many animal behaviours are not solely genetically determined, though some are, but the response of the animal's learning capabilities.

Hence a key question arises: under which conditions is the ability to learn beneficial to animals? To answer this question I will focus next on a deceptively simple model of learning by which individuals learn to associate events that occur together, for instance two stimuli, a stimulus and a response, or a response and its outcome (Mackintosh, 1974; Pearce, 2013). As previously indicated, associative learning is a fundamental cognitive process observed across species (including mollusks, insects, birds and mammals) (Carew et al., 1983; Macphail, 1982) that affects a wide variety of behaviours ranging from colour recognition (Carew et al., 1983) and spatial representation (Albasser et al., 2013), to causality judgements (Shanks, 1995) and goal-directed behaviour (Valentin et al., 2007). Of course, in addition to associative learning animals use other types of learning (e.g. social learning or perceptual learning) and ontogenetic mechanisms (e.g. habituation and phenotypic plasticity) to adapt their behaviour to the environment. Nonetheless, the pervasiveness and relevance of associative learning makes it the ideal candidate to investigate when learning is most effective. Within the wider consideration of learning as a form of adaptation to changing environments, I am particularly interested in associative learning in decision-making tasks.

From a biological perspective, learning is a mechanism for rapid adaptation (modification) of behaviour during the individual's lifetime and a distinct adaptation to changing environments in particular (Johnston, 1982). The main line of argument is that learning incurs some cost from which it follows that a constant environment should select for a genetically fixed pattern of behaviour over learned behaviour. But the relationship of learning and evolution is complex and an important aspect of learning is environmental predictability (commonly also referred to as regularity) (Staddon and Simmelhag, 1971). Clearly, there

is nothing to learn in an environment which is absolutely unpredictable. So far both factors, environmental change and regularity, have been discussed in the literature as the selective factors in the evolution of learning. A contradiction at first sight, but in the light of extreme environments a solution to the paradox would be that learning is in fact an adaptation to intermediate levels of environmental change (Johnston and Turvey, 1981).

As well as considering which aspects of the environment make a learning strategy beneficial, questions regarding the relationship between evolution and learning are of interest. To recap, I use reinforcement learning (RL) as a normative framework of associative learning and rational decision making in changing environments. RL combines the computational task of maximising rewards and the algorithmic implementation of learning without an explicit supervisory control signal (Sutton and Barto, 1998). In RL the environment is represented by moving targets rather than by a model and the learning objective is to iteratively update the targets towards their true values based on experience from interactions with the environment. Each time an individual performs an action in some state it receives a real-valued reward that indicates the immediate value of this state-action transition. Unlike in supervised learning, the learner must discover which actions yield the most reward by exploiting and exploring their relationship with the environment. These two characteristics, trial and error search and delayed rewards, are the two most important features of reinforcement learning which raises the problem of an optimal exploitation-exploration trade-off.

The last decade has seen a proliferation of research on the neural and psychological mechanisms of RL (Dayan and Daw, 2008; Doya, 2007; Maia, 2009; Niv, 2009; Rangel et al., 2008; Schultz, 2002; 2007). In particular, RL predictions are increasingly supported by experimental data describing the activity of dopaminergic neurons, mediate reward processing and reward dependent learning (Schultz et al., 1997; Montague et al., 2004; Daw and Doya, 2006; Dayan and Niv, 2008).

In this model I present fitness distributions of learning individuals in an environment designed to answer questions surrounding the initial evolution of RL mechanisms. I will compare the learning strategy with a mutating population of individuals with fixed types to investigate the cost of learning and the effects of environmental parameters on the benefits of learning. Similarly to the motivation of evolutionary games (Section 1.1.3) I consider the evolutionary dynamics as a separate layer to the model presented here. Therefore, I separate the selection process as part of the evolutionary dynamics from the model and analyse a mutation and learning process as a population's means of creating phenotypic variance. This allows the focus on static fitness distributions

which should not alter the outcome of the analysis in the biological context of finding the best strategy for the environment presented in this model. I note that this model is not a complete description of the dynamics of evolution but an adequate simplification motivated by game theoretical models. The model will allow the reasoning about the factors which might have driven the initial evolution of RL. Additionally, it allows important insights into the differences of phenotypic variance caused by mutation and learning independently of the common arguments around the speed of adaptation.

3.3.1 Model definition.

The following results of this model are a continuation of the previous model (Section 3.2.3) introducing reinforcement learning, i.e. Q-learning, to models of predator-prey interactions (Teichmann et al., 2014a). The previous model investigated the effects of aversive learning in a changing environment on an experienced predator's diet choice and energy intake. In this model I describe fitness distributions of learning individuals in the context of changing environments more generally and compare them with a simplistic mutation process in order to gain insights into the relationship between evolution and learning.

In this model the learning individual again uses Q-learning as an implementation of reinforcement learning (Watkins and Dayan, 1992). I choose Q-learning for the simplicity of its implementation of real-time error-correction learning and as it is increasingly supported by both behavioural and neural data. To recap, in Q-learning an individual uses experience following its interactions with the environment to infer optimal decisions. The learning individual utilises an action-value function to build a representation of the environment which describes the expected future payoff following a specific action in a specific state of the environment. The individual then minimises the error of the action-value function's future payoff prediction building on a growing amount of evidence from past trial-and-error interactions with the environment. These future payoff predictions are discounted by a γ factor, indicating the uncertainty of forthcoming events. Furthermore, the prediction error is modulated by a learning rate α, that is, how quickly (not necessarily how correctly) the animals learn. Finally, the individual translates the payoff predictions of the action-value function into a decision following a stochastic policy, in particular Gibb's soft-max policy. I refer to the previous introduction of Q-learning for further details on the Q-learning algorithm in Section 3.1.3.

I compare a population of learning individuals with a population of mutating individuals with fixed phenotypes. An individual of the population following the mutation strategy has a genetically determined decision policy chosen randomly

from a uniform distribution at the beginning of each generation. The important differences between the two populations are: (i) the process of simplistic mutation is random and not adaptive operating on fixed phenotypes and (ii) the learning strategy is adaptive but incurs the cost of exploration. I do not include any selection in my model as I am purely interested in the fitness distributions of both populations in a changing environment which makes the mutation frequency of the mutation process irrelevant to my model. A population dynamical approach would complicate the analysis unnecessarily as it would add further aspects of mortality, resource depletion, or interactions between individuals. Similarly to models of evolutionary games, I treat the underlying population dynamics as a separated layer and assume that it does not alter the biological relevant outcome of the analysis. I will show that it is not necessary to analyse complete population dynamics to gain insights into the benefits of learning.

I define the environment for my analysis to be stationary and ergodic and to consist of two options, a certain and an uncertain one, as shown in Figure 3.7. In my definition the certain option gives a constant fitness payoff $R = 0$ and the uncertain option returns a uniformly distributed fitness payoff $g(R)$ with the mean being zero. The value of 0 for the fitness of the constant option and the mean of the variable option is chosen for simplicity. I note that it is possible to add an arbitrary constant to either and not qualitatively change the results (to see why this is reasonable here, and when it is not, see the discussion on long-term fitness effects in Section 3.3.2). The environment is parametrised with (i) β being the number of changes of the uncertain option per generation time, (ii) ϵ being the extent of the absolute fitness change per generation time, and (iii) l being the length of the generation in interactions with the environment. The term regularity refers to the predictability of an environment within models of learning. In my model the learning individual cannot draw from any secondary source of information such as a correlation between environmental states. Therefore, I define regularity of the environment in my model by the number of interactions available for exploitation which is given by the number of interactions with a given environmental state. Hence, the regularity of the environment in my model is defined by a combination of β and l. Accordingly, the environment becomes increasingly irregular with greater values of β and smaller values of l as an individual has less interactions with a given environmental state before an environmental change occurs. I define the limits $[-a, a]$ of the Uniform distribution $g(R)$ as follows:

$$a = \frac{3}{2\beta}\,\epsilon, \tag{3.20}$$

where ϵ is the absolute average fitness change per generation derived from the

Symbol	Definition
R	The fitness payoff following an interaction with the environment.
F	The fitness of an individual at the end of a generation.
$\hat{F}$	The scale-free fitness of an individual.
l	The length of a generation in interactions with the environment.
β	The number of environmental changes per generation time affecting the regularity of the environment.
ϵ	The extend of environmental change in absolute average fitness change per generation time.
α	The learning rate of the learning individuals.
γ	The discount rate of future payoffs of the learning individuals.

Table 3.1: Parameters and their definition

triangular distribution $h(|R_i - R_{i+1}|)$ of the absolute difference of the uniformly distributed fitness payoff $g(R)$ as illustrated in Figure 3.7b. I assume that an increased frequency of environmental change β results in smoother and less pronounced single changes as reflected in the definition of Equation (3.20).

The fitness F of an individual is the sum of the fitness payoff from the interactions with the environment

$$F = \sum_{t=1}^{l} R_t. \tag{3.21}$$

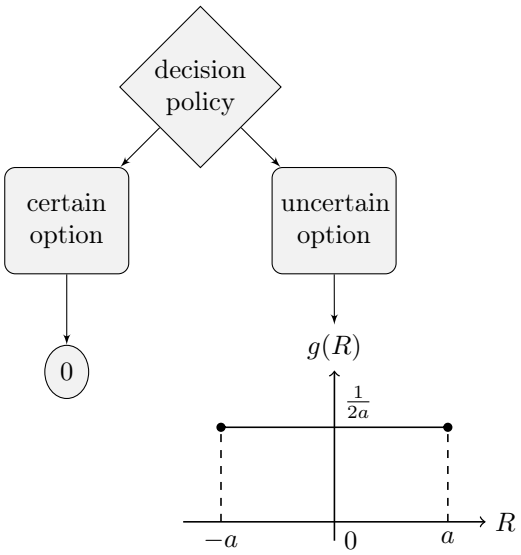
3.3.2 Results.

I present the distributions of $n = 5000$ generations interacting with their environment using a scale free variant of the fitness as follows:

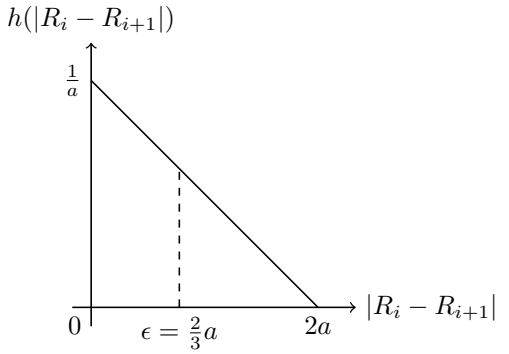
$$\hat{F} = \frac{\beta}{l\,\epsilon} F, \tag{3.22}$$

which will allow a more intuitive comparison of the two populations in respect to the parameters of the environment.

I will present the results for each population respectively as follows in the form of box-plots. The box is bounded by the first and third quartiles of the fitness distributions. The inner band shows the second quartile, the median of the fitness distributions. The whiskers of the box plots comprise 1.58 times the interquartile range (IQR). The remaining data is shown as outliers with a simple cross. The notch around the median is defined as $1.58 \times \mathrm{IQR}/\sqrt{n}$ and gives



(a) The two options of the environment with the certain option being equal to zero and the uncertain option following a Uniform fitness payoff distribution $g(R)$ with limits $-a$ and a as given by Equation (3.20).



(b) The distribution of absolute fitness change follows a triangular distribution $h(|R_i - R_{i+1}|)$ with $\epsilon = (2/3)a$ being the average absolute fitness change given the uniform distribution of fitness payoff $g(R)$ with $\beta = 1$.

Figure 3.7: The environment for my analysis with the choice of a certain and an uncertain option.

roughly a 95% confidence interval for the median of the fitness distributions.

Mutation strategy.

Figures 3.8a and 3.8b show the main characteristic of the mutation process: as the process is random and not adaptive it is independent of the number of interactions with the environment per generation l and independent of the extent of the environmental change per generation ϵ. The fitness distributions are also symmetric around fitness neutrality with mean zero. Additionally, the parameters α and γ do not apply to the mutation process.

Figure 3.8c shows the effects of the frequency of environmental changes β. The fitness distribution is unaffected for frequencies $\beta \leq 1$, i.e. when mutations occur more frequently than changes in the environment. If the frequency of environmental changes exceeds $\beta = 1$ the fitness distribution of the population of mutating individuals becomes increasingly narrow. This is a direct result of the mutation process being non-adaptive and therefore it is less likely that individuals are well suited (or poorly suited) for a number of consecutive environmental states.

Learning strategy.

The population of learning individuals uses Q-learning in order to adapt to the current state of the environment. This requires the precondition of exploration which is the sole cost of learning in my model. Additional costs of learning are difficult to quantify and I assume that during the initial evolution of learning the additional costs were probably relatively small (Johnston, 1982; Mery and Kawecki, 2004).

Figure 3.9a shows that learning requires certain environmental conditions to be beneficial: as an adaptive strategy learning benefits from a changing environment (Figure 3.9a: 1 vs. 2). The cost and benefits of exploration in the learning population can be seen (Figure 3.9a: 1) versus the mutating population (Figure 3.8a) where the learning strategy cuts off both tails of the fitness distribution and does not produce the outliers which I find in the mutation process. Additionally, learning benefits from longer generation times to exploit experience (Figure 3.9a: 3).

Figure 3.9b shows that the learning strategy is unconditionally affected by the frequency of environmental change β compared to the population of mutating individuals which is unaffected for $\beta \leq 1$ (Figure 3.8c). The effect of β on the fitness distribution of the learning individuals is not linear, and there are multiple factors underlying this effect. In environments with only very rarely occurring changes an increasing majority of the population benefits from learn-

ing (Figure 3.9b: 1). At first, an increasing frequency of environmental change increases the fraction of individuals benefiting less from learning with the fitness distribution developing a more pronounced tail of learning individuals having negative relative fitness (Figure 3.9b: 1-5). Nevertheless, environmental change benefits learning at the same time with the median of the population increasing (Figure 3.9b: 4). Secondly, a further increase of β results in the cost of consecutive exploration and consequent errors outweighing this initial benefit and the distribution increasingly aligns with the fitness distribution of the mutating population. Finally, learning does not provide any benefits in highly irregular environments interfering with any possibility of exploitation. Therefore the only difference in the fitness distributions is the shorter tails of the learning strategy, which is the result of continuous exploration (Figure 3.9b: 8 vs. Figure 3.8c: 6).

Figure 3.10 shows the effects of the extent of environmental change ϵ in combination with the frequency of environmental change β on the fitness distribution of learning individuals. It is clear that there are some effects specific to the two factors but additionally there is also an interaction between environmental change and regularity. I have already discussed the individual effects of β relating to Figure 3.9b. Regarding the effects of ϵ I can see that in environments with small values of absolute environmental change throughout a generation the fitness distribution of the learning individuals aligns with the fitness distribution of the population of mutating individuals. A learning individual prioritises continuous exploration if the environmental change is small which is a consequence of learning being an adaptation to changing environments. The shorter tails in the fitness distribution of the learning population compared to the mutating population are the result of this continuous exploration as discussed previously (Figure 3.10a: 1 vs. Figure 3.8a and Figure 3.10b: 1 vs. Figure 3.8b). An increase of ϵ has beneficial effects for all learning individuals in the population as learning requires a certain extent of environmental change to exploit. A further increase of ϵ makes mistakes during exploration more expensive which can potentially neutralise the benefits of exploiting beneficial states of the environment. The important difference between a severe extent of environmental change (ϵ) and an irregular environment (β) is that mistakes in the case of ϵ are extremely aversive and stop any further costly exploration. This is the reason that the fitness distribution of learning individuals in violently changing environments loses the negative tail compared to rapidly changing environments (Figure 3.10a: 11 vs. Figure 3.9b: 8). The combined effect of frequency and extent of environmental change shows that there is a specific combination of these two factors which hugely benefit the learning strategy (Figure 3.10b: 7).

Figure 3.11 shows that the fitness distribution of the learning population is, within a meaningful range, independent of the learning rate α and the discount

factor γ. This result should not be misinterpreted: there are specific values of α and γ best suited for achieving optimality in a specific state of the environment. But within a changing environment the distribution of fitness is independent of the specific choice of α and γ.

Long-term fitness effects.

In the context of population dynamics it becomes important to take potential long term fitness effects into account. In changing environments especially, the long term fitness of a population is largely dependent on future states of the environment.

Suppose that for some population, an individual in generation i lives for l time steps, and acquires reward $c+R_{i,j}$ at the jth of these, where c is a constant, and $R_{i,j}$ are independent, identically distributed random variables with mean 0 and variance σ^2. Such individuals thus have fitness

$$\begin{aligned} F_i &= lc + \sum_{j=1}^{l} R_{i,j} \\ &= lc\left(1 + \frac{1}{c}\bar{R}_i\right), \end{aligned} \tag{3.23}$$

where

$$\bar{R}_i = \frac{1}{l}\sum_{j=1}^{l} R_{i,j}. \tag{3.24}$$

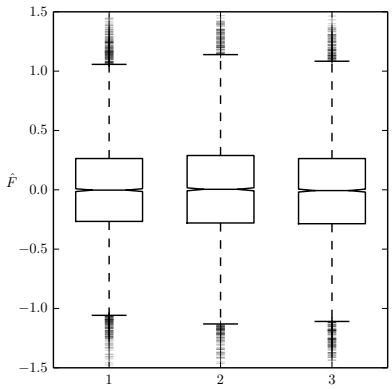
The long-term population displacement is geometric and not linear following the continuous product of the fitness of the parental and offspring populations. The logarithm of this long term fitness is given by

$$\begin{aligned} \ln\left(\prod_{i=1}^{n} F_i\right)^{1/n} &= \frac{1}{n}\sum_{i=1}^{n} \ln F_i \\ &\approx \ln(lc) + \frac{1}{c}\bar{R}_i - \frac{1}{2c^2}(\bar{R}_i)^2, \end{aligned} \tag{3.25}$$

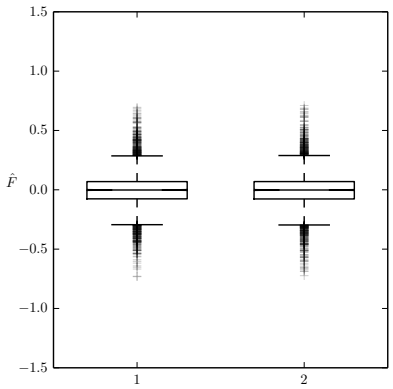
which has expectation

$$\ln l + \ln c - \frac{1}{2c^2}\frac{\sigma^2}{l}\ . \tag{3.26}$$

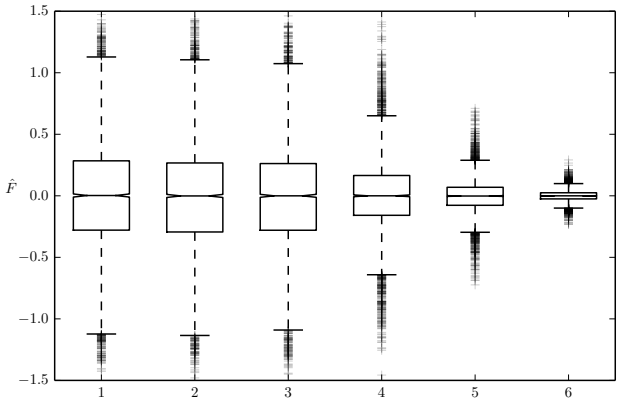
From above it is clear that the larger the value of c, and the smaller the value of σ^2, the higher the fitness of the population. For sufficiently large l, c representing the arithmetic mean is the dominant term. Nevertheless, two populations can have the same arithmetic mean fitness and very different variability of their fitness in a changing environment. If the arithmetic mean fitness term is the same for two populations, the second term comes into play, and the population



(a) The fitness distribution of the mutation strategy is independent of the extend of environmental change (ϵ) and the numbers of interactions per generation time (l). (1) $\epsilon = 0.1$, $\beta = 1$, and $l = 1000$. (2) $\epsilon = 100$, $\beta = 1$, and $l = 1000$. (3) $\epsilon = 10$, $\beta = 1$, and $l = 10$. Distributions are not significantly different using Kolmogorov-Smirnov test with all $p > 0.1$.

