Foraging choices modelled using Reinforcement Learning

Introduction

We make decisions based on reward and punishment system. When we get a reward, we have a positive feedback cycle to take the same kind of decision again and not to take the decision when we get a punishment (or no reward). We learn overtime using this feedback loop and this is called experience-based learning. There are 2 ways to take decisions, either to 'explore' or to 'exploit' whatever we know from previous experience. If we explore, we might get higher rewards but there's no guarantee, if we exploit, we get same rewards as before but there's guaranteed. The goal is to strike a balance between the two. So, what is the mechanism of this decision making? This is rather a broad question to ask so narrowing down to make decisions for one specific task(example) will give us a broader idea.

Foraging is a good phenomenon to study decision making and experience-learning. Predator needs to make decisions on the choice of prey, energy spent, time taken to travel to get the prey (parameters). Optimal foraging theory answers this by taking into account the fitness of the species overtime. Better model to understand this would be how our brains generally make decisions. This can be modelled by Reinforcement Learning.

To narrow down further, it's better to investigate one parameter at a time. In this report, we investigated how choice of prey affects foraging behavior due to learning. We referred this paper to investigate this parameter [1] and are trying to use their model to get some preliminary understanding.

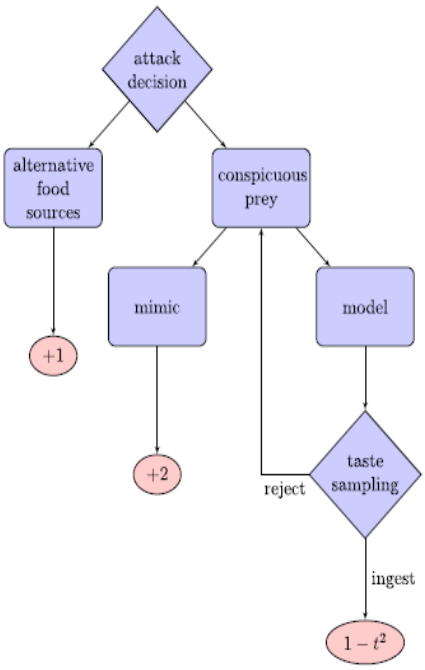
Hypothesis construction

We have one type of predator and 3 types of prey – toxic, mimics-toxic and non-toxic. Toxic prey reduces the rewards, non- toxic prey has more reward and mimics-toxic prey has more reward associated. Reward is arbitrary here. Predator starts out with no idea of the 3 types and learns overtime toxicity aversion by making decisions guided by the reward scheme described. Our hypothesis is to investigate whether ability to learn toxic aversiveness can be modelled using Reinforcement learning, thus the predator learns to hunt with better knowledge.

Model – Q - learning

Q learning is a model-free method which helps to learn from experienced rewards without the necessity of building representations of the environment. In Q learning, an 'episode' is one trial of the experiment we conduct, here from the start till end of when the predator eats. Our algorithm [2], implements many episodes until the learning is done (i.ie until a convergence condition). Each episode is described by a state (state of the system at that timepoint), action (which action is performed at that timepoint) pair under a given 'policy' of how to perform the action. We have a 'Q-table' (a state x action sized matrix) which stores all q values. Q values are representative of a scalar which represents the overall reward over the course of time starting from a given timepoint. 'Q-values' are updated with the reward at that point added with 'Temporal Difference (TD)'. TD is the difference between the target reward and the actual reward at that timepoint. We need to minimize TD to maximize learning.

Q-learning is an off-policy algorithm, which means that, while learning a so-called 'target policy', it uses a so-called '*behavior policy*' to select actions. In our model, it is 'SoftMax'. This policy selects random actions to 'explore' new predators with probabilities proportional to their current values.



$Q \leftarrow 0$
$s\_k \leftarrow s\_0$
WHILE learning DO
    $a\_k \leftarrow \pi(s\_k, Q)$
    $s\_(k+1) \leftarrow f(s\_k,\ a\_k)$
    $Q(s\_k, a\_k) \leftarrow Q(s\_k,\ a\_k) + \alpha\ (r\_(k+1) +$
        $\gamma\ \max\_a\ Q(s\_(k+1), a)\ - Q(s\_k, a\_k)\ )$
    $s\_k \leftarrow s\_(k+1)$

