Reinforcement Learning as a Model of Aposematism

Jan Teichmann[1], Eduardo Alonso[2] and Mark Broom[1]

[1]Department of Mathematics, City University London, London EC1V 0HB, UK
[2]Department of Computer Science, City University London, London EC1V 0HB, UK
E.Alonso@city.ac.uk

Abstract. 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: the prey is part of the predator's environment and vice versa. 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 and, typically, results in multiple species, which successively adapt in response to their adaptations. In particular, predator-prey co-evolution has been identified as an influential factor in the evolution of aposematism. In this paper we address the problem of formalizing predator aversive learning in the presence of aposematism by applying a reinforcement learning algorithm on a biologically plausible predator lifetime model.

Keywords: Predator-prey, Aposematism, Reinforcement Learning.

1 Introduction

The majority of species are at risk of predation in their natural habitat and are targeted by predators as part of the food web. Through the process of evolution by natural selection a predator is confronted with manifold mechanisms that have developed to avoid predation. So-called secondary defenses commonly involve the possession of toxins or deterrent substances which are not directly observable by predators. However, many defended species use conspicuous signals as warning flags in combination with their secondary defenses –what we call aposematism. There is a wide body of theory addressing the emergence and evolution of aposematism (Broom et al., 2006; Broom et al, 2008; Lee et al., 2010; Lee et al., 2011; Leimar et al., 1986; Marples et al., 2005; Ruxton et al., 2004; Ruxton et al., 2009; Yachi and Higashi, 1998). Within this context, the role of the predator as the selective agent and the mechanisms of the predator's aversive learning process are at the heart of current research (Barnett et al., 2007; Hagen et al., 2009; Sherrat, 2003). Nevertheless, there is no accepted formal model of aversive learning in foraging literature.

A normative modern framework for aversive learning can be found within the field of reinforcement learning, which provides a mapping of environmental states to an individual's action in order to maximize a reward signal in an unsupervised manner (Barto et al., 1990; Watkins, 1989; Sutton and Barto, 1998). Solving the underlying

reinforcement learning problem is crucial since natural environments are too complex to learn from examples of desired behaviour –such examples will not be 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 generalized experience from interactions with its environment. This is where a classical trade-off arises between exploration and exploitation: to maximize a reward an individual should perform the actions that it knows to be rewarding from previous experience. However, to find such actions in the first place an individual had to explore actions with unknown outcome. Therefore, the mapping of states to actions has to be obtained through trial-and-error, or goal directed learning where actions have subsequent effects on future rewards. The last decade has seen a proliferation of research on the neural and psychological mechanisms of reinforcement learning. We know from studies of neural correlates in behaving animals that reinforcement signals in the brain represent reward prediction error rather than a direct reward-reinforcement relation (Daw and Doya, 2006; Dayan and Balleine, 2002; Dayan and Daw, 2008; Dayan and Niv, 2008; Doya, 2007; Maia, 2009; Montague et al., 1996; Montague et al., 2004; Rangel et al., 2008; Schultz, 2002, 2007, 2008; Schultz et al., 1997). Temporal difference learning is a reinforcement learning methodology that reflects these insights by representing states and actions in terms of predictions about future rewards where the learning objective is to iteratively update the target values of future rewards towards their true values based on experience from interactions with the environment. However, apart from in Teichmann et al. (2014a) reinforcement learning has not been proposed to formalize aposematism within the foraging problem.

In this paper, we introduce a predator lifetime model where an individual's payoff is both dependent on the environment and additional aspects of an individual's behaviour, metabolism, and lifetime traits, which are usually abstracted away in reinforcement learning formalisms. This approach will allow us to investigate the cost of learning and the interactions of behaviour and metabolism on the learning outcome. In our approach the learning problem is to find an optimal foraging strategy under the aspects of maximizing the predator's payoff in an episodal task such as a day of foraging. Importantly, the predator's behaviour has delayed effects on its rewards so that a trajectory is not only dependent on its initial conditions but also on all the predator's actions and the subsequent state transitions. Therefore, we have to choose an episodal learning algorithm, which considers the entirety of actions and state transitions of a trajectory within its learning updates. We have used back-propagation through time (BPTT) as an efficient method to calculate the derivative of the predator's payoff function in the policy space for episodal tasks. The rest of the paper is structured as follows: in the next section we describe how reinforcement learning operates. The predator life model is introduced in detail in Section 3. Section 4 describes the reinforcement learning algorithm used to calculate optimal trajectories, and the derivatives used are formulated in Section 5. The results of the simulations are presented in Section 6. We shall conclude with a discussion of the results and further work.