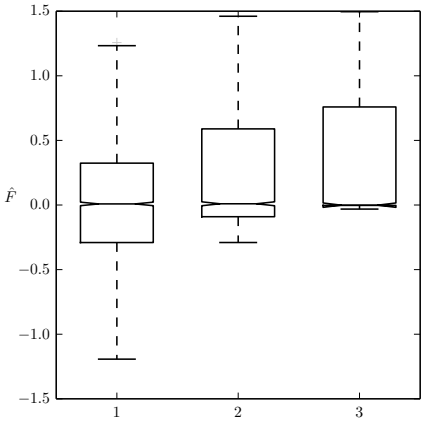


(b) Fitness distributions scale equally with the number of changes in the environment (β) independently of ϵ and l. (1) $\epsilon = 10$, $\beta = 10$, and $l = 1000$. (2) $\epsilon = 1$, $\beta = 10$, $l = 10$. Distributions are not significantly different using Kolmogorov-Smirnov test with all $p > 0.1$.

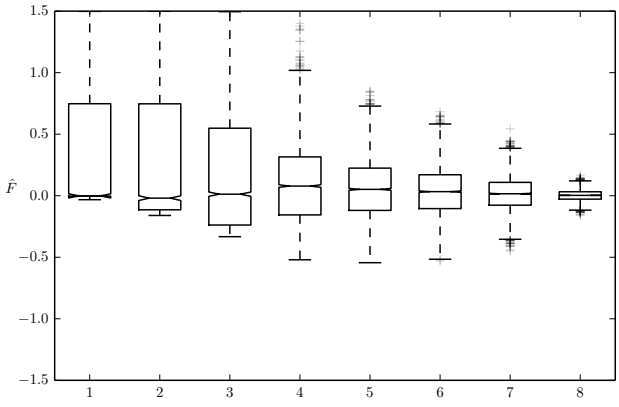


(c) Fitness distributions become narrower with increasing number of environmental changes $\beta > 1$ per generation time. Distributions are not significantly different for $\beta \leq 1$ using Kolmogorov-Smirnov test with all $p > 0.1$. (1) $\beta = 0.1$, (2) $\beta = 0.5$, (3) $\beta = 1$, (4) $\beta = 2$, (5) $\beta = 10$, and (6) $\beta = 100$. All cases have $\epsilon = 1$ and $l = 1000$.

Figure 3.8: Scale-free fitness distributions of the mutation strategy all with $n = 5000$ generations.

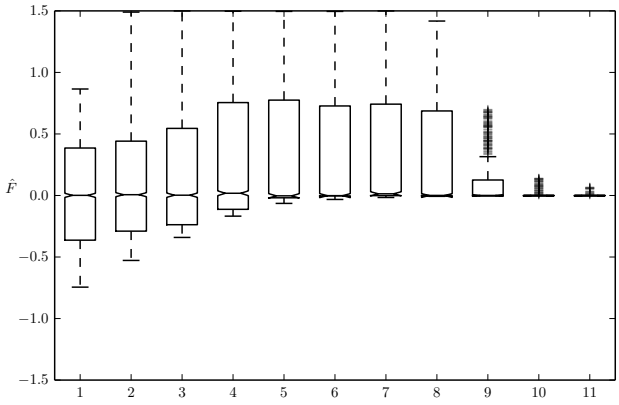


(a) Learning requires certain environmental conditions to be beneficial: firstly, learning requires environmental changes: (1) $\epsilon = 0.1$, $l = 10$ vs. (2) $\epsilon = 10$, $l = 10$. Secondly, learning benefits from longer generation times: (2) vs. (3) $\epsilon = 10$, $l = 1000$. All $\beta = 1$, $\alpha = 0.5$, and $\gamma = 0.9$.

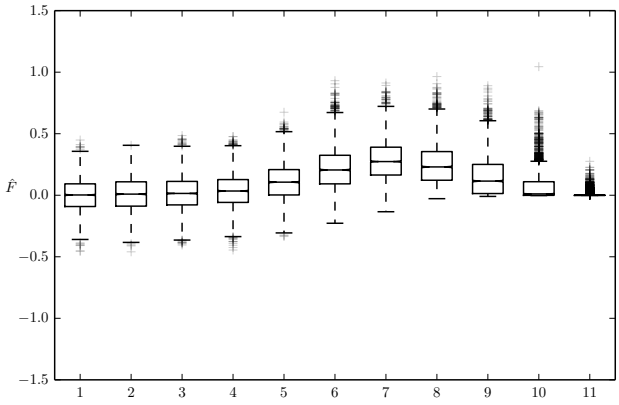


(b) Learning incurs the cost of exploration and the benefits of learning diminish in unreliable environments. (1) $\beta = 0.1$, (2) $\beta = 0.5$, (3) $\beta = 1$, (4) $\beta = 2$, (5) $\beta = 3$, (6) $\beta = 5$, (7) $\beta = 10$, and (8) $\beta = 100$. All $\epsilon = 1$, $l = 1000$, $\alpha = 0.5$, and $\gamma = 0.9$.

Figure 3.9: Scale-free fitness distributions of the learning strategy presenting isolated effects of environmental parameters. All cases have $n = 5000$ generations.

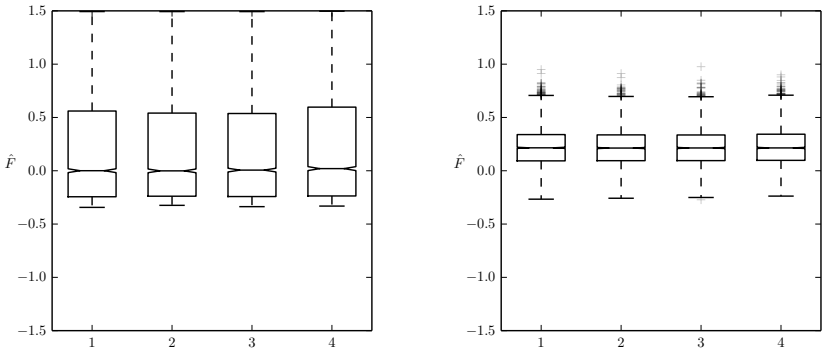


(a) $\beta = 1$



(b) $\beta = 10$

Figure 3.10: Scale-free fitness distributions of the learning strategy depending upon the extent of environmental change (ϵ). Learning benefits from a certain extend of environmental change but too severe changes incur a high cost of mistakes during the required exploration. Additionally, there is a combined effect of regularity and change. (1) $\epsilon = 0.1$, (2) $\epsilon = 0.5$, (3) $\epsilon = 1$, (4) $\epsilon = 2$, (5) $\epsilon = 5$, (6) $\epsilon = 10$, (7) $\epsilon = 20$, (8) $\epsilon = 50$, (9) $\epsilon = 100$, (10) $\epsilon = 500$, and (11) $\epsilon = 1000$. All cases have $n = 5000$ generations, $l = 1000$, $\alpha = 0.5$ and $\gamma = 0.9$.



(a) $\beta = 1$ and $\epsilon = 1$.

(b) $\beta = 10$ and $\epsilon = 10$.

Figure 3.11: Scale-free fitness distributions of the learning strategy showing the independence of α and γ. (1) $\alpha = 0.1$ and $\gamma = 0.1$, (2) $\alpha = 0.9$ and $\gamma = 0.1$, (3) $\alpha = 0.9$ and $\gamma = 0.1$, and (4) $\alpha = 0.9$ and $\gamma = 0.9$. All cases have $n = 5000$ generations and $l = 1000$. Distributions are not significantly different using Kolmogorov-Smirnov test with all $p > 0.1$.

with the smaller variance has the higher fitness. For smaller l, the contribution of the variance may be sufficiently large to reverse the effect of larger arithmetic mean fitness, and mean that the population with the smaller arithmetic mean is actually the fitter.

In summary, the application of the arithmetic mean of the fitness at the moment of reproduction is a suitable descriptor of the short-term success of a population. But as I discussed above, in changing environments the arithmetic mean is not necessarily a representative long-term fitness descriptor. Two populations can have the same arithmetic mean fitness and very different variability of their fitness in a changing environment. This discrepancy can be addressed with the geometric mean fitness as described.

3.3.3 Discussion.

In this model I look at the fitness distribution of individuals using a reinforcement learning strategy, i.e. Q-learning, in connection with different aspects of a changing environment: regularity, frequency, and size of change. My model compares the fitness distributions of individuals using the learning strategy with the performance of a population under simplistic mutation of fixed phenotypes in order to gain insights into the benefits of learning. The main findings are:

- a random mutation process is non-adaptive and consequently the fitness distribution of a population under a simplistic mutation process is inde-

pendent of most aspects of a changing environment in my model (I note that this is because of the situation selected, where there are two options of identical long term value),

- the fitness distribution of the mutating population is symmetrical around fitness neutrality in the environment of my model where individuals are generally well adapted,
- learning requires environmental change and longer generation times to be beneficial,
- learning is optimal for specific combinations of regularity and size of environmental change,
- the fitness distribution of learning individuals in changing environments is independent of the learning rate α and the discount factor γ, and
- regularity is the only environmental factor which impacts whether learning is generally advantageous.

The motivation of this model lies in questions around the benefits of learning when individuals are generally well adapted to their environment through evolution. In particular, I am interested in conditions enabling the initial evolution of learning. I apply Q-learning as an implementation of reinforcement learning for its simplicity and its increasing support from both behavioural and neural data. I assumed that the environment only allows inference from trial-and-error and learning cannot draw from additional sources of information such as the correlation between environmental states in the initial evolution of learning. As a result I do not model the environment with regard to absolute states but in relation to a relative fitness difference between two options. With reference to animals being generally well-adapted to their environment these options are only relevant in the light of evolutionary selection if they have the same long term value (Nowak, 2006). The difference between the two options in my model is the changing fitness payoff of the uncertain option. This condition does not apply strictly to the learning strategy. For learning one of the options could be a little worse than the other on average as long as there is enough variation allowing for sufficient exploitation. Additionally, the mutating population is affected by long term fitness effects. In changing environments the arithmetic mean is not a representative descriptor for fitness in a population dynamical context. Two populations can have the same arithmetic mean fitness and very different variability of their fitness in a changing environment. The variability of the fitness can have a dominating effect on the long-term fitness in a population dynamical

context, especially when generation times are short. Correspondingly, the population displacement occurs geometrically and not linearly being the continuous product of the fitnesses of the parental and offspring populations. It can be shown that if the arithmetic mean of two populations are the same, the population which has a more variable fitness has lower overall fitness. Section 3.3.2 on long-term fitness effects provides a more detailed investigation of the fitness in this context.

I do not present an evolutionary theory of learning in itself. But I show that a simple reinforcement strategy which is increasingly backed by experimental studies of neural correlates is beneficial for a vast range of environmental parameters. In particular, the fact that the success of the learning strategy is independent of technical parameters of learning, i.e. the learning rate α and the discount factor γ, is a new reassuring insight. These are technical parameters which allow the tuning of over-fitting and the extent of exploration for a specific learning task and have great importance in the field of computing. But in a biological context of changing environments these technical learning parameters become negligible. This significantly reduces the complexity of the initial evolution of reinforcement learning. For a coherent evolutionary theory I would also have to consider the effects the environment has on the development and the phenotype of individuals, phenotypic plasticity, in addition to the environments role in the fitness function. Phenotypic plasticity is a genetically defined process and refers to all environmentally induced changes which derive from a change in gene expression. Such changes may or may not be permanent and include behavioural changes. The performance of a plastic behavioural strategy in my model would depend greatly on the cost of plasticity but might be better than the fixed phenotype of individuals in the mutating population (DeWitt et al., 1998). There is a wide range of literature looking into plasticity as a form of adaptation to changing environments which addresses very similar questions, e.g. the evolution of plasticity as an adaptation to changing environments and their benefits (Pigliucci, 2001; Via et al., 1995). Nevertheless, plasticity is distinct from learning, which relies on cognitive processes.

It has been widely acknowledged that the benefit of learning is the ability to adapt to a changing environment faster than the time scale on which evolution operates (Johnston, 1982; Ackley and Littman, 1991). This is certainly an important population dynamical argument. But the focus on population dynamics of previous studies emphasises the individual performance on a specific subset of tasks and therefore the importance of technical parameters. Additionally, it raises further questions, e.g. depletion of the surrounding resources and the mortality and interactions of individuals.

My general model of learning using reinforcement learning does not include interactions between individuals and is not a population dynamical model. Nevertheless, my results reproduce many of the widely accepted theories of learning in the context of change and regularity (Stephens, 1991). In my model the benefits of learning originate in the ability of exploitation rather than the speed of adaptation itself.

Considering the effects of selection the mutating population in my model has a constant relative arithmetic mean fitness of zero. The environmental changes only affect the fitness variability of the mutating population in a symmetric fashion. As the arithmetic payoff of both options in my model are equal, selection would increase long-term fitness of the mutating population by discarding the uncertain option from the action space of the mutation process in order to reduce fitness variability. This provides an alternative interpretation of why learning is a distinct adaptation to changing environments alongside the cost argument: a simplistic mutation process cannot exploit environmental change without the introduction of increased fitness variability at the same time.

Taking the consequences of selection into account, my results show that reinforcement learning is a promising starting point for the initial evolution of learning. The only environmental factor which impacts the general success of learning is regularity. If selection cannot discard the uncertain option from the action space of the mutational process, learning is always beneficial as it has lower fitness variability even in extremely irregular environments when compared to the mutating population. If selection can in fact discard the uncertain option in the case of the mutating population then learning becomes disadvantageous in irregular environments.

Chapter 4

A predator lifetime model.

The previous chapter introduced reinforcement learning to models of predator-prey interactions. I used the Q-learning algorithm to address the question of how predators generalise information from their encounters with potentially aposematic prey into foraging behaviour. In the previous chapter, Q-learning showed to be an elegant solution to optimal behaviour in changing and uncertain environments and its analytical tractability allowed an application in mathematical models of aposematism and predator-prey interactions. Q-learning is also shown to be an advantageous adaptation to changing environments within a biological context in general regardless of technical parameters. Following the promising results of the previous models I will develop a more complete lifetime model of predators utilising reinforcement learning in this chapter.

4.1 The motivation to learn.

The previous chapter introduced conditioning as a way for animals to predict and respond to events in their environment. In particular, in operant conditioning, the type of feedback an animal receives depends on the actions it performs. Operant conditioning is thus closely related to the optimal-control problem and reinforcement learning theory in computer science which was the motivation to apply Q-learning to models of predator-prey interactions.

However, an important question remains: what are the biologically and psychologically relevant components of a reward? Biological models are generally concerned with *(Darwinian) fitness* as the core of evolutionary theory. On the individual level, fitness describes the ability of an animal to survive and reproduce within its environment. However, the fitness definition also applies transparently to the genetic level through the individual's contribution to the gene pool. The fitness of an individual manifests itself through its phenotype

which connects both levels as the subject of natural selection (Fisher, 1930; Huxley, 1942). From this evolutionary perspective the previous application of reinforcement learning (Section 3.3) assumes a monotonically increasing functional relation between rewards and fitness. In such a scenario optimisation of rewards seems like a straightforward choice. However, little attention has been paid to this assumption in the previous chapter. The success of reward mediated learning as a widely observable adaptation to the environment can easily deceive the observer into believing that this assumption is generally true. Nevertheless, this chapter is going to open Pandora's box in an attempt to quantify the link between reward driven behaviour and fitness driven selection. This is by far the most ambitious and controversial part of this thesis and this attempt to unify reward and fitness is bound to fail. Very little is still known about the relation of behavioural and genetic traits. In particular, the evolutionary dynamics of phenotypic variance in animal behaviour are poorly understood. To fully understand the evolution of animal behaviour it requires both mechanistic and functional approaches. The mechanistic approach tries to quantify the influence of genetic and environmental factors on the phenotype whereas the functional approach tries to describe how the interaction of phenotypes and their environment affects fitness. However, functional approaches towards understanding behaviour have received very little attention as summarised in the review by Dingemanse and Réale (2005). The following model will explore the functional approach and builds on computational theories of reinforcement learning which do not implement any psychological elements of rewards themselves. I will have to revert to a creative definition of the environment to simulate the effects which psychological elements might have on a predator's foraging behaviour in Section 4.3. Even though the model will not be able to provide a complete description of the functional relations of rewards and fitness it will, firstly, point out the importance of the functional component in understanding animal behaviour and, secondly, it will show how RL provides an interesting methodology to do so.