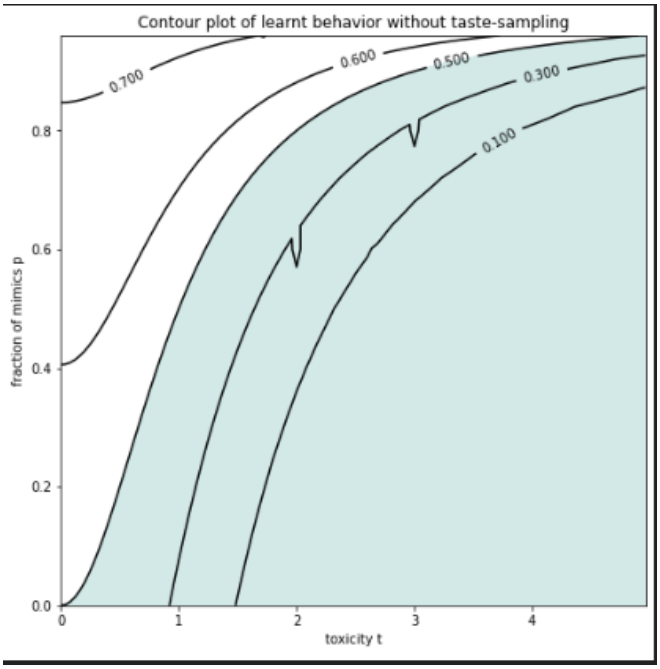
Fig. A1. Q-learning algorithm in pseudo-code.

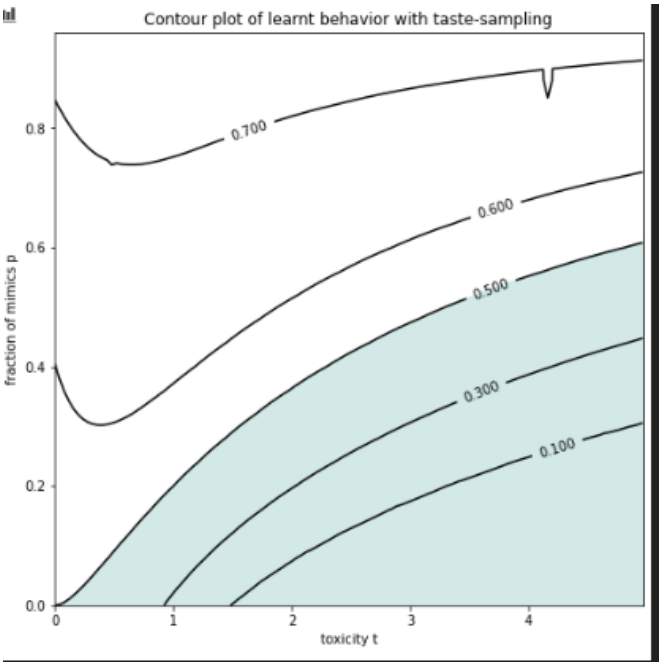
Model overview [1]

Implementation of the model can be found on <github>[2] and paper describing it can be found here[1].

Discussion

Our model enables an unbiased predator to learn toxicity aversiveness. Given a toxicity and graph our model produces gives a probability with which the predator attacks. From the plots we can see that during high toxicity and a smaller number of mimics, the predator is more toxic aversive. Also, during high number of mimics and a smaller number of mimics, it is less toxic aversive. This is our sanity check to ensure that the model behaves like we expected it to.





With taste sampling means the predator ‘explores’ more than ‘exploits. Without sampling means the predator ‘exploits’ more than explores. The shaded region shows toxic aversity.

If the predator explores, then it is more likely to choose aversive preys than if it exploits as shown in the plots.

We conclude that with only using toxicity to classify the type of prey we can use Reinforcement learning to model the foraging behavior/choice of the predator.

Further tasks

1. To explore parameter that influence foraging behavior.
2. To include versatility of prey
3. Use this model as a benchmark to see how real-life animals behave with respect to it.
4. Implement more realistic Reward scalars and qualitatively explain it.
5. What are the neural mechanisms that govern decision making of this type?

References

1. **The application of temporal difference learning in optimal diet models**
   https://www.sciencedirect.com/science/article/abs/pii/S0022519313004189
2. Implementation of the modelhttps://github.com/theLamentingGirl/ForagingByReinforcementLearning
3. Reinforcement learning: the good, the bad and the ugly.
   https://www.researchgate.net/publication/23175588_Reinforcement_learning_The_Good_The_Bad_and_The_Ugly
4. https://www.freecodecamp.org/news/an-introduction-to-q-learning-reinforcement-learning-14ac0b4493cc/