2 Reinforcement Learning

A typical reinforcement learning scenario is an animal inhabiting a state space $S \subseteq \mathbb{R}^n$, such that at k iteration it "lives" in state $s_k \in S$. The state space represents any features the modeller considers relevant, typically a collection of stimuli, but can also include internal constructs. At each iteration, the animal chooses an action u_k (from an action space $u_k \in U$), which takes it to the next state according to a *model function*

$$s_{k+1} = f(s_k, u_k) \tag{1}$$

and gives it an immediate scalar r_k, represented by the *reward function*

$$r_{k+1} = r(s_k, u_k) \tag{2}$$

The animal keeps acting, forming a trajectory of states $(s_0, s_1, ...)$ indefinitely or until a given terminal state is reached. In this problem the animal must learn to choose actions that maximize the expectation of the total long-term reward, the return, received from any given start state s_0. Formally, the problem is to find an *policy* $p(s,z)$, where z is the parameter of a function approximator (typically, a neural network), which calculates actions

$$u_k = p(s_k, z) \tag{3}$$

such that the following *value function* is maximized

$$V(s_0, z) = \sum_k \gamma^k r_k \tag{4}$$

subject to Equations (1), (2) and (3), where $\gamma \in [0,1]$ is a constant discount factor that specifies the relative importance of long-term costs over short-term ones.

3 The 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. In this model an individual predator is characterized by its state s_k at iteration k. The state is given by

$$s_k = \{T, A, X, Y\} \tag{5}$$

with T being the time of an iteration k within an episode, A the age of the predator, and X and Y the axes of the spatial location of the predator within its environment at iteration k. 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 well-understood Gaussian distribution function

$$g_i(X,Y) = p_i \exp\left(-\left(\frac{(X - x_{i,0})^2}{2\sigma_{i,x}^2} + \frac{(Y - y_{i,0})^2}{2\sigma_{i,y}^2} \right) \right) \tag{6}$$

with $(x_{i,0}, y_{i,0})$ being the center of the prey population with density p_i and $(\sigma_{i,x}, \sigma_{i,y})$ the spread of the prey. The model assumes that the prey is aposematic with both models (venomous animals) and potential mimics (non-venomous animals which mimic the defenses of venomous animals). 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, given by

$$u_k = \{e_x, e_y\} \tag{7}$$

with e_x, e_y representing 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 and is the result of the predator's interaction with the environment. Thereby, the predator's actions have subsequent effects on its environment through locomotion and the predator's spatial location within the environment according to the reinforcement learning model. 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

$$r_{k+1} = \dot{V} = \underbrace{\lambda(A_k) R(s_k) - t_0 \dot{T}}_{\text{state dependent}} \overbrace{- |E(u_k)|}^{\text{action dependent}} \tag{8}$$

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

$$R(s_k) = \sum_i g_i(s_k) d(t_i)(r - t_i^2) \tag{9}$$

where

$$d(t) = \frac{1}{1 + d_0 t} \tag{10}$$

is 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, for example, by

$$I(A) = \frac{1}{1 + A} \tag{11}$$

The environment of this model is Markovian defined by the state transition function

$$f(s,u)_{k \to k+1} = \begin{pmatrix} \dot{T} = 1 + \sum_i g_i(s_k)\left(d(t_i)\left(t_h + t_t t_i^2\right) + t_s\right) \\ \dot{A} = (1/\lambda_0)\dot{T} \\ \dot{X} = \tanh(c_0 e_x) \\ \dot{Y} = \tanh(c_0 e_y) \end{pmatrix} \tag{12}$$

The transition of time $(\dot{T})$ between iterations occurs in unit time steps reflecting abasal 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 a unit step of one. The functions of the model are governed by single parameters, which allow the trade off between the different aspects of the predator's behaviour, lifetime traits, and environment $(x_0, y_0, t_0, d_0, I_0, c_0)$. These parameters define the lifetime model. The only term missing is the subjective payoff. The assumption 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 for prey type i such as to reproduce the observed foraging behaviour of the predator. Additionally to the lifetime model, we define a final instantaneous cost Υ 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 $(X = 0, Y = 0)$:

$$\Upsilon(s)_l = \begin{cases} -r_l \sqrt{X^2 + Y^2} & \text{if} \quad \sqrt{X^2 + Y^2} > e \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

with $-r_l$ being a punishment for not returning to the den at the end of an episode.