What is the discrepancy between maximising biological fitness through natural selection and maximising rewards through reinforcement learning? In reinforcement learning the behaviour of individuals is modulated by its consequences: a desirable outcome, positive reinforcement, increases the probability of the behaviour and an undesirable outcome, positive punishment, decreases the probability of behaviour. Many of the behavioural studies have shown that energy content of food is an example of a strong positive reinforcer which supports the idea of flavour-calorie learning in the context of foraging behaviour. In such studies rats were given the choice between two differently flavoured non-nutritious solutions. If the rat consumes the positively reinforced solution

it leads to the intra-gastric administration of an energy-rich agent versus an infusion with water in the case of the consumption of the neutral solution. The outcome of such discriminative learning studies was a clear preference for the conditioned excitatory stimulus for a large number of energy-rich reinforcers such as sucrose, glucose, starch, and fats (Sclafani, 1990; 2004).

The question of whether behaviour is always optimal can be addressed from two perspectives: (i) does behaviour maximise positive reinforcement and (ii) is behaviour optimal in relation to maximising fitness in a biological context. In particular, the relation between maximising rewards in flavour-calorie learning and increasing fitness is not unconditionally positive. Even though obesity is not a common phenomenon of wild animals, laboratory studies have shown that also animals are generally prone to health and fitness costs in scenarios of unrestrained reward maximisation in flavour-calorie learning. Additionally, realised behaviour can be very subjective and variable with many well documented examples of apparently non-optimal behaviour where reinforcement learning seems to fail. See Breland and Breland (1961) for a wide range of accounts on conditioning animals for shows and TV and how the notion of animal instinct sets boundaries to conditioning by reinforcement. One commonly observed divergence from optimal behaviour in operant conditioning experiments is *risk aversion*: given a certain and an uncertain option most subjects tend to show a preference for the certain option even if the uncertain option has a higher expected payoff. But even after the removal of the uncertainty suboptimal behaviour can be observed in so called *self-control* experiments: subjects have to resist a mediocre payoff in order to get a greater payoff. Even though it would be optimal to resist the mediocre payoff in order to get the better payoff most subjects choose the mediocre payoff if the greater payoff involves some time delay (Staddon and Cerutti, 2003).

Of course all such behaviour is only apparently suboptimal under an isolated reward maximisation point of view which does not take the biological context of such evolved behaviour into account. It is therefore crucial to understand animal behaviour in its evolutionary context using mechanistic and functional approaches.

It is understood that rewards can be divided into three specific psychological components: (i) learning (e.g. knowledge produced by associative conditioning), (ii) affect (so called *liking*, hedonic impact), and (iii) motivation (so called *wanting*, incentive salience) (Berridge et al., 2009). All three components can cause the reward fitness relation to be non-monotonic and aspects of affect and motivation are commonly ignored within the computational theories of reinforecement learning. As discussed earlier, studies have shown that the associative learning component of rewards is closely related to reward prediction

and dopaminergic neurons in the nucleus accumbens (Schultz, 2002). However, the nature of subjective and objective affective reactions (liking) involves opioid neurotransmitters and GABAergic neurons in the nucleus accumbens (Berridge, 2003). Most rewards which are liked are also wanted but the processes of subjective desire or objective motivation (wanting) are distinct from the processes of liking with their own neural substrates, mesolimbic dopamine amongst others (Dayan and Balleine, 2002). Within the brain all three components of rewards are interacting with each other and I hypothesise that the processes of liking and wanting are the main contributions to a potential non-monotony of the reward and fitness relation (Robinson and Berridge, 2003). In this regard, the endogenous opiate system in particular seems to play a crucial role in defining the incentive value of foods (Berridge, 1996).

It is evident that the rewards within the computational theories of reinforcement learning are a great simplification of the true psychological and neuronal nature of rewards within the brain. The close relationship of RL with optimal control problems makes it applicable to biological models of fitness but seems to be an inappropriate choice for models of individual behaviour. There are various implementations of RL which address the true nature of rewards such as models of planning and motivational states in actor-critic models (Dayan and Balleine, 2002; Niv et al., 2006). Summarising, the application of RL to biological systems and the consequential link between rewards and fitness seems unproblematic itself. The discrepancy seems to lie in the importance of 'wanting' and 'liking' in the realisation of individual behaviour which does not map monotonically to fitness. The next section tries to define a lifetime model for predators which can be interpreted on both levels: (i) the level of realised behaviour of a single individual driven by rewards including components of affect and motivation and (ii) the level of behavioural repertoires of a population driven by fitness. The discussion will compare the results and insights from both levels to find similarities and differences between reward motivated objectives of individual behaviour and the evolution of behavioural repertoires driven by fitness.

4.2 A predator lifetime model.

This section introduces the lifetime model of an individual predator and the definition of the individual's payoff based on its environment and additional aspects of its behaviour, metabolism, and lifetime traits which have been abstracted away in the previous chapter. In this model an individual predator is

Symbol	Definition
s_k	The state vector describing the environment at iteration k.
u_k	The action vector at iteration k.
T	The continuous time variable.
$\dot{T}$	The transition function of time T.
A	The age of a predator.
$\dot{A}$	The transition function of age A.
V	The total payoff of predator at the end of its lifetime.
$\dot{V}$	The state and action dependent payoff.
X, Y	The spatial location of the predator.
e_x, e_y	The investment of the predator into locomotion.
$g_i(X, Y)$	The dispersion of prey population i within the environment.
p_i	The density of prey population i.
$R(s)$	The state dependent reward term.
$d(t)$	The probability of ingesting a prey individual of toxicity t after taste-sampling.
$\lambda(A)$	The age-agility of predators of age A.
T_l	The length of an episode of foraging given by the cut-off time T_l.

Table 4.1: Parameters and their definition.

characterised by its state vector s_k at iteration k. The state vector is given by

$$s_k = \{T, A, X, Y\}, \tag{4.1}$$

with T being the time of an iteration k within an episode, A being the age of the predator, and X, Y being the spatial location of the predator within its environment at iteration k. An episode in this model corresponds to a day of foraging with the length of an episode given by the cut-off time T_l and an episode being defined as $T < T_l$.

The predator finds itself in an environment defined by the availability of different food sources. The dispersion of each prey population i within the environment is described by a Gaussian function

$$g_i(X, Y) = p_i \exp\left(-\left(\frac{(X - x_{i,0})^2}{2\sigma_{i,x}^2} + \frac{(Y - y_{i,0})^2}{2\sigma_{i,y}^2}\right)\right), \tag{4.2}$$

with $(x_{i,0}, y_{i,0})$ being the centre of the prey population with density p_i and $(\sigma_{i,x}, \sigma_{i,y})$ being the spread of the prey.

The model assumes that the prey is aposematic with potential mimics being present. (The model can also be used to include non-aposematic prey. The limitation to aposematic prey is solely to simplify the forthcoming analysis of the model.) The predator feeds on prey it encounters as it cannot distinguish

between models and mimics based on their appearance. However, the predator has the option to move around freely in its environment to avoid encounters with possibly aversive prey based on its experience. The predator's locomotion is defined by its action vector u_k which is given by

$$u_k = \{e_x, e_y\}\,, \tag{4.3}$$

with e_x, e_y being the energy invested into locomotion at iteration k.

The value function V describes the total payoff of a predator at the end of an episode (a day of foraging in this model) and is the result of a predator's environment and its actions. Thereby, the predator's actions have subsequent effects on the composition of the prey population of its surroundings through locomotion and the predator's spatial location within the environment according to the reinforcement learning model (Figure 3.2). In this section the predator's value function V is defined by the sum of its payoffs and is not directly equivalent to Darwinian fitness following the previous discussion. The following Section 4.4 will discuss necessary modifications of the model to interpret the total payoff V as Darwinian fitness. The subsequently received payoff for the predator being in a specific state s_k and taking action u_k at iteration k is given by the payoff function as follows

$$r_{k+1} = \dot{V} = \underbrace{\lambda(A_k)R(s_k) - t_0\dot{T}}_{\text{state dependent}} \overbrace{-|E(u_k)|}^{\text{action dependent}}\,, \tag{4.4}$$

with $t_0\dot{T}$ being the metabolic cost of the predator, $-|E(u_k)|$ being the absolute energy expenditure of a predator's actions, and $R(s)$ being the state specific payoff given as follows

$$R(s_k) = \sum_i g_i(s_k)d(t_i)(r - t_i^2), \tag{4.5}$$

with r being the baseline reward of a prey item and

$$d(t) = \frac{1}{1 + d_0 t} \tag{4.6}$$

being the probability of ingesting a prey individual of toxicity t after taste-sampling. The model has the option to include age related effects such as an age dependent agility of the predator given as follows

$$\lambda(A) = \frac{1}{1 + A}\,. \tag{4.7}$$

The environment of this model is Markovian defined by a state transition function which is given as follows:

$$f(s,u)_{k\to k+1} = \begin{pmatrix} \dot{T} = 1 + \sum_i g_i(s_k)\left(d(t_i)\left(t_h + t_t t_i^2\right) + t_s\right) \\ \dot{A} = (1/\lambda_0)\dot{T} \\ \dot{X} = \tanh(c_0 e_x) \\ \dot{Y} = \tanh(c_0 e_y) \end{pmatrix}. \tag{4.8}$$

I use the dot notation to describe the functional change between iterations ($k \to k+1$) as a shorthand for the derivative $\dot{f} = df/dk$. The transition of time ($\dot{T} = dT/dk$) between iterations ($k \to k+1$) occurs in unit time steps reflecting a basal metabolic expenditure and the additional costs of foraging such as the sampling of prey items t_s, the handling of prey t_h, and the recovery from ingested toxins $t_t t^2$. The predator ages ($\dot{A}$) linearly with time. The predator's locomotion results in a change of its spatial location ($\dot{X}, \dot{Y}$) depending on the predator's energy investment e_x, e_y with the maximal spatial displacement per iteration being an unit step of one. The functions of the model follow the same motivations as in the previous chapter and are governed by single parameters which allow the trade off between the different aspects of the predator's behaviour, lifetime traits, and environment ($x_0, y_0, t_0, d_0, \lambda_0, c_0$).

In the next Section 4.3, I introduce a foraging simulator for individual predators based on the lifetime model presented here. The aim of the individual based simulation is to gain a better understanding of the psychological components of rewards which characterise the individual's realised behaviour such as affect and motivation. In Section 4.4, the model is modified to reflect the Darwinian fitness component of the rewards based on behavioural repertoires and the assumption of a co-evolution of predator and prey under stabilising selection. Finally, the discussion in Section 4.5 will present and analyse the differences in the results.

4.3 A TD learning based foraging simulator.

This section combines the previously introduced lifetime model with TD learning into a foraging simulator. It is possible to observe realised behaviour of predators interacting with their prey within their natural environment which they are adapted to. With regard to the lifetime model there are different factors which can be measured or quantified such as the metabolic rate of a predator, the handling time, sampling time and toxin recovery time after ingesting aposematic prey. Additionally, the prey can be quantified by the toxicity of prey individuals and the density of a prey population. Also, it should be possible to quantify the average energy expenditure of a predator into different behaviours

but especially into locomotion based on the size of a predator's territory and its average travelling distance.

These parameters define my lifetime model. The only term missing is the subjective payoff r in Equation (4.5). The assumption and motivation for the simulator is that this subjective payoff can be reverse engineered from the observed foraging behaviour of the predator using a reinforcement learning algorithm. The aim is to find the subjective value of the payoff r for prey type i in order to reproduce the observed foraging behaviour of the predator.

Additionally to the lifetime model, the simulation defines a final instantaneous cost Ψ of the terminal state s_l with l being the final iteration of an episode based on the spatial distance of the predator from its den at $(X = 0, Y = 0)$:

$$\Psi(\vec{s})_l = \begin{cases} -r_l\sqrt{X^2 + Y^2} & \text{if } \sqrt{X^2 + Y^2} > \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{4.9}$$

with $-r_l$ being a punishment for not returning to the den at the end of an episode. The lifetime model defines the predator's payoff in terms of subjective rewards. However, the final instantaneous cost Ψ adds an additional Darwinian fitness component to the model. This will allow to simulate the discrepancy of maximising rewards and maximising fitness within the computational RL algorithms. Within a biological context I suspect such a final cost Ψ to be step-like around the predator's den. If a predator has to feed offspring staying behind in the den the cost of almost returning will not decrease smoothly within the proximity of the den. (There are smooth penalty functions which are also biologically meaningful, e.g. defining an increasing penalty for returning late to the den instead of a precise cut-off at the end of an episode. However, such a penalty function is difficult to implement with the episodal RL algorithms used in this chapter. Additionally, this functional shape has been chosen to simulate psychological effects of rewards as I will explain in detail in Section 4.3.5.)

The simulator (Figure 4.1) has been written in `C++` utilising two implementations of reinforcement learning: 1) back-propagation through time (BPTT), a hill climbing method on the value function, and 2) value gradient learning (VGL), a hill climbing method on the target gradients themselves. Additionally, the simulator makes use of artificial neural networks as universal function approximators in order to implement the different elements of the learning problem such as the behavioural policy. I refer the reader to the Section 4.3.1 which follows for a detailed introduction of artificial neural networks (ANN) as universal function approximators. The learning algorithms will build on the existing concepts of efficiently calculating the derivatives of the network function from ANNs by extending it to the optimisation problem of episodal tasks as in this

predator lifetime model.

4.3.1 Artificial neural networks.

An artificial neural network (ANN) is a computational model inspired by the information processing functionality of the brain. But how does the brain compute? Generally, the central elements of computation are processing, transmission, and storage. Within the brain the neuron is the central computing element. Neurons receive signals and produce responses. The transmission of information at the neural level involves electrical signals – so called action potentials – based broadly on ions and semipermeable membranes, and chemical signals at the synapses. In the brain the storage of information corresponds to learning which occurs at the synapses. These synapses are at the interface between neurons and regulate the transmission of information from neuron to neuron.

An ANN widely corresponds to the processing paradigm of neural networks with the nodes of the ANN being the central computing element similar to the neuron. In fact, ANNs are nothing but networks of primitive functions where the chain of function compositions transforms an input to an output. The composition of the computational model is contained implicitly in the interconnections of the nodes and is referred to as the *network function*.

Each node comprises a primitive function transforming its input into an output (Figure 4.2). Typically, the inputs of a node have an associated weight w_i by which the input x_i is multiplied. The node integrates all its inputs – usually by adding the different inputs – followed by the evaluation of its primitive function f. The primitive function f computed in the node can be any function but common choices are differentiable functions such as the sigmoid function. Models of ANNs mainly differ in their choice of the primitive function, the topology of the network, and rarely in the timing of the evaluation of the primitive function.

In *feed-forward* ANNs the network is composed of distinctive layers where each neuron only receives input from neurons of the previous layer. Accordingly, a feed-forward network has a distinct input and output layer with the intermediate layers being referred to as hidden layers (Figure 4.3). The second class of ANNs are *recurrent* networks where connections between nodes form directed cycles.

The network function of an ANN can be understood as a universal function approximation. However, the difference between ANNs and a Taylor or Fourier series is that the function to be approximated is given not explicitly but implicitly, through a representative set of input-output examples. It will be the task of the learning algorithm to adjust the parameters of the ANN to reflect the

```
#include "config.hpp"
#include "predator.hpp"
#include "environment.hpp"

int main(int argc, char* argv[])
{
  Config * CONFIG = new Config();
  CONFIG->use_defaults();

  // Adding a food source to the environment
  Array<double,1> mu(2);
  mu = 5;
  Array<double,1> sig(2);
  sig = 5;
  Gaussian gaus(mu, sig);
  ptr_food_source food_source(new FoodSource<Gaussian>
        (CONFIG, gaus,
         2.0, // toxicity
         2.0, // payoff
         0.1, // handling_time
         0.5)); // density
  ptr_environment environment(new Environment(CONFIG));
  environment->add_food_source(food_source);

  // creating a predator(CONFIG, Environment, use_rprop,
  //                     learning method)
  Predator predator(CONFIG, environment, false, 1);
  std::vector<int> layout_actor;
  layout_actor = {4,10,2};
  predator.set_Actor(layout_actor);
  std::vector<int> layout_critic;
  layout_critic = {4,10,4};
  predator.set_Critic(layout_critic);

  while (!stopFlag) {
      predator->run_episode();
  }
  delete CONFIG;
  delete predator;
  return 1;
}
```

Figure 4.1: Code-fragment defining the elements of the predator lifetime simulator. The complete source code of the simulator can be found on GitHub (Teichmann, 2014).

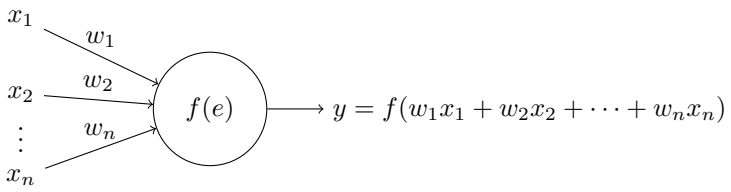


Figure 4.2: An abstract neuron representing a node in an artificial neural network. The neuron is evaluating its primitive function $f(e)$ whereas the neurons excitement e is given by the weighted w inputs x.

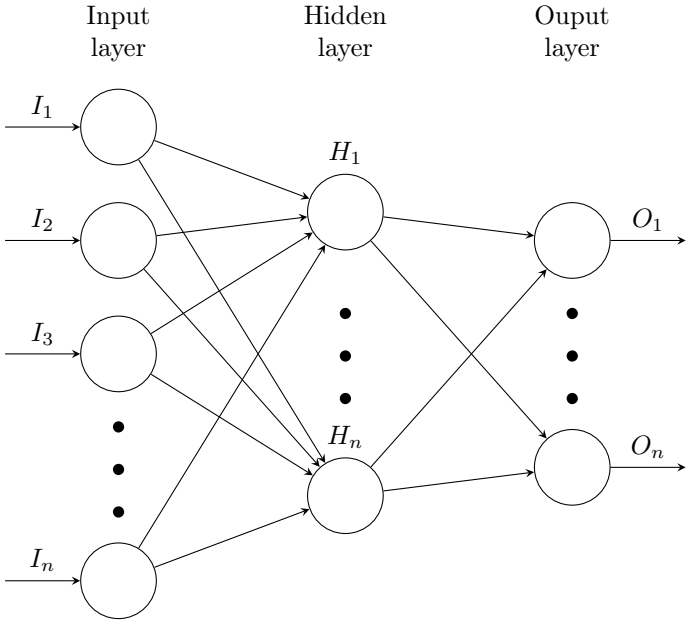


Figure 4.3: A feed-forward artificial neural network characterized by its distinctive layers.

input-output examples and to extrapolate to new input patterns in an optimal manner. The learning algorithm is an adaptive method by which the network self-organises to reflect the function to be approximated. The computational effort directly relates to the number of parameters and therefore to the topology of the network and increases substantially for more complicated ANNs. It was not until the proposal of *back-propagation* as a learning algorithm (Werbos, 1974) that the application of ANNs gained momentum and it has been the most widely used algorithm for neural network learning ever since.

The back-propagation algorithm uses gradient descent on the error function of an ANN in weight space. Thus, the weights of an ANN which minimise its error function are considered to be the solution of the learning problem. As a precondition for gradient descent the error function of an ANN needs to

be continuous and differentiable. Since the ANN is simply the composition of its primitive functions the error function becomes differentiable if the networks primitive functions are differentiable themselves.

In the back-propagation algorithm an ANN is initialised randomly with weights. Next, the gradient of the error function is computed recursively and the weights of the ANN are adjusted accordingly using gradient descent. Because an ANN is a complex chain of a sequential function composition the chain rule plays a most important role in calculating the gradient of the network function's error. The back-propagation algorithm implements the chain rule for the recursive calculation of the gradient of the error function in weight space in a very efficient manner.

Learning in an ANN with back-propagation consists of two stages: in the first stage, the feed-forward step, the information progresses form the input layer throughout the network towards the output layer. Each node of the network evaluates its primitive function $f_j(e)$ and emits the result y_j to the connected nodes in the subsequent layer. Additionally, each node calculates and stores the derivative of its primitive function $df_j(e)/de$. The second stage, the back-propagation step, consists in reversing the flow of information throughout the network whereby a unit input propagates from the output layer towards the input layer with the activation of each neuron now being the back-propagation term δ_j. At each node the back-propagation term δ_j is multiplied by the stored derivative of the node's primitive function from the previous feed-forward step which gives the gradient in weight space $(df_j(e)/de)\delta_j$. Finally, the weights are updated using gradient descent as given by

$$w'_{i,j} = w_{i,j} + \alpha y_i \frac{df_j(e)}{de}\delta_j, \tag{4.10}$$

with α being the learning rate and $w_{i,j}$ being the weight of the feed-forward connection from neuron i in the previous layer to neuron j in the subsequent layer.

In the case of batch or off-line learning the weight changes are aggregated over the complete set of input/ output examples and the ANN is updated after all examples have been presented to the ANN. In so called on-line learning the weights are updated sequentially after the presentation of each input/ output example.

In the lifetime model presented here, the ANN is used in an episodal learning problem where the ANN is applied repeatedly to generate a trajectory. The weight update of the ANN occurs at the end of a completed trajectory which will now not only depend on the current input/ output example but also on the prevailing inputs and outputs of the ANN which subsequently determined

the current example. Therefore, the reinforcement learning algorithm has to address the weight update of an ANN in episodal tasks.

4.3.2 Back-propagation through time.

The previous chapter introduced Q-learning as an algorithm to solve the reinforcement learning problem. In Q-learning the individual receives a subsequent reward which it uses to form a target of the value of the occurred state transition. Thereby, the individual bases its target formation on the prevailing estimates of discounted future rewards, also known as bootstrapping. I used Gibb's soft-max policy as the behaviour policy to address the exploration-exploitation trade-off under the precondition of continuous learning and exploration. In the lifetime model presented here the individual takes episodal decisions and is trying to optimise the value of an episode of foraging under the constraints of locomotion, metabolism, the composition of the prey population, and returning to the den at the end of every episode. Many aspects of the reinforcement learning problem in the lifetime model are equivalent to elements and concepts of the Q-learning algorithm of the previous chapter.

However, in this model the behaviour policy has been implemented with an artificial neural network which is also called the *actor*. Within the episodal task the individual repeatedly applies the actor to generate a trajectory of actions and consequent state transitions and rewards. The usage of the ANN as universal function approximator for the behaviour policy and the episodal nature of the reinforcement learning problem in this model requires a different type of RL algorithm.

As discussed previously, the control problem in reinforcement learning is about finding an optimal behaviour policy. Reformulating the RL problem using an actor, the aim becomes to find the parametrisation $\vec{z}$ of the actor $\pi(\vec{s}, \vec{z})$ which maximises the total value or minimises the overall temporal difference error for a complete trajectory based on $V(\vec{s}, \pi(\vec{s}, \vec{z}))$. This can be achieved using hill climbing on the total value of a complete trajectory itself with respect to $\vec{z}$, i.e. $\Delta\vec{z} = \alpha(\partial V/\partial\vec{z})$, which is also called a policy gradient with *back-propagation through time* (BPTT) being an efficient implementation of the optimisation problem for episodal tasks as in this model here (Werbos, 1990; Fairbank, 2013). BPTT uses the actor $\pi(\vec{s}, \vec{z})$ which has been implemented as an artificial neural network with weights $\vec{z}$ as a universal function approximator and is equivalent to the behaviour policy of previous models. As such, BPTT is an off-line learning algorithm which issues a weight update $\Delta\vec{z}$ at the end of an episode. Differently to the previous chapter, the delayed effects of actions in the RL problem means that the outcome of a trajectory is not only dependent on its

initial conditions but also on all the actions of an individual and the subsequent state transitions. Therefore, an episodal learning algorithm has to consider the entirety of actions and state transitions of a trajectory within its updates. Thereby, the trajectory of a complete episode is unrolled backwards using the Markovian properties of the environment with the component $(\partial V/\partial \vec{z})_k$ being computed from the prevailing quantity $(\partial V/\partial \vec{z})_{k+1}$, i.e. the policy gradient of the value function is computed backwards in time starting at the end of an episode (eq: (4.12)). This property gives the methodology its name. As introduced earlier, back-propagation is an efficient way of calculating the derivative of the network function in artificial neural networks. Back-propagation through time is the extension of that methodology to efficiently calculate the derivative of the network function in episodal tasks where the neural network has been applied multiple times to create a trajectory of states and payoffs – similarly to recurrent neural network problems – including the previously introduced concepts of discounting and bootstrapping. Hence, the derivative of the overall network function is the sum of the discounted incremental gradients at each iteration of the trajectory with their calculation expanding as follows: at the beginning of the BPTT algorithm the partial gradients of the value function are initialised: $(\partial V/\partial \vec{z})_l \leftarrow \vec{0}$ and $(\partial V/\partial \vec{s})_l \leftarrow (\partial \Psi/\partial \vec{s})_l$ with Ψ (eq (4.9)) being the final instantaneous cost of the terminal state s_l and l being the final iteration in an episode of finite length.

The following RL algorithms are presented using the *trajectory-shorthand notation* as introduced by Fairbank (2013). The subscript k refers to a specific iteration of an episode with the corresponding states s_k and actions u_k, thus $(r)_k := r_{k+1}(\vec{s}_k, \vec{u}_k)$ and $(\partial V/\partial \vec{z})_k$ is $\partial V(\vec{s}_k, \vec{z})/\partial \vec{z}$ evaluated at $(\vec{s}_k, \vec{z})$.

Following the initialisation, the algorithm processes the trajectory of an episode backwards in time starting from the second last iteration to the first iteration in the episode. At each step the algorithm adds the partial policy gradient of the current iteration to the overall policy gradient of the value function for an episode $\partial V/\partial \vec{z}$ as follows:

$$\left(\frac{\partial V}{\partial \vec{z}}\right)_k \leftarrow \left(\frac{\partial V}{\partial \vec{z}}\right)_{k+1} + \gamma^k \underbrace{\left(\frac{\partial \pi(\vec{s}, \vec{z})}{\partial \vec{z}}\right)_k}_{\text{actor}} \underbrace{\left(\left(\frac{\partial r}{\partial \vec{u}}\right)_k + \gamma \left(\frac{\partial f}{\partial \vec{u}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}\right)}_{\text{behavioural target gradient}}, \quad (4.11)$$

which gives the iterative calculation of the gradient on the total value of any given trajectory in the parameter space of the actor. The single contributions to the overall gradient are discounted by a factor γ. The iterative contributions to

the gradient in the parameter space of the actor are the product of the partial derivative of the actor and the behavioural target gradient. The target gradient is given by the effects of a predator's behaviour on the short term reward payoff $\partial r/\partial \vec{u}$ and on the long term value due to changes in the environmental state caused by the predator's behaviour $\gamma(\partial f/\partial \vec{u})_k(\partial V/\partial \vec{s})_{k+1}$. The state dependent value contribution $\partial V/\partial \vec{s}$ derives from the Markovian properties of the environment as follows:

$$\left(\frac{\partial V}{\partial \vec{s}}\right)_k = \underbrace{\left(\frac{\partial r}{\partial \vec{s}}\right)_k + \gamma \left(\frac{\partial f}{\partial \vec{s}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}}_{\text{environmental target gradient}} + \underbrace{\left(\frac{\partial \pi(s,\vec{z})}{\partial \vec{s}}\right)_k}_{\text{actor}} \underbrace{\left(\left(\frac{\partial r}{\partial \vec{u}}\right)_k + \gamma \left(\frac{\partial f}{\partial \vec{u}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}\right)}_{\text{behavioural target gradient}}. \quad (4.12)$$

Because of the recurrent nature of the episodal learning task where the predator's behaviour has consequent effects on the environmental state, the state dependent value contribution itself is the sum of the state dependent environmental target gradient and the behavioural target gradient. Thus, the BPTT algorithm is bootstrapping the value function just as the Q-learning algorithm does.

The final weight update gives the implementation of hill climbing on the value function V with respect to the policy gradient of $\pi(\vec{s},\vec{z})$ as follows:

$$\vec{z} \leftarrow \vec{z} + \alpha \frac{\partial V}{\partial \vec{z}}, \quad (4.13)$$

with α being the learning rate.

Summarising, the BPTT algorithm can be understood as propagating the policy gradient of the value function with respect to a future state $(\partial V/\partial \vec{s})_{k+1}$ backwards in time through the actor, the state transition function, and the payoff function to obtain the policy gradient of the value function $(\partial V/\partial \vec{s})_k$ of the previous state. As BPTT utilises the Markovian properties of the environment using the state transition function for the propagation of the state dependent gradient backwards through time it is a model-based methodology. BPTT as a simple hill-climbing algorithm on the value function has robust convergence proofs (Fairbank, 2013). Additionally, its implementation using an actor and the application to episodal tasks is relatively simple. However, the BPTT algorithm has some shortcomings: as every gradient based optimisation it is sensitive towards local optima. Especially in combination with the exploration-exploitation dilemma the BPTT algorithm might converge to a suboptimal trajectory.

4.3.3 Value gradient learning.

In the previous section I introduced the BPTT algorithm to find the parametrisation of an actor which maximises the value of a complete trajectory in episodal tasks. As such BPTT is a policy gradient method with respect to $\vec{z}$, the parametrisation of the actor $\pi(\vec{s}, \vec{z})$.

An alternative would be the usage of a gradient with respect to the state space $\vec{s}$. Such a method is called *value-gradient learning* or VGL for short (Fairbank, 2013). The aim of VGL is to learn the value gradient

$$G(\vec{s}, \vec{w}) = \frac{\partial V(\vec{s}, \vec{w})}{\partial \vec{s}}, \tag{4.14}$$

where the $\vec{w}$ is the parametrisation of the value gradient function which is also called the *critic* and which, similarly to the actor, has been implemented with an artificial neural network as universal function approximator. Hence, VGL requires two ANNs one implementing the behaviour policy, called the actor, and one implementing the value-gradient, called the critic. The reason for choosing an ANN to implement the critic lies in the fact that the gradient of the critic defined by Equation (4.14) $\partial G/\partial \vec{w}$ would require second-order back-propagation $\partial G/\partial \vec{w} = \partial^2 V/(\partial \vec{w} \partial \vec{s})$ and using an ANN to approximate $G(\vec{s}, \vec{w})$ requires only first-order back-propagation.

The aim of VGL is to learn the value gradient G over the state space $\mathbb{S}$ equivalent to $\partial V/\partial \vec{s}$; which is the same as learning $V(\vec{s}, \vec{w})$ with the addition of a constant.

Similarly to BPTT, the VGL algorithm unrolls the trajectory of a complete episode backwards in time to calculate the overall value gradient as the sum of discounted incremental gradients at each iteration of the trajectory. The VGL algorithm unfolds as follows: the VGL algorithm is initialised at the terminal iteration of a trajectory with $G'_l \leftarrow (\partial \Psi/\partial \vec{s})_l$, $\Delta \vec{z} \leftarrow \vec{0}$, and $\Delta \vec{w} \leftarrow (\partial G/\partial \vec{w})_l (G'_l - G_l)$ with $\Psi(\vec{s}_l)$ (eq. (4.9)) being the final instantaneous cost of the terminal state $\vec{s}_l$ and l being the final iteration in an episode.

Following the initialisation, the VGL algorithm unrolls the trajectory of an episode similarly to the BPTT algorithm backwards in time beginning from the second last iteration to the first iteration in the episode. At each step the algorithm aggregates the discounted incremental value gradient of each iteration to the overall experience based value gradient of the complete episode $\partial V(\vec{s}, \vec{w})/\partial \vec{s}$ as follows:

$$G'_k \leftarrow \underbrace{\left(\frac{\partial r}{\partial \vec{s}}\right)_k + \gamma \left(\frac{\partial f}{\partial \vec{s}}\right)_k \vec{p}}_{\text{environmental target gradient}} + \underbrace{\left(\frac{\partial \pi(\vec{s},\vec{z})}{\partial \vec{s}}\right)_k}_{\text{actor}} \underbrace{\left(\left(\frac{\partial r}{\partial \vec{u}}\right)_k + \gamma \left(\frac{\partial f}{\partial \vec{u}}\right)_k \vec{p}\right)}_{\text{behavioural target gradient}}, \quad (4.15)$$

which is almost identical to the state dependent value contribution term of the BPTT algorithm. However, VGL uses a critic to learn the value gradient additionally to the bootstrapping in the previous algorithms. The combination of bootstrapping and an additional critic is given by $\vec{p}$ with

$$\vec{p} \leftarrow \lambda \underbrace{G'_{k+1}}_{\text{experience}} + (1-\lambda) \underbrace{G_{k+1}}_{\text{learning critic}} . \quad (4.16)$$

The parameter λ in the previous Equation (4.16) refers to a concept called *eligibility traces*. From a mechanistic perspective an eligibility trace corresponds to a temporary memory of previously visited states, taken actions, and occurred rewards. In such a backwards view the currently observed state and its corresponding value is not just critical for improving the latest action taken by the individual but also for previous actions which in consequence led to the current state. However, Equation (4.16) shows that the actual trace is calculated in a forward direction. On the one hand, the extreme case of $\lambda = 1$ results in the trace being solely based on G', the experience based value gradient, which corresponds to a Monte-Carlo methodology. On the other hand, the extreme case of $\lambda = 0$ results in the trace being solely based on G, the learning critic itself which corresponds to bootstrapping as in the one-step temporal difference Q-learning algorithm presented in the previous chapter. In general, introducing eligibility traces stabilises learning and increases convergence in episodes with many iterations and with delayed rewards which is the case in the lifetime model of a foraging predator presented here. I expand on this further in the Discussion.

At each iteration of the algorithm within an episode the incremental weight changes are calculated as follows:

$$\Delta \vec{w} \leftarrow \Delta \vec{w} + \left(\frac{\partial G}{\partial \vec{w}}\right)_k \underbrace{(G'_k - G_k)}_{\text{reinforcement error}} \quad (4.17)$$

for the parameters of critic with $(G'_k - G_k)$ being the reinforcement error between

experienced gradient G' and learned gradient G and

$$\Delta\vec{z} \leftarrow \Delta\vec{z} + \underbrace{\left(\frac{\partial\pi(\vec{s},\vec{z})}{\partial\vec{s}}\right)_k}_{\text{actor}} \underbrace{\left(\left(\frac{\partial r}{\partial\vec{u}}\right)_k + \gamma\left(\frac{\partial f}{\partial\vec{u}}\right)_k G_{k+1}\right)}_{\text{behavioural target gradient}}, \tag{4.18}$$

for the parameters of the actor which is identical to the recurrent element of the state dependent value contribution in BPTT (Equation (4.12)) giving the behavioural effects on the consequent state transition and the long term value. Thus, the actor again uses bootstrapping to learn the value gradient.

The final weight updates for actor and critic occur when the algorithm aggregated the weight changes of each iteration of an episode as follows:

$$\vec{z} \leftarrow \vec{z} + \alpha\Delta\vec{z} \tag{4.19}$$

and

$$\vec{w} \leftarrow \vec{w} + \beta\Delta\vec{w} \tag{4.20}$$

with α and β being learning rates.

The open question still remaining is why use VGL? I already hinted that eligibility traces are beneficial in long running episodes with delayed rewards. Another aspect is the differences in the exploration-exploitation trade-off between BPTT and VGL. Even though there has been no explicitly defined exploration in BPTT the algorithm is always locally exploring different policies as part of the methodology. For global optimality a hill climbing algorithm such as BPTT has to explore the entire state space $\mathbb{S}$ and even for local optimality the BPTT algorithm has to evaluate all adjacent trajectories. The BPTT algorithm therefore requires extensive exploration to learn the value along every adjacent trajectory. This becomes unnecessary if the algorithm learns the value-gradient instead (Figure 4.4). The weight update of the VGL algorithm contains additional information about adjacent trajectories implicitly through the state dependent gradient of the value function as the learning objective. This improves overall convergence and reduces the computational cost of exploration.

4.3.4 Derivatives used by the learning algorithms.

As both BPTT and VGL algorithms are model based they require a number of derivatives of the underlying lifetime model. The lifetime model is implemented as a Markovian decision process and the propagation of incremental gradients backwards through time in both algorithms requires the state and action dependent derivatives of the state transition function f, $\partial f(\vec{s}_k,\vec{u}_k)/\partial\vec{s}$ and

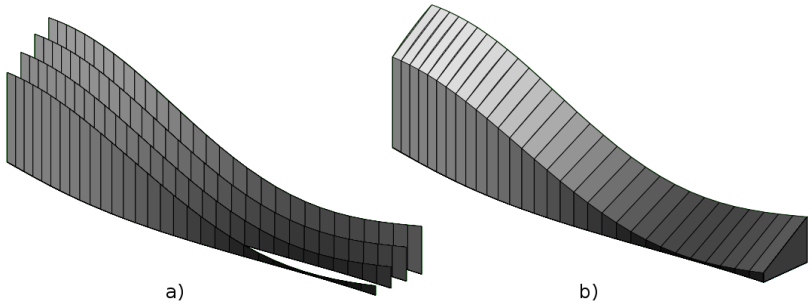


Figure 4.4: The reduction of necessary exploration when using b) VGL over a) BPTT. The VGL algorithm learns a wider range of the state space than the BPTT algorithm as the VGL weight update contains more information about adjacent trajectories. Figure by M. Fairbank taken from Fairbank (2013).