Within a biological context such a final cost can be thought of as a 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.

4 Reinforcement Learning using BPTT

As discussed previously, the control problem in reinforcement learning is about finding an optimal behaviour policy (aka an actor). Reformulating reinforcement learning using an actor becomes the problem of finding the parametrization $\vec{z}$ of the actor $\pi(\vec{s},\vec{z})$ that maximizes the total value –or minimizes the overall prediction 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 optimization problem for episodal tasks (Werbos, 1974). BPTT uses the actor $\pi(\vec{s},\vec{z})$, traditionally implemented as an artificial neural network with weights $\vec{z}$ as a universal function approximator and which is equivalent to the behaviour policy. As such, BPTT is an off-line learning algorithm that issues a weight update $\Delta\vec{z}$ at the end of an episode. The delayed effects of actions in the reinforcement learning 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. 14). This property gives the methodology its name. It is well-known that 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 concepts of discounting and bootstrapping. Therefore, 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 initialized: $(\partial V / \partial \vec{z})_l \leftarrow \vec{0}$ and $(\partial V / \partial \vec{s})_l \leftarrow (\partial \Psi / \partial \vec{s})_l$, with Υ (Eq. 13) being the final instantaneous cost of the terminal state s_l, and l being the final iteration in an episode of finite length.

Following the initialization 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 $(\partial V / \partial \vec{z})_k$ as follows:

$$\left(\frac{\partial V}{\partial \vec{z}}\right)_k \leftarrow \left(\frac{\partial V}{\partial \vec{z}}\right)_{k+1} + \gamma^k \underbrace{\left(\frac{\partial \pi(\vec{s},\vec{z})}{\partial \vec{z}}\right)_k}_{actor} \underbrace{\left(\left(\frac{\partial r}{\partial \vec{u}}\right)_k + \gamma\left(\frac{\partial f}{\partial \vec{u}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}\right)}_{behavioral\ gradient} \tag{14}$$

with the following state dependent value contribution deriving from the Markovian properties of the environment

$$\left(\frac{\partial V}{\partial \vec{s}}\right)_k = \underbrace{\left(\left(\frac{\partial r}{\partial \vec{s}}\right)_k + \gamma\left(\frac{\partial f}{\partial \vec{s}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}\right)}_{environmental\ gradient} + \underbrace{\left(\frac{\partial \pi(\vec{s},\vec{z})}{\partial \vec{z}}\right)_k}_{actor} \underbrace{\left(\left(\frac{\partial r}{\partial \vec{u}}\right)_k + \gamma\left(\frac{\partial f}{\partial \vec{u}}\right)_k \left(\frac{\partial V}{\partial \vec{s}}\right)_{k+1}\right)}_{behavioral\ gradient} \tag{15}$$

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

$$\vec{z} \leftarrow \vec{z} + \alpha \frac{\partial V}{\partial \vec{z}} \tag{16}$$

with α representing a learning rate. Summarizing, the BPTT algorithm can be understood as propagating the policy gradient of the value function with respect to a future state $\left(\partial V / \partial \vec{s}\right)_{k+1}$ backwards in time through the actor, the state-transition function, and the payoff function to obtain the policy gradient of the value function $\left(\partial V / \partial \vec{s}\right)_k$ of the previous state. As BPTT utilizes 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 (Werbos, 1990).

5 Derivatives Used by the BPTT Algorithm

As BPTT is model-based it requires 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 the algorithm requires the state $\partial f(\vec{s}_k,\vec{u}_k)/\partial \vec{s}$ and action $\partial f(\vec{s}_k,\vec{u}_k)/\partial \vec{u}$ dependent derivatives of the state-transition function f, where

$$\frac{\partial f(\vec{s}_k,\vec{u}_k)}{\partial \vec{u}} = \left\{\frac{\partial \dot{T}}{\partial \vec{u}}, \frac{\partial \dot{A}}{\partial \vec{u}}, \frac{\partial \dot{X}}{\partial \vec{u}}, \frac{\partial \dot{Y}}{\partial \vec{u}}\right\} \tag{17}$$

and

$$\frac{\partial f(\vec{s}_k,\vec{u}_k)}{\partial \vec{s}} = \left\{\frac{\partial \dot{T}}{\partial \vec{s}}, \frac{\partial \dot{A}}{\partial \vec{s}}, \frac{\partial \dot{X}}{\partial \vec{s}}, \frac{\partial \dot{Y}}{\partial \vec{s}}\right\} \tag{18}$$