$\partial f(\vec{s}_k, \vec{u}_k)/\partial \vec{u}$ respectively, as follows:

$$\frac{\partial f(\vec{s}_k, \vec{u}_k)}{\partial \vec{u}} = \left\{ \frac{\partial \dot{T}}{\partial \vec{u}}, \frac{\partial \dot{A}}{\partial \vec{u}}, \frac{\partial \dot{X}}{\partial \vec{u}}, \frac{\partial \dot{Y}}{\partial \vec{u}} \right\} \tag{4.21}$$

and

$$\frac{\partial f(\vec{s}_k, \vec{u}_k)}{\partial \vec{s}} = \left\{ \frac{\partial \dot{T}}{\partial \vec{s}}, \frac{\partial \dot{A}}{\partial \vec{s}}, \frac{\partial \dot{X}}{\partial \vec{s}}, \frac{\partial \dot{Y}}{\partial \vec{s}} \right\}. \tag{4.22}$$

Furthermore, both algorithms require the state and action dependent derivative of the payoff function r_{k+1}, $\partial r_{k+1}(\vec{s}_k, \vec{u}_k)/\partial \vec{s}$ and $\partial r_{k+1}(\vec{s}_k, \vec{u}_k)/\partial \vec{u}$ respectively, as follows:

$$\frac{\partial r_{k+1}(\vec{s}_k, \vec{u}_k)}{\partial \vec{u}} = \left\{ \frac{\partial \dot{V}}{\partial \vec{u}} \right\} \tag{4.23}$$

and

$$\frac{\partial r_{k+1}(\vec{s}_k, \vec{u}_k)}{\partial \vec{s}} = \left\{ \frac{\partial \dot{V}}{\partial \vec{s}} \right\}. \tag{4.24}$$

Concerning the previously defined lifetime model the derivatives of the state transition function f are as follows: the state dependent time transition $\partial \dot{T}/\partial \vec{s}$ is defined by the predator's spatial location within the environment and its interactions with the prey present. Otherwise time progresses at constant rate

and independently of age and time itself, i.e.

$$\frac{\partial \dot{T}}{\partial \vec{s}} = \sum_i p_i \frac{\partial g_i(X,Y)}{\partial \vec{s}} \left(t_s + d(t_i)\left(t_{h,i} + t_t t_i^2\right)\right)$$
$$= \begin{cases} \frac{\partial \dot{T}}{\partial T} & = 0 \\ \frac{\partial \dot{T}}{\partial A} & = 0 \\ \frac{\partial \dot{T}}{\partial X} & = \sum_i p_i \left(-g_i(X,Y)(X - x_{i,0})/\sigma_{i,x}^2\right)\left(t_s + d(t_i)\left(t_{h,i} + t_t t_i^2\right)\right) \\ \frac{\partial \dot{T}}{\partial Y} & = \sum_i \underbrace{p_i \left(-g_i(X,Y)(Y - y_{i,0})/\sigma_{i,y}^2\right)}_{\text{encounter with prey}} \underbrace{\left(t_s + d(t_i)\left(t_{h,i} + t_t t_i^2\right)\right)}_{\text{prey handling}}. \end{cases} \tag{4.25}$$

From the definition of $\dot{T}$ in Equation (4.8) the prey specific handling time is a constant term which results in the state dependent time transition being solely affected by the spatially defined chance of encounter with a prey type i following Equation (4.2).

The state dependent derivative of the predator's ageing follows directly from the time transition in Equation (4.25):

$$\frac{\partial \dot{A}}{\partial \vec{s}} = \begin{cases} \frac{\partial \dot{A}}{\partial T} & = 0 \\ \frac{\partial \dot{A}}{\partial A} & = 0 \\ \frac{\partial \dot{A}}{\partial X} & = (1/\lambda_0)\frac{\partial \dot{T}}{\partial X} \\ \frac{\partial \dot{A}}{\partial Y} & = (1/\lambda_0)\frac{\partial \dot{T}}{\partial Y}. \end{cases} \tag{4.26}$$

As such age is simply a scaled aggregation of time.

Other relevant derivatives of the state transition function f are the action dependent changes in the predator's spatial location within the environment. The predator can invest energy e_x, e_y into locomotion with respect to X and Y respectively. The locomotion of the predator itself is bound by a unit length per iteration as follows:

$$\frac{\partial X}{\partial \vec{u}} = \begin{cases} \frac{\partial X}{\partial e_x} & = c_0(1 - (\tanh(c_0 e_x))^2) \\ \frac{\partial X}{\partial e_y} & = 0 \end{cases}$$
$$\frac{\partial Y}{\partial \vec{u}} = \begin{cases} \frac{\partial Y}{\partial e_x} & = 0 \\ \frac{\partial Y}{\partial e_y} & = c_0(1 - (\tanh(c_0 e_y))^2). \end{cases} \tag{4.27}$$

Equation (4.27) follows from Equation (4.8) and shows that the predator can choose the spatial components of its movement independently. Thus, there is no "wind" or "drag" in my model.

By definition of the lifetime model the remaining derivatives of the state transition function f are independent of the state or the predator's actions:

$$\frac{\partial \dot{T}}{\partial \vec{u}} = \frac{\partial \dot{A}}{\partial \vec{u}} = \frac{\partial \dot{X}}{\partial \vec{s}} = \frac{\partial \dot{Y}}{\partial \vec{s}} = \vec{0}, \tag{4.28}$$

with the spatial location of the predator being solely affected by the predator's action. Additionally, time and age progress independently to the predator's investment into locomotion within each iteration.

Next I give the state and action dependent derivatives related to the value function V which is given by the sum of discounted payoffs r along the trajectory of an episode. The derivatives of the incremental changes to the value of an episode along a trajectory are given as follows:

$$\frac{\partial \dot{V}}{\partial \vec{s}} = \begin{cases} \frac{\partial \dot{V}}{\partial T} & = 0 \\ \frac{\partial \dot{V}}{\partial A} & = \dot{\lambda} R(s_k) \\ \frac{\partial \dot{V}}{\partial X} & = \lambda(A) \frac{\partial R(s_k)}{\partial X} - \left(t_0 \frac{\partial \dot{T}}{\partial X} \right) \\ \frac{\partial \dot{V}}{\partial Y} & = \lambda(A) \frac{\partial R(s_k)}{\partial Y} - \left(t_0 \frac{\partial \dot{T}}{\partial Y} \right), \end{cases} \tag{4.29}$$

following the state dependent payoff in Equation (4.4). The time T affects the payoff through the metabolic rate only at a constant rate. It is the age A of a predator which has a varying effect on a predator's payoff. The main components of the state dependent value derivative are the spatial elements where $\partial R(\vec{s})/\partial \vec{s}$ is the derivative of the state dependent payoff from interacting with prey given by

$$\begin{aligned} \frac{\partial R(\vec{s})}{\partial \vec{s}} &= \sum_i p_i \frac{\partial g_i(X,Y)}{\partial \vec{s}} d(t_i)(r_i - t_i^2) \\ &= \begin{cases} \frac{\partial R(\vec{s})}{\partial T} & = 0 \\ \frac{\partial R(\vec{s})}{\partial A} & = 0 \\ \frac{\partial R(\vec{s})}{\partial X} & = \sum_i p_i \left(-g_i(X,Y)(X - x_{i,0})/\sigma_{i,x}^2 \right) d(t_i)(r_i - t_i^2) \\ \frac{\partial R(\vec{s})}{\partial Y} & = \sum_i \underbrace{p_i \left(-g_i(X,Y)(X - x_{i,0})/\sigma_{i,x}^2 \right)}_{\text{encounter with prey}} \underbrace{d(t_i)(r_i - t_i^2)}_{\text{prey payoff}} \end{cases} \end{aligned} \tag{4.30}$$

and

$$\dot{\lambda}(A) = -\frac{1}{(A+1)^2}. \tag{4.31}$$

Just as in the case of the state dependent time transition given by Equation (4.25) the state dependent derivative of the payoff R defined in Equation (4.5) is solely affected by the spatially defined chance to encounter prey with the prey specific payoff being a constant.

A predator's actions affect the value of an episode also directly through the absolute amount of energy invested into locomotion during each iteration $|E(u_k)|$, Equation (4.4), as follows:

$$\frac{\partial \dot{V}}{\partial \vec{u}} = \begin{cases} \frac{\partial \dot{V}}{\partial e_x} & = -\operatorname{sgn}(e_x) \\ \frac{\partial \dot{V}}{\partial e_y} & = -\operatorname{sgn}(e_y), \end{cases} \tag{4.32}$$

with the signum function being defined as

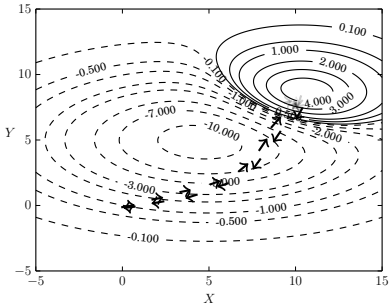
$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases} \tag{4.33}$$

4.3.5 Results.

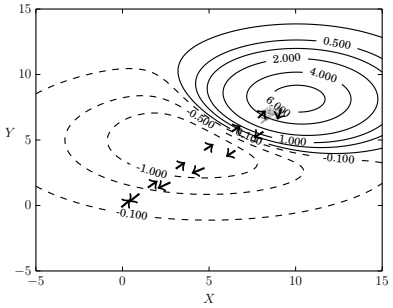
In this chapter I have presented a model of a predator's foraging behaviour which incorporates aspects of a predator's life history traits which have been abstracted away from the models in the previous chapter, such as metabolic costs, locomotion, prey handling, and toxin recovery. The model presented is designed for episodal tasks of finite length such as a day of foraging in an environment with aposematic prey and Batesian mimics. The written simulator generates trajectories of a predator throughout its environment using either back-propagation through time (BPTT) or value gradient learning (VGL) as reinforcement learning algorithms. The simulator is written in `C++` and available for download from GitHub (Teichmann, 2014).

This behavioural model using reinforcement learning assumes payoffs in the form of rewards which do not necessarily reflect the fitness component of a prey item and which are subjective to the individual predators. The aim of the simulator is to find the payoff which generates an observed trajectory or preference for a specific prey type of an individual predator defined by its life history traits.

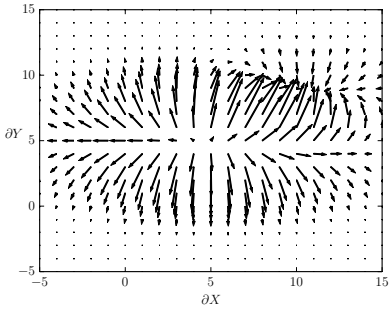
Unfortunately, the simulator does not always generate meaningful trajectories where the predator returns to den. This occurs for different reasons: generally, the learning side of the simulator depends on a great number of additional parameters. Firstly, the actor and critic are implemented as artificial neural networks which are generally difficult to train. In the case of VGL both networks are trained at the same time but are also dependent on each other which can prevent convergence. This is a technical issue which can be circumvented in future work as shown in Fairbank (2013). Additionally, the sign change of the desired functional response of the actor to allow the returning of



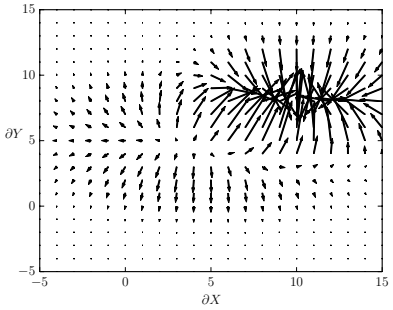
(a) The state dependent reward payoff $R(s_k)$ (Equation (4.5)) for a predator not utilising taste sampling: $d_0 = 0$ and $t_s = 0$.



(b) The state dependent reward payoff $R(s_k)$ (Equation (4.5)) for a predator utilising taste sampling: $d_0 = 1$ and $t_s = 0.1$.

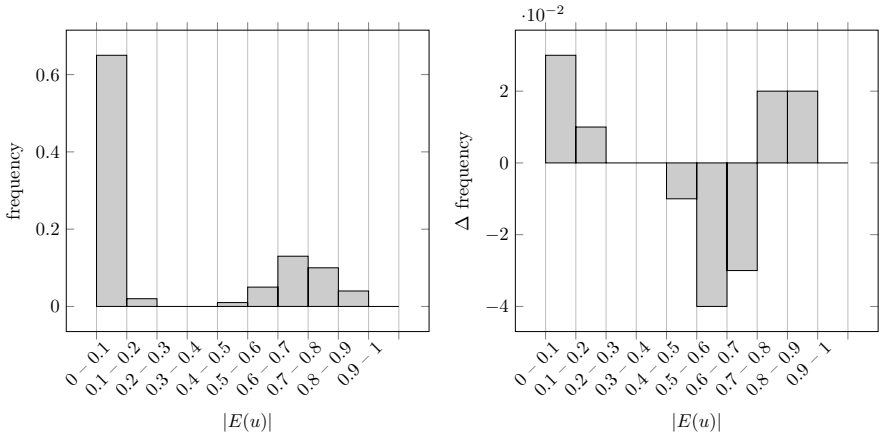


(c) The partial derivative of the reward with respect to the spatial position (Equation (4.30)) of the predator not utilising taste sampling: $d_0 = 0$ and $t_s = 0$.



(d) The partial derivative of the reward with respect to the spatial position (Equation (4.30)) of the predator utilising taste sampling: $d_0 = 1$ and $t_s = 0.1$.

Figure 4.5: The state dependent reward of an exemplary environment with aposematic prey and Batesian mimics. Additionally, (a) and (b) show a realised trajectory of the simulator for a predator using the BPTT algorithm. All cases have $t_h = 0.1$, $t_t = 0.2$, $p_1 = p_2 = 0.5$, $x_{1,0} = y_{1,0} = 5$, $\sigma_{1,x} = 5$, $\sigma_{1,y} = 2.55$, $t_1 = 5$, $r_1 = 1$, $x_{2,0} = 10$, $y_{2,0} = 8$, $\sigma_{2,x} = 2$, $\sigma_{2,y} = 2$, $t_2 = 0$, $r_2 = 15$, $T_l = 80$.



(a) The predator's locomotion optimises the energy expenditure $|E(u)|$ as defined in Equation (4.4) with $\max_e d\dot{X}/de = d\dot{Y}/de = 0.5$ and $\sqrt{\dot{X}(0.5)^2 + \dot{Y}(0.5)^2} = 0.7$.

(b) This plot shows the difference of behavioural energy expenditure $|E(u)|$ in the second half compared to the first half of an episode (Δ). The predator prefers to feed longer and return to its den using a greater step size than the optimal of 0.7.

Figure 4.6: The locomotion profile of a predator not utilising taste sampling with an episode of length $T_l = 80$. (The values are of a single trajectory as shown in Figure 4.5a for a simulation using the BPTT algorithm.)

the predator to its den is a challenging learning problem for an artificial neural network. There is also the question of choosing an appropriate network layout with the right number of nodes in the hidden layer and learning rates. Secondly, the learning problem of the model is by design ill-posed. The model defines a final instantaneous cost which makes the predator return to its den at the end of an episode. Initially, the trajectories returning to the den are suboptimal, usually ending short. Further iterations improve the trajectory in order to avoid the final cost Ψ and to increase the payoff along the trajectory. However, as soon as the learning algorithm successfully finds a trajectory returning to the den the final cost is avoided with $\Psi_l = 0$. At this point the aversiveness of the final cost Ψ_l starts to decay with the continuous rewards from feeding, tempting the predator to overstay in its feeding grounds. Consequently, the trajectory collapses completely with the predator not returning to its den any longer. In summary, the nature of the final cost Ψ results in a continuous loop between trajectories returning to the den in order to avoid the final cost and trajectories staying in the feeding ground until the end of the episode. Consequently, the simulator does not converge on optimality regarding finding a stable trajectory which maximises the overall payoff. This instability could be avoided with a continuous final cost Ψ. On the one hand, the observed instability was desired

to simulate the previously discussed psychological components of rewards which interfere with a linear reward fitness relationship. On the other hand, a continuous penalty function also has pitfalls: in a model with a cut-off at the end of an episode and a continuous final cost Ψ the predator is weighting the rewards from feeding against the final cost and is consequently not fully returning to the den. Such trajectories might be optimal from a computing point of view but make little sense in a biological context. I presume that in a biological context this final cost function has a step like behaviour around the den, e.g. when the individual has offspring to feed in the den the final cost of stopping close to the den will not be smooth. An alternative smooth penalty function which is also biologically meaningful would be an increasing cost for returning late to the den. Unfortunately, such a cost function is difficult to implement with the episodal RL algorithms used by this model here. Both algorithms, BPTT and VGL, are off-line learning algorithms which only issue a weight update at the end of an episode. If the end of an episode is defined by a predator's return rather than by a cut-off time, episodes will be long or even open-ended, particularly in the beginning of the simulation, due to the infinite state space $\mathbb{S}$. Clearly, the solution would be the implementation of psychological components of rewards within a computational RL algorithm using on-line learning. Unfortunately, such an extension was not within the scope of this thesis and has to be addressed in future work.

I conclude that the definition of the penalty function Ψ in my model causes the observed instability of the simulator which can be interpreted as the simulated effects of the interaction of the three components of rewards: associative learning, wanting, and liking. When the predator returns to the den successfully the parts of wanting and liking of the rewards outweigh the aversiveness of the final cost which is decaying. It becomes obvious that the instability of the simulator driven by maximising rewards along the trajectory (excluding the final cost) does not maximise the overall value of a complete trajectory (including the final cost).

The results in Figure 4.5 and Figure 4.6 show trajectories which are close to an optimal trajectory and which were found running the BPTT learning algorithm continuously saving trajectories which increased the overall payoff V for an episode. The environment is composed of an aposematic prey population and a population of Batesian mimics. The predator cannot distinguish between them visually and has to utilise experience from ingesting prey individuals to find a rewarding feeding ground. The trajectory of a predator which is not utilising taste-sampling (Figure 4.5a) shows avoidance of the aversive prey population taking a non-direct route to the population of mimics. The pre-condition of exploration to successfully form and maintain aversion and the non-direct route

result in a very low value for the locomotion parts of the trajectory. In order to make the predator exploit the population of mimics the length of an episode had to be high, with $T_l = 80$ in this simulation. The taste-sampling predator takes a more direct route towards the population of mimics and experiences a much higher value for the locomotion parts of the trajectory (Figure 4.5b).