Furthermore, BPTT requires the state and action dependent derivatives, respectively, $\partial r_{k+1}(\vec{s}_k,\vec{u}_k)/\partial \vec{s}$ and $\partial r_{k+1}(\vec{s}_k,\vec{u}_k)/\partial \vec{u}$ of the payoff function r_{k+1}. In regards to the lifetime model, the derivatives of the state transition function f are

$$\frac{\partial \dot{T}}{\partial \vec{s}} = \sum_i p_i \frac{\partial g_i(X,Y)}{\partial \vec{s}}\left(t_s + d(t_i)\left(t_{h,i} + t_i t_i^2\right)\right) = \tag{19}$$

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

where 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 present prey. Otherwise time progresses constantly and independently of age and time. The state dependent derivative of the predator's ageing follows directly from the time transition. 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. 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, i.e.

$$\frac{\partial \dot{T}}{\partial \vec{u}} = \frac{\partial \dot{A}}{\partial \vec{u}} = \frac{\partial \dot{X}}{\partial \vec{s}} = \frac{\partial \dot{Y}}{\partial \vec{s}} = \vec{0} \tag{20}$$

with the spatial location of the predator being solely affected by the predator's action. Additionally, time and age progress independently from the predator's investment into locomotion within each iteration. Next, consider the state and action dependent derivatives related to the value function V, which are given by the sum of discounted payoffs along the trajectory of an episode. The derivatives of the incremental changes to the value of an episode along a trajectory are

$$\frac{\partial \dot{V}}{\partial \vec{s}} = \begin{cases} \frac{\partial \dot{V}}{\partial T} = 0 \\ \frac{\partial \dot{V}}{\partial A} = \dot{\lambda} R(s_k) \\ \frac{\partial \dot{V}}{\partial X} = \lambda(A) \frac{\partial R(s_k)}{\partial X} - \left(t_0 \frac{\partial \dot{T}}{\partial X} \right) \\ \frac{\partial \dot{V}}{\partial Y} = \lambda(A) \frac{\partial R(s_k)}{\partial Y} - \left(t_0 \frac{\partial \dot{T}}{\partial Y} \right) \end{cases} \tag{21}$$

where $\partial R(\vec{s}) / \partial \vec{s}$ is the derivative of the state dependent payoff from interacting with prey given by

$$\frac{\partial R(\vec{s})}{\partial \vec{s}} = \sum_i p_i \frac{\partial g_i(X,Y)}{\partial \vec{s}} d(t_i)\left(r_i - t_i^2\right) = \tag{22}$$

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

and

$$\dot{\lambda}(A) = -\frac{1}{(A+1)^2} \tag{23}$$

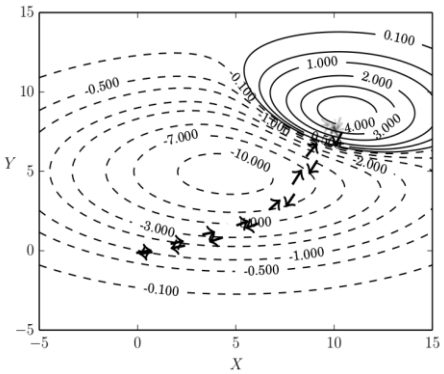
The absolute amount of energy invested into locomotion each iteration affects the value of an episode as follows:

$$\frac{\partial \dot{V}}{\partial \vec{u}} = \begin{cases} \frac{\partial \dot{V}}{\partial e_x} = -\operatorname{sgn}(e_x) \\ \frac{\partial \dot{V}}{\partial e_y} = -\operatorname{sgn}(e_y) \end{cases} \tag{24}$$

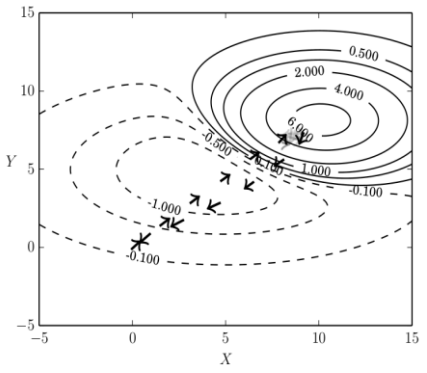
6 Results

The results in Fig. 1 and Fig. 2 depict the trajectories which are close to an optimal trajectory and which were found running the learning algorithm continuously saving trajectories that 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 utilize experience from ingesting prey individuals to find a rewarding feeding ground. The trajectory of a predator which is not utilizing taste-sampling (Fig. 1a) shows avoidance of the aversive prey population taking a non-direct route to the population of mimics. The pre-condition of exploration for successful aversion formation 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 $l = 80$ in the 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 (Fig. 1b).