Figure 4.6 shows the locomotion profile for the non-taste sampling predator for the trajectory presented in Figure 4.5a using BPTT. The predator's locomotion in general is optimised towards efficiency maximising the displacement per energy expenditure $\max_{e_x} d\dot{X}/de_x$ and $\max_{e_y} d\dot{Y}/de_y$ which is at $e_x = e_y = 0.5$ in this simulation with a diagonal locomotion of $\sqrt{\dot{X}(0.5)^2 + \dot{Y}(0.5)^2} = 0.7$ being most efficient. There is a trade-off in this simulation as the population of mimics is not located on the diagonal and due to the presence of an aposematic prey population (Figure 4.6a). Additionally, the predator over-stays in the feeding grounds with the second half of the trajectory showing a more rapid locomotion than the first half (Figure 4.6b).

Even though it is difficult to get meaningful trajectories from the simulator due to the great number of learning related parameters and the instability of the learning task the results presented show some interesting properties:

- in a biological context the trajectories of the predator are unstable due to effects which can be attributed to wanting and liking of rewards in individual based models,
- the element of the model which is generally optimised is the efficiency of locomotion (the behavioural expenditure),
- however, rewards can interfere with this general optimisation of behavioural expenditure. A non-taste sampling predator, for example, avoids the aposematic prey population in order to minimize its metabolic costs from toxin ingestion and
- the predator shows a tendency to over-stay in the feeding grounds and returns to the den with above optimal energy expenditure for locomotion.

4.4 From rewards to Darwinian fitness.

In the previous section I presented a simulator based on my predator lifetime model generating trajectories of individuals based on reinforcement learning and payoffs reflecting subjective rewards. The simulator showed interesting properties around the psychological components of rewards such as wanting and liking which cause a suboptimal maximisation of payoffs along a trajectory ignoring the fitness relevant final cost of not returning to the den at the end of

an episode. On the one hand, the model showed a general tendency to optimise behavioural expenditure. On the other hand, the simulator showed elements which interfered with optimal behaviour such as the predator overstaying in the feeding ground, or avoiding aversive prey to minimise metabolic costs of ingesting toxins.

However, the core of evolutionary models is Darwinian fitness as discussed in the initial motivation of this chapter. In an evolutionary context models such as OFT look at optimal behaviour with regard to maximising fitness. It becomes obvious that the findings of the previous section contradict the idea of maximising fitness with behaving animals showing clear reward driven motivations. Nevertheless, the evolution of behavioural repertoires should maximise fitness.

It is apparent that rewards reflect some fitness component and that there should be a general relation between strength of rewards and fitness. The motivational question was: would it be possible to determine the fitness component of rewards from the environmental set-up and the behaviour of predators.

In a situation with two types of prey a predator could completely prefer one type over the other out of subjective choice. Especially in the previous simulation of an aposematic prey population and a population of mimics the predator shows clear preference for the non-defended prey. However, it is the consequent co-evolution of predator and prey and their ongoing arms-race which allow such complex systems of defence like aposematism in the first place. This implies that prey avoided by an individual predator is also potential prey for that predator within an evolutionary context. In particular the precondition of exploration and the presence of inexperienced predators results in aversive prey experiencing some level of predation as seen in the results of the previous Chapter 3.

It can be assumed that predators generally show evolved behaviour adapted to their environment and that without the occurrence of new mutants selection is of a stabilising nature: the end-result is a stable system of balanced interactions of co-evolved predators and all their prey. I will use the stability argument to infer the fitness components of the previous subjective rewards as follows: for the very reasons of co-evolution and stability some kind of fitness related quantity of interacting predators and prey is assumed to be balanced. The observed environment is interpreted as an evolutionarily stable snapshot without the presence of any mutants with fitness advantages/disadvantages. Figure 4.7 shows the evolutionary model with $t_0\dot{T}$ representing the metabolic cost of the predator, $E(u)$ being the behavioural expenditure (including amongst others locomotion and reproduction costs), and R being the influx of some fitness quantity from predation. I leave the units of the terms open but they could be interpreted as a form of energy and generally $t_o\dot{T} < 0$ and $R > 0$. Under the

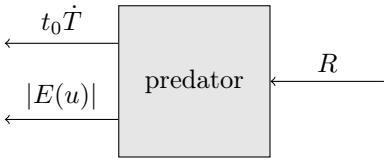


Figure 4.7: The evolutionary model of predator-prey interactions including elements of the predator lifetime model. The predator has an metabolic cost $t_0\dot{T}$ and a behavioural expenditure $|E(u)|$. The costs are balanced by the influx of some fitness related quantity R.

assumption of interim stability without the presence of mutants it follows that

$$t_0\dot{T} - |E(u)| + R = 0. \tag{4.34}$$

If that condition is not met and the l.h.s. of Equation (4.34) is positive the population of predators would grow and if the l.h.s. is negative the population of predators would shrink. In a coupled system of co-evolution this would lead to changing selective pressure on the prey population which is assumed to be stabilising (or either predator or prey would go extinct). For simplicity I assume that the system has reached a stable point of balanced interactions between predator and prey. This seems reasonable under the assumption of a process of co-evolution.

The next step is to reformulate the state transition function using population averages in order to move from individual preferences for rewards to Darwinian fitness. This follows the assumption of this model that evolution acts on the overall behavioural repertoire which maximises fitness and that the average of observed individual behaviour gives an indication of the behavioural repertoire. Additionally, it seems reasonable to assume that the state transition function defined in terms of energy relates to fitness. For the predator, costly behaviour such as having big territories or complicated mating behaviour $E(u)$ will have a negative influence on reproduction. The same is true for predators which spend a long time on handling prey or recovering from toxins $t_o\dot{T}$. Therefore, the state transition function itself can be interpreted in terms of some energy-based fitness quantity assuming a relation between energy and fitness.

Following the previous definition of the lifetime model in Equation (4.2), the total available prey from the prey population i is given by the integral over the prey dispersion as follows:

$$G_i = \int_x \int_y g_i(x, y)\ dx\, dy = 2p_i\pi\sigma_{i,x}\sigma_{i,y}. \tag{4.35}$$

Substituting (4.35) into the previous payoff function of the lifetime model (4.5)

gives the predator's fitness payoff in terms of Darwinian fitness as follows:

$$R = \sum_i G_i d(t_i)(r^* - t_i^2), \tag{4.36}$$

with r^* being the assumed fitness component of the reward payoff of the previous section. I presume that the fitness component r^* of the reward to the predator is related to the fitness of the prey. For example if the unit of fitness is energy the predator has a high fitness influx R from a prey which also had a great amount of energy reserves for reproduction. Finally, if r^* relates to the fitness of prey this value has to be equal for different types of prey under the assumption of stability. If the fitness of a type of prey would be greater than the fitness of other prey types it would be advantageous for the predator to feed exclusively on this prey. It is apparent that fitness in such an interpretation is not equivalent to the number of offspring. r^* is better interpreted as a kind of energy quantity from which individuals can allocate towards the cost of predator defences (Section 1.2.1), reproduction, metabolic costs of toxin ingestion, or behavioural expenditures. I refer to the Discussion (Section 4.5) for an interpretation of r^* in the context of aposematic prey and mimics.

In summary, solving Equation (4.34) for the fitness component r^* for a predator prey interaction with just a single type of prey results in:

$$r^* = \frac{1}{G\, d\, \lambda(A)} \left(E(u) + t_0 + G\left(t_0 t_s + d(t^2\lambda(A) + t_o(t_t + t_h))\right)\right). \tag{4.37}$$

Consequently, r^* needs to be higher to sustain stability when

(i) a predator feeds on toxic prey,

(ii) when the prey requires lengthy handling,

(iii) prey is rare,

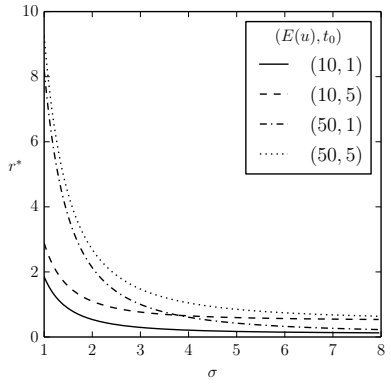
(iv) the predator has a high metabolic rate t_0,

(v) the predator utilises costly behaviour, or

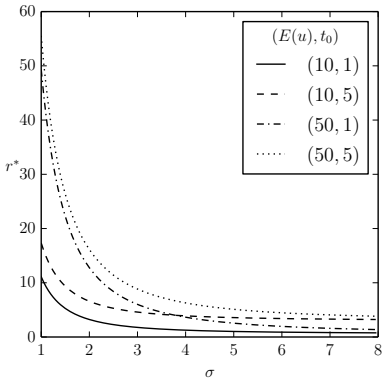
(vi) when predators live longer.

On the predator's side r^* can be termed the nutritional value of prey within this context.

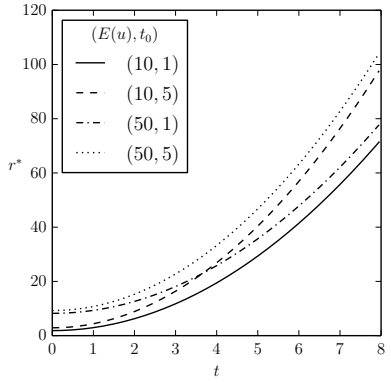
The results in Figure 4.8 show the effects of different aspects of the lifetime model on the nutritional value r^* in the context of a single prey type. We see that an increasing prey abundance σ reduces the required nutritional value of prey for a stable predator population. Nevertheless, there is a minimal nutritional value



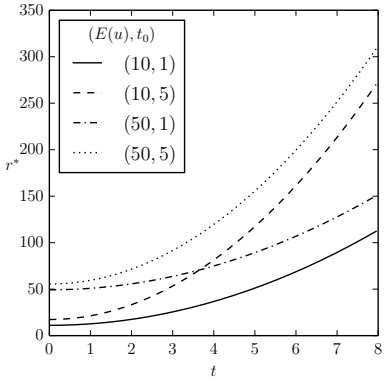
(a) The fitness component r^* of a single prey type with respect to the prey-population's abundance σ, the predator's behavioural costs $|E(u)|$, and the predator's metabolic rate t_0. Without ageing $\lambda(A = 0) = 1$.



(b) The fitness component r^* of a single prey type with respect to the prey-population's abundance σ, the predator's behavioural costs $E(u)$, and the predator's metabolic rate t_0. With age distribution $A = 5$.



(c) The fitness component r^* of a single prey type with respect to the prey-population's toxicity t, the predator's behavioural costs $E(u)$, and the predator's metabolic rate t_0. With $\sigma = 1$, without ageing $\lambda(A = 0) = 1$.



(d) The fitness component r^* of a single prey type with respect to the prey-population's toxicity t, the predator's behavioural costs $E(u)$, and the predator's metabolic rate t_0. With $\sigma = 1$ and age distribution $A = 5$.

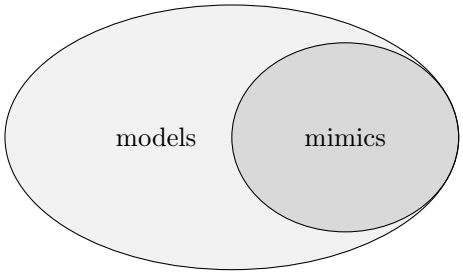
Figure 4.8: Effects of a single aposematic prey population on the required fitness component r^* for there to be an equilibrium of a stable predator-prey environment. Without taste-sampling: $d(t) = 1$, $t_s = 0$, $t_h = 0.1$, and $t_t = 0.1$.

prey must have which depends on the metabolic rate of the predator t_0 and which is independent of the predator's behavioural expenditure $E(u)$ and the prey's abundance σ (Figure 4.8a and 4.8b). If prey is rare the predator's behavioural expenditure $E(u)$ has a much greater impact on r^* than its metabolic rate t_0. This result is in agreement with the predictions of the optimal foraging theory in Section 3.2.1, if prey is rare predator's are believed to be generalists and forage on every prey item they encounter. The metabolic costs of ingesting toxins or mediocre prey is less relevant to a predator's fitness in such a situation than the behavioural expenditure of searching for an alternative prey item. The age distribution or longevity of predators acts as a simple multiplicative factor in this context. In the case of longevity prey has to be more nutritious but the functional shape with regard to prey abundance σ is identical.

Figures 4.8c and 4.8d show the effects of prey toxicity t on the nutritional requirement r^*. Generally, increasing prey toxicity t requires higher nutritional values r^* for stability. In the case of less toxic prey the predator's behavioural expenditure $E(u)$ has again a greater impact on r^* than its metabolic rate. With increasing prey toxicity the predator's metabolic rate has greater impact on r^*. The simulation showed the same result with the non-taste sampling predator avoiding the aversive prey population in order to reduce its metabolic cost from ingesting toxins. This result extends the definition of the optimal foraging theory and shows that predators are predicted to be specialists and avoid prey if their toxicity imposes high metabolic costs for the predator.

With regard to prey toxicity the age distribution or longevity of predators acts not just as a simple factor as in the case of prey abundance. Generally, longevity increases the required nutritional value of prey. Additionally it affects the impact of the prey's toxicity on r^* which weakens for the predator's metabolic costs in the case of low prey toxicity and increases for the predator's behavioural expenditure in the case of high prey toxicity. Consequently, older predators are predicted to be specialists when prey is highly defended and generalists when prey is only weakly defended. A prediction of this model is that the greater the longevity of a predator the clearer should the classification into specialists or generalists become.

As in the previous section using the simulator, I will move to a predator feeding on an aposematic prey in the presence of a Batesian mimic with the following figure illustrating such an environment with two prey populations:



.

The predator cannot distinguish between the two prey populations and has to use experience obtained from a precondition of exploration. As such both prey populations experience some levels of predation. Moving the evolutionary model from Figure 4.7 to multiple food sources i gives the following condition under the assumption of stability:

$$0 = \sum_i R_i - t_0 \dot{T} - E(u), \tag{4.38}$$

with the predator now having multiple sources of fitness influx R_i. As discussed previously, I assume that both types of prey have the same r^* whereas the models allocate parts of their energy inventory towards the cost of their anti-predator defences and mimics have to allocate greater amounts towards reproduction to compensate for higher levels of predation especially in the case of predators able to taste-sample their prey. Solving Equation (4.38) for r^* in the model-mimic system results in:

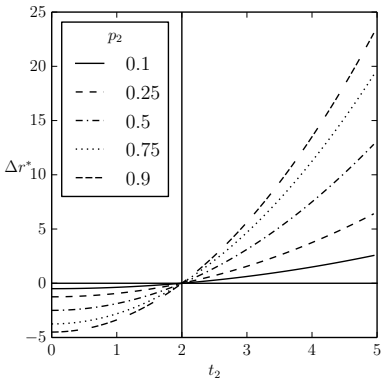
$$r^* = \frac{1}{\sum_i G_i \, d_i \, \lambda(A)} \times \left(E(u) + t_0 + \sum_i G_i \left(t_0 t_s + d_i \left(t_i^2 \lambda(A) + t_o (t_{t,i} + t_{h,i}) \right) \right) \right), \tag{4.39}$$

simply expanding the previous solution in Equation (4.37) to multiple sources of fitness influx R_i.

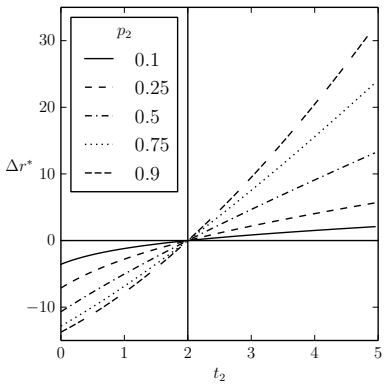
Figure 4.9 shows the results of Equation (4.39) as a function of different parameters of the lifetime model. The overall prey abundance $\sum_i G_i$ is held constant in all charts. Figures 4.9a and 4.9b show the effects of a second aposematic type of prey in comparison to an environment with only one aposematic prey type. With the second aposematic prey being less toxic than the first prey type it overall lowers r^* and vice versa when the second type is more toxic. An increasing fraction of the second prey type p amplifies the effects on r^*. Addi-

tionally, taste-sampling also amplifies the effect of the second prey type on r^*. However, the impact of taste-sampling is greater if the second prey type is less toxic than the first prey type.

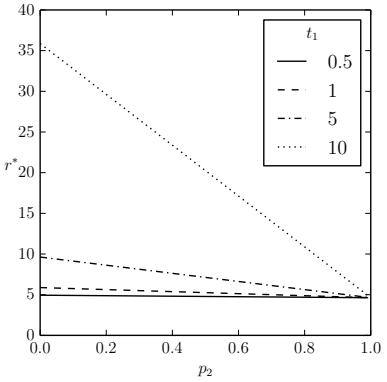
Figures 4.9c and 4.9d show the effects of mimics. Generally, the presence of mimics lowers r^* and mimics have an increasing impact on r^* with increasing toxicity of the aposematic model t_1. In the case of non-taste-sampling predators the effect of mimics on r^* is linear with respect to the fraction of mimics in the overall prey population p_2. Taste-sampling in predators generally increases r^* for stability and the effect of mimics on r^* becomes non-linear with increasing impact in case of mimics being rare.



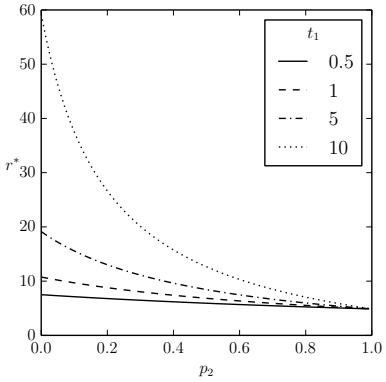
(a) Effects of a second aposematic prey population with respect to its level of defence t_2 and density p_2. The horizontal line $\Delta r^* = 0$ and the vertical line $t_2 = 2$ indicate no differences. Without taste-sampling $d(t) = 1$ and $t_s = 0$. $t_1 = 2$ and $p_1 = 1 - p_2$.