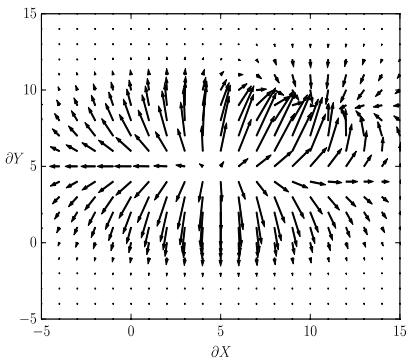
Fig. 2 shows the locomotion profile for the non-taste sampling predator. The predator's locomotion in general is optimized towards efficiency maximizing the displacement per energy expenditure $\max_{e_x} d\dot{X} / de_x$ and $\max_{e_y} d\dot{Y} / de_y$, which is at $e_x = e_y = 0.5$ in the simulation with a diagonal locomotion of length 0.7 being therefore 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. Additionally, the predator is over-staying in the feeding grounds with the second half of the trajectory showing a more rapid locomotion than the first half.



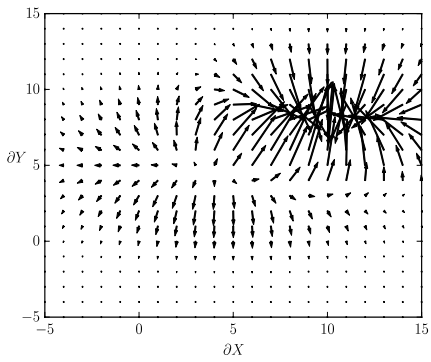
(a) The state dependent reward payoff $R(s_k)$ for a predator not utilizing taste sampling: $d_0 = 0$ and $t_s = 0$.



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

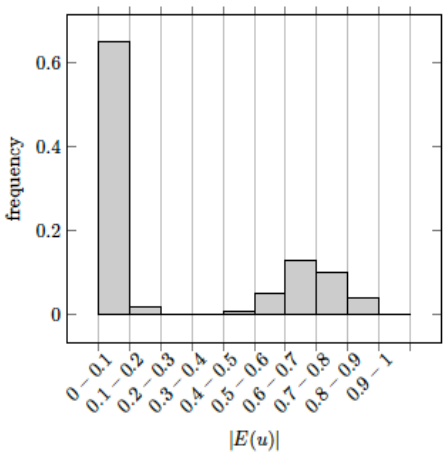


(c) The partial derivative of the reward with respect to the spatial position of the predator not utilizing taste sampling: $d_0 = 0$ and $t_s = 0$.

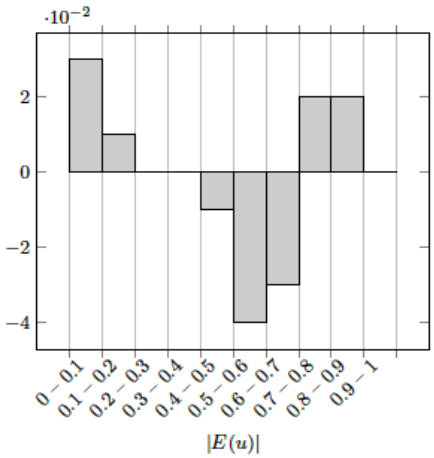


(d) The partial derivative of the reward with respect to the spatial position of the predator utilizing taste sampling: $d_0 = 1$ and $t_s = 0.1$.

Fig. 1. The state dependent reward of an exemplary environment with aposematic prey and Batesian mimics. All with $t_h = 0.1$, $t_t = 0.2$, $p_1 = p_2 = 0.5$, $x_{1,0} = y_{1,0} = 5$, $\mathcal{S}_{1,x} = 5$, $\mathcal{S}_{1,y} = 2.55$, $t_1 = 5$, $r_1 = 1$, $x_{2,0} = 10$, $y_{2,0} = 8$, $\mathcal{S}_{2,x} = 2$, $\mathcal{S}_{2,y} = 2$, $t_2 = 0$, $r_2 = 15$, $l = 80$.



(a) The predator's locomotion optimizes the energy expenditure with $\max_e d\dot{X}/de = 0.5$ and $\sqrt{\dot{X}(0.5)^2 + \dot{Y}(0.5)^2} = 0.7$.



(b) This plot shows the difference of the second half compared to the first half of an episode. The predator prefers to feed longer and return to its den using a greater step size than the optimal of 0.7.

Fig. 2. The locomotion profile of a predator not utilizing taste sampling with an episode of length $l = 80$.

These results show some interesting properties: (a) In a biological context the trajectories of the predator are unstable; (b) the element of the model which is generally optimized is the efficiency of locomotion (the behavioural expenditure); (c) a non-taste sampling predator avoids the aposematic prey population in order to minimize its metabolic costs from toxin ingestion; and (d) The predator shows a tendency to over-stay in the feeding grounds and returns to the den with above optimal energy expenditure for locomotion.

7 Discussion

In this paper we have presented a predator lifetime model including life history traits that have been traditionally abstracted away in the literature such as metabolic costs, locomotion, prey handling, and toxin recovery. The 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.

We applied a reinforcement learning algorithm trained using back-propagation through time (BPTT), to simulate behaviour of single individuals driven by rewards. BPTT address learning in episodal tasks based on experience including discounted future rewards. We used an artificial neural network as a universal function approximator in order to implement the policy. The learning task for the simulator is defined in a way to address the discussion of when behaviour is optimal (Teichmann et al., 2014b). On the one hand, the environment in the simulation contains rewards and punishment and optimal behaviour should maximize positive reinforcement. On the other hand, the environment contains a fitness related element in the form of an instantaneous final cost in case the predator does not return to its den at the end of an episode. From a biological context this cost 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 instability due to the interference of maximizing positive reinforcement along the trajectory (excluding the fitness cost) and maximizing the value of a complete trajectory (including the fitness cost). The simulator oscillates between two states: (i) a state of maximizing 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 maximizing the value of the complete trajectory including the final cost where the predator successfully returns to its den. However, the simulation shows that the predator generally optimizes the efficiency of its behavioural expenditure. That 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 corresponds in fact with biological findings (Nonacs, 2001). Summarizing, as far as we know the paper presents the first attempt to formalize foraging behaviour as a learning problem. In order to do this we have extended the traditional reinforcement learning framework with relevant ethological features, making it biologically plausible (à la Berridge 1996, 2003; Berridge et al., 2009). Hence our model is a contribution to foraging theory as

well as to reinforcement learning research. Of course, our results are preliminary, and future work will include investigating the effects of alternative terminal cost functions and refining the lifetime model. For instance, the model could be further developed by integrating a Darwinian fitness function. Finally, we will also consider whether we can eliminate, at least partially, fluctuations around optimal trajectories by using variations of the learning algorithm as such.

References

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

Barto, A.G., Sutton, R.S., Watkins, C.J.C.H.: Learning and sequential decision making. In: Gabriel, M., Moore, J.W. (eds.) Learning and Computational Neuroscience: Foundations of Adaptive Networks, pp. 539-602, MIT Press, Cambridge, Mass (1990).

Berridge, K.C.: Food reward: brain substrates of wanting and liking. Neuroscience & Biobehavioral Reviews 20(1), 1–25 (1996).

Berridge, K.C.: Pleasures of the brain. Brain and Cognition 52(1), 106–128 (2003).

Berridge, K.C., Robinson, T.E., Aldridge, J.W.: Dissecting components of reward: 'liking', 'wanting', and learning. Current Opinion in Pharmacology 9(1), 65–73 (2009).

Broom, M., Ruxton, G.D., Speed, M.P.: Evolutionarily stable investment in anti-predatory defences and aposematic signaling. In: Deutsch, A., Bravo de la Parra, R., de Boer, R.J., Diekmann, O., Jagers, P., Kisdi, E., Kretzschmar, M., Lansky, P., Metz, H. (eds.) Mathematical Modeling of Biological Systems, Volume II: Epidemiology, Evolution and Ecology, Immunology, Neural Systems and the Brain, and Innovative Mathematical Methods, pp. 37–48, Springer, Berlin (2008).