(b) Effects of a second aposematic prey population with respect to its level of defence t_2 and density p_2. The horizontal line $\Delta r^* = 0$ and the vertical line $t_2 = 2$ indicate no differences. With taste-sampling $d_0 = 1$ and $t_s = 0.1$. $t_1 = 2$ and $p_1 = 1 - p_2$.



(c) The effects of mimics within an aposematic prey population with respect to the mimics density p_2 and the models toxicity t_1. Without taste-sampling $d(t) = 1$ and $t_s = 0$. $p_1 = 1 - p_2$.



(d) The effects of mimics within an aposematic prey population with respect to the mimics density p_2 and the models toxicity t_1. With taste-sampling $d_0 = 1$ and $t_s = 0.1$. $p_1 = 1 - p_2$.

Figure 4.9: Effects of aposematic prey expressed as the relative change in the fitness component Δr^* in a stable predator-prey environment with multiple prey populations when compared to an environment with a single prey-population. With $t_h = 0.1$, $t_t = 0.1$, $t_0 = 2.5$, $E(u) = 25$, $\lambda(A = 0) = 1$, and $\sigma = 1$. The total prey abundance $\sum_i G_i$ is held constant.

4.5 Discussion.

In this chapter I presented a predator lifetime model including life history traits which had been abstracted away in the previous chapters, such as metabolic costs, locomotion, prey handling, and toxin recovery. The presented model was defined in such a way that it can be interpreted in a psychological context of subjective behaviour driven by reward motivated objectives and an evolutionary context of a behavioural repertoire which is driven by fitness and co-evolution between predator and prey.

I applied two reinforcement learning algorithms, i.e. back-propagation through time (BPTT) and value gradient learning (VGL), to simulate behaviour of single individuals driven by rewards. Both algorithms address learning in episodal tasks based on experience including discounted future rewards. I used artificial neural networks as universal function approximators for the policy implementation (the actor) and in case of VGL also for the implementation of the value function (the critic). BPTT is computationally cheaper as it does not require a critic for learning. However, BPTT requires a great amount of exploration of the state space for convergence as each learning iteration contains only information about the value of the current trajectory. VGL requires an additional critic. However, VGL is learning the gradient of the value function directly which provides additional information about the value of adjacent trajectories. This reduces the amount of exploration and makes VGL less sensitive towards local optima.

The learning task for the simulator is defined in a way to address the initial discussion of when behaviour is optimal. On the one hand, the environment in the simulation contains rewards and punishment and optimal behaviour should maximise positive reinforcement. On the other hand, the environment contains a fitness related element in the form of an instantaneous final cost in the case when the predator does not return to its den at the end of an episode. From a biological context this penalty function is a steep step-like function: if the predator has to feed offspring in its den then being close to the den will not gradually reduce the cost of not returning. The trajectories from the simulator show a great instability due to the interference of maximising positive reinforcement along the trajectory (excluding the fitness cost) and maximising the value of a complete trajectory (including the fitness cost). The simulator oscillates between two states: (i) a state of maximising rewards along the trajectory excluding the final cost where the predator stays in the feeding ground and does not return to its den and (ii) a state of maximising the value of the complete trajectory including the final cost where the predator successfully returns to its den. I interpreted this as a manifestation of wanting and liking in the simula-

tions of individual behaviour (I note that this effect has been constructed with the choice of the final cost function as the reinforcement learning algorithms used in the simulation do not incorporate psychological effects of rewards and are purely of associative nature).

The simulation shows that the predator generally optimises the efficiency of its behavioural expenditure. But again the rewards interfere with the optimal behaviour as the predator overstays in the feeding grounds and uses above optimal energy for its locomotion on its return to the den. This result coincides with the observations of many studies testing the quantitative predictions of foraging behaviour under the marginal value theorem (See Nonacs (2001) for a summary).

Furthermore, the simulation of a predator which does not utilise taste-sampling shows avoidance of the aversive prey population. The exposure to higher toxicity intake makes the metabolic cost a crucial factor. This result is the same as in the analysis of the fitness quantity r^* which also predicts that the optimisation of the metabolic cost of a predator dominates the behavioural expenditure in environments with highly defended prey.

The fitness quantity is obtained by assuming a stabilising co-evolution between predator and prey and can be interpreted as a form of energy. On the predator's side this might be the nutritional value of prey and on the prey's side it might be interpreted as an energy inventory which the prey can allocate towards the costs of defence and reproduction. Aposematic prey allocates greater amounts towards the cost of its defence whereas the mimics have to allocate greater amounts towards their reproduction due to higher risks of predation from experienced predators. The presence of mimics generally lowers the value of r^* for such a system to be stable. If models and mimics co-exist with an unchanged r^* the prediction is that the models are better defended than in scenarios without mimics. If mimics and models co-exist but with unchanged levels of defence then models are predicted to be smaller and have lower nutritional value than in a system without mimics. Taste-sampling as a strategy increases r^* if mimics are rare or if models are only moderately well defended. However, the impact of taste-sampling is non-linear especially in systems with highly defended models. In such situations taste-sampling lowers the value of r^*. Consequently, under the assumption of a fixed value for r^* and stability, a predator evolves a taste-sampling strategy because mimics are less common or models are better defended than in a comparable stable environment where predators do not utilise taste-sampling.

Another interesting aspect is the effect of different age distributions: in general, longevity in predators increases r^* for stability. The effects are linear with regard to prey abundance but non-linear with regard to prey toxicity where

behavioural expenditure gains increasing impact in the case of defended prey and older predators. Alternatively, metabolic costs have an increased impact in the case of non-defended prey. This finding predicts that longevity creates a clear classification of predators into generalists or specialists depending on the toxicity of their prey.

Reflecting on the initial motivation of this chapter, the lifetime model concludes with a recognised discrepancy between reward driven behaviour and maximising fitness. The definition of a final instantaneous cost as part of the lifetime model simulated the interference of fitness related components of the rewards with the additional psychological aspects of wanting and linking. Particularly within a biological context, reinforcement learning generates unstable trajectories due to interfering aspects of maximising rewards along a trajectory and maximising the value of a complete trajectory (including fitness components like the instantaneous final cost in this model). The lifetime model showed that the predator in the simulation generally optimises its behavioural expenditure (locomotion in this model) at low toxin intake. However, rewards, on the one hand, can interfere with this optimisation with the predator overstaying in its feeding grounds. On the other hand, a predator in the simulation optimises its metabolic cost rather than its behavioural expenditure at increasing toxin intake from highly defended prey. This results are equivalent with the predictions of the evolutionary lifetime model and optimal foraging theory.

I conclude that behavioural repertoires allow the derivation of a fitness related quantity r^* under the assumption of co-evolution and stabilising selection. On the predator's side this quantity is related to the nutritional value of prey and on the prey's side it relates to an energy inventory which can be allocated amongst others towards the cost of defences or reproduction.

Summarising, the main conclusions of this chapter are that the simulation of an individual predator allows subjective reward driven trajectories which interfere with the maximisation of the overall value of a trajectory and contradict predictions of the evolutionary lifetime model. However, many aspects of subjective trajectories are also predicted by behavioural repertoires under selective pressure and stabilising co-evolution of predators and their prey. These are as follows:

- the behavioural expenditure has a greater impact than metabolic costs when prey is rare and undefended in which case predators are predicted to be generalists,
- alternatively, the metabolic costs have a greater impact when prey is abundant or highly defended where predators are predicted to be specialists.
- Longevity in predators generally increases the nutritional value of prey

items required for stability.

- Additionally, longevity increases the importance of the behavioural expenditure in case of highly defended prey and the impact of metabolic costs if prey is undefended. This finding suggests that longevity creates a clear classification of predators into generalists or specialists.
- Finally, mimics generally lower r^* which leads to less nutritional prey or better defended models for stability under the assumption of a fixed value for r^*
- and predators utilise taste-sampling if mimics are rare or models are highly toxic.

In conclusion, the results of the two interpretations of the lifetime model, rewards and fitness respectively, allow to distinguish which parts of observed individual behaviour are caused by subjective preferences for rewards including psychological aspects of wanting and liking, and which are actually part of an evolved behavioural repertoire which maximises fitness. Even though Section 4.3 does not conclude fully satisfactorily, the lifetime of this chapter has shown some great opportunities for such a combined approach. Many predictions made by the lifetime model are reasonable or could be tested, i. e. the effects of longevity on predator classification into generalists and specialists.

Chapter 5

Conclusions.

> A conclusion is the place where you got tired thinking.
>
> (Martin H. Fischer)

5.1 Summary.

The aim of this thesis was to develop models of aposematism addressing open questions beyond its initial evolution. I presented a wide body of work on two main aspects revolving around: 1.) the properties of aposematic solutions in finite populations and their stability under the influence of drift, and 2.) reinforcement learning as an implementation of the predator's aversive learning to generalise from encounters with prey items to optimal foraging behaviour.

I started this thesis from the prey perspective by introducing a new and more flexible methodology for assessing the evolutionary dynamics and stability of the co-evolution of secondary defences and signalling of such defences in chapter 2. I extended a previous game-theoretical framework with a discussion of finite population size and the resulting drift as an additional evolutionary force affecting aposematic solutions and their stability. The main conclusions were:

- Drift is an important aspect of real population systems and leads to an increased inter-population diversification of aposematic solutions. It results in a wide region of possible levels of defence and signalling of such defence comparable to an evolutionarily stable set where strategies can change within the solution space due to approximately neutral drift, but resist invasion from outside.

- For aposematism to evolve in finite populations the number of mutants needs to be relatively large. However, stability of aposematic solutions

is tightly bound to selection against rare prey types (anti-apostatic selection) due to the predator's aversive learning which prevents cheating and intra-population differences (e.g. automimicry and continuous variation in defences). In particular, an aposematic prey population has to look alike or its level of aversiveness decreases and aposematism loses its advantage.

- An accelerated learning process of strong aversion in predators allows for a negative correlation between the strength of signals and defence where aversive prey can reduce their defences with increasing conspicuousness of their signals.

In particular the last two conclusions illustrate the importance of the predator's learning process. The details of the predator's generalisation of aversive information are not only decisive for stability, they are critical to the relation of defence and signals in aposematism.

Following these insights the thesis investigated the predator's aversive learning within models of aposematism. As the selective agent the predator co-evolves with its prey in a continues arms-race and the evolution and details of its aversive learning process attracted my scientific attention for the majority of this thesis.

Chapter 3 introduced operant conditioning as a generalised theory of associative learning and Q-learning as a methodology implementing the reinforcement learning problem where a predator receives feedback depending on its actions. To begin with I applied Q-learning to an optimal diet model to investigate the link between aposematism, aversive learning, and optimal foraging behaviour. The main conclusions were as follows:

- Temporal difference learning is a suitable and elegant approach to optimal foraging in the context of aposematic prey, Batesian mimics, and uncertain environments.

- A pre-condition of continuous exploration in changing environments or situations of conflicting rewards results in aversive prey experiencing some level of predation as part of the predator's aversive memory formation. This can cause foraging behaviour which is conditionally suboptimal in a stationary environment. However, continuous exploration is a good policy precisely because environments are inherently uncertain.

- The model reproduced many expected results of established models and allowed new insights into the effects of uncertainty and changing environments.

Following the successful application of Q-learning to optimal foraging in the presence of aposematic prey I looked into the initial evolution of associative

learning. Comparing the fitness distributions of an adaptive learning strategy with a non-adaptive strategy of random mutations with regard to regularity, frequency, and size of changes it showed that:

- A learning strategy incurs the cost of exploration and requires environmental change and longer generation times to be beneficial. Learning is optimal for specific combinations of regularity and size of environmental change.
- Regularity is the only environmental factor which impacts whether learning is generally advantageous.
- The fitness distributions of the learning strategy are independent of technical parameters of reinforcement learning such as learning and discount rate within a biological context of changing environments.

Chapter 3 showed that Q-learning can be applied successfully to models of aposematism to investigate how a predator includes aversive information from an encounter with aposematic prey into generalised foraging behaviour. Additionally, Q-learning was shown to be a generally advantageous adaptation to changing environments allowing its initial evolution.

However, the chapter abstracted away many aspects of the aposematic predator-prey interaction. In particular, it made the broad assumption of a monotonically increasing functional relation between rewards and fitness. Chapter 4 added some components of established models of aposematism such as life history traits of the predator and prey handling to the previous model of optimal foraging. Additionally, it discussed optimal behaviour in both contexts of maximising rewards and maximising fitness using a predator lifetime model. In summary the main conclusions were:

- There is a widely recognised discrepancy between individual behaviour driven by reward motivated objectives and the maximisation of fitness. This phenomenon was observable in the simulated trajectories of the predator lifetime model where an instability caused by wanting and liking of rewards intervenes with, firstly, the aversiveness of not meeting a constraint of returning to the den and, secondly, with the optimal behavioural expenditure of overstaying in the feeding ground.
- The predator traits off against the metabolic cost and its behavioural expenditure depending on the toxicity of aposematic prey and the presence of Batesian mimics.
- The generalisation to behavioural repertoires and the assumption of the co-evolution of predators and their prey under stabilising selection allows

us to infer a fitness related energy quantity as a subcomponent of the rewards in the predator lifetime model. This quantity can be interpreted as a nutritional value of prey items and as an energy inventory from which prey can allocate towards the cost of their defence and reproduction.

- In both contexts, individual reward motivated behaviour and fitness driven behavioural repertoires, the model makes similar predictions about the impact of metabolic cost and behavioural expenditure on the predator's optimal foraging behaviour.
- Additionally, the model makes interesting predictions about the effects of taste-sampling and longevity on the composition of a stable aposematic predator-prey system.

5.2 Future work.

Even though this thesis gives important insights into new aspects of aposematism, naturally, it cannot be complete. In particular, it only touched the surface of the role of the predator as the selective agent. There are many options for future work which are, amongst others:

- Adding population dynamical interactions to build a complete evolutionary model of learning.
- Adding explicit psychological elements of rewards to a reinforcement learning method to investigate the impact of wanting and liking on associative learning.
- Include constraints from innate behaviour such as returning to the den as a limitation to the state and action space in the RL model.
- Compare the exploration patterns of a learning predator with Levy-flight patterns.
- Allowing the aposematic prey to evolve the amount of energy it allocates towards the costs of aposematism and reproduction in the model of co-evolving behavioural repertoires.
- Building a complete model with co-evolving defence and signalling of such a defence and co-evolving predators with their prey using associative learning to generalise information from encounters with prey into optimal foraging behaviour.
- Investigating different reward fitness relations and their consequences on the co-evolution of predators and their prey in the context of aposematism.

- Formulating a consistent definition of evolutionary stability of aposematic solutions including associative learning and the pre-condition of exploration.

Bibliography

Ackley, David and Michael Littman. 1991. *Interactions between learning and evolution*, Artificial life II **10**, 487–509.

Albasser, Mathieu M., Julie R. Dumont, Eman Amin, Joshua D. Holmes, Murray R. Horne, John M. Pearce, and John P. Aggleton. 2013. *Association rules for rat spatial learning: The importance of the hippocampus for binding item identity with item location*, Hippocampus **23**, 1162–1178.

Alonso, Eduardo, Mark D'inverno, Daniel Kudenko, Michael Luck, and Jason Noble. 2001. *Learning in multi-agent systems*, The Knowledge Engineering Review **16**, 277–284.

Alonso, Eduardo and Esther Mondragón. 2006. *Associative learning for reinforcement learning: where animal learning and machine learning meet*, Proceedings of the fifth symposium on adaptive agents and multi-agent systems.

Alonso, Eduardo and Nestor Schmajuk. 2012. *Special issue on computational models of classical conditioning guest editors' introduction*, Learning & Behavior **40**, 231–240.

Bagnell, Andrew J., Sham M. Kakade, Jeff G. Schneider, and Andrew Y. Ng. 2003. *Policy search by dynamic programming*, Advances in neural information processing systems.

Barnett, Craig A., Melissa Bateson, and Candy Rowe. 2007. *State-dependent decision making: educated predators strategically trade off the costs and benefits of consuming aposematic prey*, Behavioral Ecology **18**, 645–651.

Barto, Andrew G., Richard S. Sutton, and Chris Watkins. 1989. *Learning and sequential decision making*, Learning and computational neuroscience.

Barton, Nicholas H. and Brian Charlesworth. 1984. *Genetic revolutions, founder effects, and speciation*, Annual Review of Ecology and Systematics **15**, 133–164.

Berns, Gregory S., Samuel M. McClure, Giuseppe Pagnoni, and Read P. Montague. 2001. *Predictability modulates human brain response to reward*, The Journal of Neuroscience **21**, 2793–2798.

Berridge, Kent C. 1996. *Food reward: brain substrates of wanting and liking*, Neuroscience & Behavioral Reviews **20**, 1–25.

———. 2003. *Pleasures of the brain*, Brain and Cognition **52**, 106–128.

Berridge, Kent C., Terry E. Robinson, and Wayne J. Aldridge. 2009. *Dissecting components of reward: liking, wanting, and learning*, Current Opinion in Pharmacology **9**, 65–73.

Blaisdell, Aaron P., James C. Denniston, Hernán I. Savastano, and Ralph R. Miller. 2000. *Counterconditioning of an overshadowed cue attenuates overshadowing*, Journal of Experimental Psychology: Animal Behavior Processes **26**, 74–83.

Blount, Jonathan D., Michael P. Speed, Graeme D. Ruxton, and Philip A. Stephens. 2009. *Warning displays may function as honest signals of toxicity*, Proceedings of the Royal Society B: Biological Sciences **276**, 871–877.

Bouton, Mark E. 2007. *Learning and behavior: A contemporary synthesis*, Sinauer Associates.

Breland, Keller and Marian Breland. 1961. *The misbehavior of organisms*, American Psychologist **16**, 681–689.

Broom, Mark, Graeme D. Ruxton, and Michael P. Speed. 2008. *Evolutionarily stable investment in anti-predatory defences and aposematic signalling* in Mathematical modeling of biological systems. Birkhäuser Boston.

Broom, Mark, Mike P. Speed, and Graeme D. Ruxton. 2006. *Evolutionarily stable defence and signalling of that defence*, Journal of Theoretical Biology **242**, 32–34.

Carew, Thomas J., Robert D. Hawkins, and Eric R. Kandel. 1983. *Differential classical conditioning of a defensive withdrawal reflex in aplysia californica*, Science **219**, 397–400.

Charnov, Eric L. 1976. *Optimal foraging, the marginal value theorem*, Theoretical Population Biology **9**, 129–136.

Christiansen, Freddy B. 1991. *On conditions for evolutionary stability for a continuously varying character*, American Naturalist **138**, 37–50.

Clark, Colin W. and Marc Mangel. 2000. *Dynamic state variable models in ecology: Methods and applications: Methods and applications*, Oxford University Press, USA.

Darst, Catherine R. 2006. *Predator learning, experimental psychology and novel predictions for mimicry dynamics*, Animal Behavior **71**, 743–748.

Darst, Catherine R., Molly E. Cummings, and David C. Cannatella. 2006. *A mechanism for diversity in warning signals: conspicuousness versus toxicity in poison frogs*, Proceedings of the National Academy of Sciences **103**, 5852–5857.

Daw, Nathaniel D. and Kenji Doya. 2006. *The computational neurobiology of learning and reward*, Current Opinion in Neurobiology **16**, 199–204.

Dayan, Peter and Bernard W. Balleine. 2002. *Reward, motivation, and reinforcement learning*, Neuron **36**, 285–298.

Dayan, Peter and Nathaniel D. Daw. 2008. *Decision theory, reinforcement learning, and the brain*, Cognitive, Affective, & Behavioral Neuroscience **8**, 429–453.

Dayan, Peter and Yael Niv. 2008. *Reinforcement learning: the good, the bad and the ugly*, Current Opinion in Neurobiology **18**, 185–196.

DeWitt, Thomas J., Andrew Sih, and David S. Wilson. 1998. *Costs and limits of phenotypic plasticity*, Trends in Ecology & Evolution **13**, 77–81.

Dingemanse, Niels J. and Denis Réale. 2005. *Natural selection and animal personality*, Behavior **142**, 1159–1184.

Doya, Kenji. 2007. *Reinforcement learning: Computational theory and biological mechanisms*, Human Frontiers Science Program Journal **1**, 30–40.

Edmunds, Malcolm. 1974. *Defence in animals: a survey of anti-predator defences*, Longman Harlow.

Ellegren, Hans. 2009. *A selection model of molecular evolution incorporating the effective population size*, Evolution **63**, 301–305.

Fairbank, Michael. 2013. *Value-gradient learning*, Ph.D. Thesis.

Fisher, Ronald A. 1930. *The genetical theory of natural selection*, Clarendon Press.

Gamberale-Stille, Gabriella and Tim Guilford. 2004. *Automimicry destabilizes aposematism: predator sample-and-reject behaviour may provide a solution*, Proceedings of the Royal Society B: Biological Sciences **271**, 2621–2625.

Gamberale-Stille, Gabriella and Birgitta S. Tullberg. 1996. *Evidence for a peak-shift in predator generalization among aposematic prey*, Proceedings of the Royal Society B: Biological Sciences **263**, 1329–1334.

_____. 2001. *Fruit or aposematic insect? context-dependent colour preferences in domestic chicks*, Proceedings of the Royal Society B: Biological Sciences **268**, 2525–2529.

Gillespie, John H. 2001. *Is the population size of a species relevant to its evolution?*, Evolution **55**, 2161–2169.

Glimcher, Paul W., Ernst Fehr, Colin Camerer, and Russell Alan Poldrack. 2008. *Neuroeconomics: Decision Making and the Brain*, Academic Press.

Guilford, Tim. 1994. *Go-slow signalling and the problem of automimicry*, Journal of Theoretical Biology **170**, 311–316.

Hagen, Edward H., Richard J. Sullivan, Robert Schmidt, Genela Morris, Richard Kempter, and Peter Hammerstein. 2009. *Ecology and neurobiology of toxin avoidance and the paradox of drug reward*, Neuroscience **160**, 69–84.

Hall, Geoffrey. 2002. *Associative structures in pavlovian and instrumental conditioning* in Stevens' Handbook of Experimental Psychology. John Wiley & Sons.

Haselgrove, Mark and Lee Hogarth. 2013. *Clinical applications of learning theory*, Psychology Press.

Holen, Øistein H. 2013. *Disentangling taste and toxicity in aposematic prey*, Proceedings of the Royal Society B: Biological Sciences **280**, 201–209.

Houston, Alasdair I. and John McNamara. 1982. *A sequential approach to risk-taking*, Animal Behaviour **30**, 1260–1261.

Huxley, Julian. 1942. *Evolution. The Modern Synthesis*, George Allen & Unwin.

Janzen, Daniel H. 1980. *When is it coevolution*, Evolution **34**, 611–612.

Johnston, Amy N. B. and Thomas H. J. Burne. 2008. *Aposematic colouration enhances memory formation in domestic chicks trained in a weak passive avoidance learning paradigm*, Brain Research Bulletin **76**, 313–316.

Johnston, Timothy D. 1982. *The selective costs and benefits of learning: an evolutionary analysis*, Advances in the Study of Behavior **12**, 65–106.

Johnston, Timothy D. and M. T. Turvey. 1981. *A sketch of an ecological metatheory for theories of learning*, The psychology of Learning and Motivation **14**, 147–205.

Keehn, John D. 1959. *The effect of a warning signal on unrestricted avoidance behaviour*, British Journal of Psychology **50**, 125–135.

Krebs, John R. 1980. *Optimal foraging, predation risk and territory defence*, Ardea **68**, 83–90.

Lee, Thomas J., Nicola M. Marples, and Michael P. Speed. 2010. *Can dietary conservatism explain the primary evolution of aposematism?*, Animal Behaviour **79**, 63–74.

Lee, Thomas J., Michael P. Speed, and Philip A. Stephens. 2011. *Honest signaling and the uses of prey coloration*, American Naturalists **178**, 1–9.

Leimar, Olof, Magnus Enquist, and Birgitta S. Tullberg. 1986. *Evolutionary stability of aposematic coloration and prey unprofitability: A theoretical analysis*, American Naturalists **128**, 469–490.

Lev-Yadun, Simcha and Kevin S. Gould. 2007. *What do red and yellow autumn leaves signal?*, The Botanical Review **73**, 279–289.

Lindström, Leena, Rauno V. Alatalo, Anne Lyytinen, and Johanna Mappes. 2001. *Strong antiapostatic selection against novel rare aposematic prey*, Proceedings of the National Academy of Sciences **98**, 91–102.

Longson, Chris G. and Jean M. P. Joss. 2006. *Optimal toxicity in animals: predicting the optimal level of chemical defences*, Functional Ecology **20**, 731–735.

MacArthur, Robert H. and Eric R. Pianka. 1966. *On optimal use of a patchy environment*, American Naturalist **100**, 603–609.

Mackintosh, Nicholas John. 1974. *The psychology of animal learning*, Academic Press.

_____. 1994. *Animal learning and cognition*, Academic Press.

Macphail, Euan M. 1982. *Brain and intelligence in vertebrates*, Claredon Press.

Maia, Tiago V. 2009. *Reinforcement learning, conditioning, and the brain: Successes and challenges*, Cognitive, Affective, & Behavioral Neuroscience **9**, 343–364.

Mangel, Marc and Colin W. Clark. 1986. *Towards a unified foraging theory*, Ecology **67**, 1127–1138.

Mappes, Johanna, Nicola Marples, and John A. Endler. 2005. *The complex business of survival by aposematism*, Trends in Ecology and Evolution **20**, 598–603.

Marples, Nicola M., David J. Kelly, and Robert J. Thomas. 2005. *Perspective: The evolution of warning coloration is not paradoxical*, Evolution **59**, 933–940.

Masel, Joanna. 2011. *Genetic drift*, Current Biology **21**, 837–838.

Matsushima, Toshiya, Ai Kawamori, and Tiaza Bem-Sojka. 2008. *Neuro-economics in chicks: Foraging choices based on amount, delay and cost*, Brain Research Bulletin **76**, 245–252.

Maynard Smith, John. 1974. *The theory of games and the evolution of animal conflicts*, Journal of Theoretical Biology **47**, 209–221.

Mazur, James E. and Theresa A. Ratti. 1991. *Choice behavior in transition: Development of preference in a free-operant procedure*, Animal Learning & Behavior **19**, 241–248.

McNamara, John M., Richard F. Green, and Ola Olsson. 2006. *Bayes' theorem and its applications in animal behaviour*, Oikos **112**, 243–251.

McNamara, John M. and Alasdair I. Houston. 1985. *Optimal foraging and learning*, Journal of Theoretical Biology **117**, 231 –249.

Mery, Frederic and Tadeusz J. Kawecki. 2004. *An operating cost of learning in drosophila melanogaster*, Animal Behaviour **68**, 589–598.

Mitchell, Tom. 1997. *Machine learning*, McGraw-Hill Education.

Montague, Read P., Peter Dayan, and Terrence J. Sejnowski. 1996. *A framework for mesencephalic dopamine systems based on predictive hebbian learning*, The Journal of Neuroscience **16**, 1936–1947.

Montague, Read P., Steven E. Hyman, and Jonathan D. Cohen. 2004. *Computational roles for dopamine in behavioural control*, Nature **431**, 760–767.

Moran, Patrick A. P. 1962. *The statistical processes of evolutionary theory*, Clarendon Press.

Nash, John. 1951. *Non-cooperative games*, Annals of mathematics, 286–295.

Niv, Yael, Daphna Joel, and Peter Dayan. 2006. *A normative perspective on motivation*, Trends in Cognitive Sciences **10**, 375–381.

Niv, Yael. 2009. *Reinforcement learning in the brain*, Journal of Mathematical Psychology **53**, 139–154.

Nonacs, Peter. 2001. *State dependent behavior and the marginal value theorem*, Behavioral Ecology **12**, 71–83.

Nowak, Martin A. 2006. *Evolutionary dynamics: exploring the equations of life*, Belknap Press.

Ollason, Janet G. 1980. *Learning to forage optimally*, Theoretical Population Biology **18**, 44 –56.

Osborn, Henry F. 1896. *From the greeks to darwin: an outline of the development of the evolution idea*, Macmillan.

Parker, Geoffrey A. 1978. *Searching for mates*, Behavioural Ecology **1**, 214–244.

Pearce, John M. and Mark E. Bouton. 2001. *Theories of associative learning in animals*, Annual Review of Psychology **52**, 111–139.

Pearce, John M. 2013. *Animal learning and cognition: an introduction*, Psychology Press.

Perry, Gad and Eric R. Pianka. 1997. *Animal foraging: past, present and future*, Trends in Ecology & Evolution **12**, 360–364.

Peters, Jan and Stefan Schaal. 2008. *Natural actor-critic*, Neurocomputing **71**, 1180–1190.

Pigliucci, Massimo. 2001. *Phenotypic plasticity: beyond nature and nurture*, Johns Hopkins University Press.

Platt, Michael L. and Scott A. Huettel. 2008. *Risky business: the neuroeconomics of decision making under uncertainty*, Nature Neuroscience **11**, 398–403.

Poulton, Edward B. 1890. *The colours of animals: their meaning and use, especially considered in the case of insects*, D. Appleton and Company.

Pyke, Graham H. 1984. *Optimal foraging theory: a critical review*, Annual Review of Ecology and Systematics, 523–575.

Rangel, Antonio, Colin Camerer, and Read P. Montague. 2008. *A framework for studying the neurobiology of value-based decision making*, Nature Reviews Neuroscience **9**, 545–556.

Robinson, Michael H. 1969. *Defenses against visually hunting predators*, Evolutionary Biology **3**, 225–259.

Robinson, Terry E. and Kent C. Berridge. 2003. *Addiction*, Annual Review of Psychology **54**, 25–53.

Rummery, Gavin A. and Mahesan Niranjan. 1994. *On-line Q-learning using connectionist systems*, Cambridge University Press.

Ruxton, Graeme D., Tom N. Sherratt, and Mike P. Speed. 2004. *Avoiding attack: The evolutionary ecology of crypsis, warning signals and mimicry*, Oxford University Press.

Ruxton, Graeme D., Mike P. Speed, and Mark Broom. 2009. *Identifying the ecological conditions that select for intermediate levels of aposematic signalling*, Evolutionary Ecology **23**, 491–501.

Jones, Rebecca, Sian Davis, and Mike P. Speed. 2013. *Defence cheats can degrade protection of chemically defended prey*, Ethology **119**, 52–57.

Schachtman, Todd R. and Steve S. Reilly. 2011. *Associative learning and conditioning theory: Human and non-human applications*, Oxford University Press.

Shettleworth, Sara J. 1999. *Cognition, evolution, and behavior*, Oxford University Press.

Schultz, Wolfram. 2002. *Getting formal with dopamine and reward*, Neuron **36**, 241–263.

_____. 2007. *Multiple dopamine functions at different time courses*, Annual Review of Neuroscience **30**, 259–288.

_____. 2008. *Introduction. neuroeconomics: the promise and the profit*, Philosophical Transactions of the Royal Society B: Biological Sciences **363**, 3767–3769.

Schultz, Wolfram, Peter Dayan, and P. Read Montague. 1997. *A neural substrate of prediction and reward*, Science **275**, 1593–1599.

Sclafani, Anthony. 1990. *Nutritionally based learned flavor preferences in rats* in Taste, experience, and feeding. American Psychological Association.

_____. 2004. *Oral and postoral determinants of food reward*, Physiology and Behavior **81**, 773–779.

Shanks, David R. 1995. *The psychology of associative learning*, Cambridge University Press.

Sherratt, Thomas N. 2002. *The coevolution of warning signals*, Proceedings of the Royal Society of London. Series B: Biological Sciences **269**, 741–746.

_____. 2003. *State-dependent risk-taking by predators in systems with defended prey*, Oikos **103**, 93–100.

Sih, Andrew and Bent Christensen. 2001. *Optimal diet theory: when does it work, and when and why does it fail?*, Animal Behaviour **61**, 379–390.

Skelhorn, John and Candy Rowe. 2006. *Prey palatability influences predator learning and memory*, Animal Behaviour **71**, 1111 –1118.

Speed, Mike P. and Graeme D. Ruxton. 2005. *Aposematism: what should our starting point be?*, Proceedings of the Royal Society B: Biological Sciences **272**, 431–438.

_____. 2007. *How bright and how nasty: explaining diversity in warning signal strength*, Evolution **61**, 623–635.

Speed, Mike P. 2000. *Warning signals, receiver psychology and predator memory*, Animal Behaviour **60**, 269–278.

Speed, Mike P., Graeme D. Ruxton, John D. Blount, and Philip A. Stephens. 2010. *Diversification of honest signals in a predator–prey system*, Ecology Letters **13**, 744–753.

Staddon, John E. R. and Virginia L. Simmelhag. 1971. *The supersitition experiment: A reexamination of its implications for the principles of adaptive behavior*, Psychological Review **78**, 3–43.

Staddon, John E. R. and Daniel T. Cerutti. 2003. *Operant conditioning*, Annual Review of Psychology **54**, 115–144.

Stephens, David W. 1991. *Change, regularity, and value in the evolution of animal learning*, Behavioral Ecology **2**, 77–89.

_____. 1986. *Foraging theory*, Princeton University Press.

Stephens, David W. and Eric L. Charnov. 1982. *Optimal foraging: some simple stochastic models*, Behavioral Ecology and Sociobiology **10**, 251–263.

Stephens, David W. and John R. Krebs. 1987. *Foraging theory*, Princeton University Press.

Summers, Kyle and Mark E. Clough. 2001. *The evolution of coloration and toxicity in the poison frog family (dendrobatidae)*, Proceedings of the National Academy of Sciences **98**, 6227–6238.

Sutton, Richard S. and Andrew G. Barto. 1998. *Reinforcement learning: An introduction*, Cambridge University Press.

Svádová, Kateřina, Alice Exnerová, Pavel Štys, Eva Landová, Jan Valenta, Anna Fučíková, and Radomír Socha. 2009. *Role of different colours of aposematic insects in learning, memory and generalization of naïve bird predators*, Animal Behaviour **77**, 327–336.

Taylor, Christine, Drew Fudenberg, Akira Sasaki, and Martin A. Nowak. 2004. *Evolutionary game dynamics in finite populations*, Bulletin of Mathematical Biology **66**, 1621–1644.

Teichmann, Jan. 2014. *A foraging simulator based on td learning: Source available at https://github.com/teichmaj/PredatorReinforcementLearningSimulator.*

Teichmann, Jan, Mark Broom, and Eduardo Alonso. 2014a. *The application of temporal difference learning in optimal diet models*, Journal of theoretical biology **340**, 11–16.

_____. 2014b. *The evolutionary dynamic of aposematism: a numerical analysis of co-evolution in finite populations*, Mathematical Models of Natural Phenomena **9**, 148–164.

Thomas, Randal J., Nicola M. Marples, Innes C. Cuthill, Masaya Takahashi, and Edward A. Gibson. 2003. *Dietary conservatism may facilitate the initial evolution of aposematism*, Oikos **101**, 458–466.

Tollrian, Ralph and Drew C. Harvell. 1999. *The ecology and evolution of inducible defenses*, Princeton University Press.

Valentin, Vivian V., Anthony Dickinson, and John P. O'Doherty. 2007. *Determining the neural substrates of goal-directed learning in the human brain*, The Journal of Neuroscience **27**, 4019–4026.

Via, Sara, Richard Gomulkiewicz, Gerdien De Jong, Samuel M. Scheiner, Carl D. Schlichting, and Peter H. Van Tienderen. 1995. *Adaptive phenotypic plasticity: consensus and controversy*, Trends in Ecology & Evolution **10**, 212–217.

Watkins, Chris. 1989. *Learning from delayed rewards*, Ph.D. Thesis.

Watkins, Chris and Peter Dayan. 1992. *Q-learning*, Machine Learning **8**, 279–292.

Weibull, Jörgen W. 1997. *Evolutionary game theory*, MIT Press.

Werbos, Paul. 1974. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, Ph.D. Thesis.

_____. 1990. *Backpropagation through time: what it does and how to do it*, Proceedings of the IEEE **78**, 1550–1560.

Whitlock, Michael C. 2000. *Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection*, Evolution **54**, 1855–1861.

Willi, Yvonne, Patrick Griffin, and Josh Van Buskirk. 2012. *Drift load in populations of small size and low density*, Heredity **110**, 296–302.

Williams, David R. and Herbert Barry. 1966. *Counter conditioning in an operant conflict situation*, Journal of Comparative and Physiological Psychology **61**, 154.

Yachi, Shigeo and Richard M. Higashi. 1998. *The evolution of warning signals*, Nature **394**, 882–884.