Broom, M., Speed, M.P., Ruxton, G.D.: Evolutionarily stable defence and signalling of that defence. Journal of Theoretical Biology 242, 32–34 (2006).

Daw, N.D., Doya, K.: The computational neurobiology of learning and reward. Current Opinion in Neurobiology 16, 199–204 (2006).

Dayan, P., Balleine, B.W.: Reward, motivation, and reinforcement learning. Neuron 36(2), 285–298 (2002).

Dayan, P., Daw, N.D.: Decision theory, reinforcement learning, and the brain. Cognitive, Affective & Behavioral Neuroscience 8, 429–453 (2008).

Dayan, P., Niv, Y.: Reinforcement learning: the good, the bad and the ugly. Current Opinion in Neurobiology 18, 185–196 (2008).

Doya, K.: Reinforcement learning: Computational theory and biological mechanisms. Human Frontiers Science Program 1, 30–40 (2007).

Guilford, T.: "Go-slow" signalling and the problem of automimicry. Journal of theoretical biology 170, 311–316 (1994).

Hagen, E.H., Sullivan, R.J., Schmidt, R., Morris, G., Kempter, R., Hammerstein, P.: Ecology and neurobiology of toxin avoidance and the paradox of drug reward. Neuroscience 160, 69 –84 (2009).

Lee, T.J., Marples, N.M., Speed, M.P.: Can dietary conservatism explain the primary evolution of aposematism?. Animal Behaviour 79, 63 –74 (2010).

Lee, T.J., Speed, M.P., Stephens, P.A.: Honest signaling and the uses of prey coloration. The American Naturalist 178(1), E1–E9 (2011).

Leimar, O., Enquist, M., Sillen-Tullberg, B.: Evolutionary stability of aposematic coloration

and prey unprofitability: A theoretical analysis. The American Naturalist 128(4), 469–490 (1986).
Maia, T.V.: Reinforcement learning, conditioning, and the brain: Successes and challenges. Cognitive, Affective, & Behavioral Neuroscience 9, 343–364 (2009).
Marples, N.M., Kelly, D.J., Thomas, R.J.: Perspective: The evolution of warning coloration is not paradoxical. Evolution 59, 933–940 (2005).
Montague, P.R., Dayan, P., Sejnowski, T.J.: A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of neuroscience 16, 1936–1947 (1996).
Montague, P.R., Hyman, S.E., Cohen, J.D.: Computational roles for dopamine in behavioural control. Nature 431, 760–767 (2004).
Nonacs, P.: State dependent behavior and the marginal value theorem. Behavioral Ecology 12(1), 71–83 (2001).
Rangel, A., Camerer, C., Montague, P.R.: A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience 9, 545–556 (2008).
Ruxton, G.D., Sherratt, T.N., M.P. Speed, M.P.: Avoiding attack: The evolutionary ecology of crypsis, warning signals and mimicry. Oxford University Press, Oxford (2004).
Ruxton, G.D., Speed, M.P., Broom, M.: Identifying the ecological conditions that select for intermediate levels of aposematic signaling. Evolutionary Ecology 23, 491–501 (2009).
Schultz, W.: Getting formal with dopamine and reward. Neuron 36, 241–263 (2002).
Schultz, W.: Multiple dopamine functions at different time courses. Annual Review of Neuroscience 30, 259–288 (2007).
Schultz, W.: Neuroeconomics: the promise and the profit. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 3767–3769 (2008).
Schultz, W., Dayan, P., Montague P.R.: A neural substrate of prediction and reward. Science 275, 1593–1599 (1997).
Sherratt, T.N.: State-dependent risk-taking by predators in systems with defended prey. Oikos 103, 93–100 (2003).
Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. Cambridge University Press, Boston (1998).
Teichmann, J., Broom, M., Alonso, E.: The application of temporal difference learning in optimal diet models. Journal of Theoretical Biology 340, 11–16 (2014a).
Teichmann, J., Broom, M., Alonso, E.: The evolutionary dynamic of aposematism: a numerical analysis of co-evolution in finite populations. Mathematical Modelling of Natural Phenomena 9, 148-164 (2014b).
Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. Thesis. King's College, Cambridge University, London (1989).
Werbos, P.J.: Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. Thesis. Harvard University, Cambridge, Mass (1974).
Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE 78(10) 1550–1560 (1990).
Yachi, S., Higashi, M.: The evolution of warning signals. Nature 394, 882–884 (1998).