Cogmaster – Methods in Computational Neuroscience

Problem set #2 QUANTITATIVE MODELS OF BEHAVIOR

Deadline: Friday 10th of March, 2023 at 14:00h

PROBLEM 1 *Rescorla-Wagner Model*

Several classical conditioning experiments can be qualitatively reproduced by the Rescorla-Wagner model (see Chapter 9.1-9.2 of the Dayan & Abbott). To explain these experimental observations, we assume that the objective of an animal is to predict the presence of certain events such as a food reward. Let us denote the presence or absence of a such a reward (also called unconditioned stimulus or UCS) by $r = 1$ or $r = 0$, respectively. Other events such as stimuli (also called conditioned stimuli or CS) may or may not predict the occurence of this reward, and we denote their presence by $u = 1$ and their absence by $u = 0$. The organism's task is then to predict if a reward is present, depending on whether the stimulus was present. We denote the animal's prediction by v and write

$$v = wu \tag{1}$$

where w is a free parameter that the animal needs to learn. After every trial of a conditioning experiment, this parameter is "learned" or updated using the Rescorla-Wagner learning rule,

$$w \to w + \epsilon\delta u \tag{2}$$

where ϵ is the learning rate, $\delta = r - v$ is the prediction error, and u is the stimulus.

(a) Generate a set of stimuli u_i and rewards r_i. We use the index i to number the trials. Assume that during the first 25 trials, both stimuli and reward are present, and during the next 25 trials, only the stimulus is present. Plot the rewards and stimuli as a function of the trial index.

(b) Use the learning rule above to adapt w during the experiment. Choose the learning rate $\epsilon = 0.1$. Plot v_i as a function of the trial.

(c) Change the learning parameter ϵ. What happens?

(d) *Partial conditioning*. In each trial, the stimulus is present, but the presence of the reward is a random event with probability 0.4. Plot the presence/absence of the reward as a function of the trial index. Redo the simulation from (b). What happens?

(e) *Blocking*. Assume that there are two stimuli, u_1 and u_2, and two parameters, w_1 and w_2 to learn, and that the animal's prediction is given by $v = w_1u_1 + w_2u_2$. During the first 25 trials, only one stimulus and the reward are present, during the next 25 trials, both stimuli and the reward are present. What happens?

(f) *Overshadowing* Assume again that there are two stimuli and two parameters to learn. However, now both stimuli are present from the beginning, as is the reward. What happens if one of the learning rates is larger (e.g. $\epsilon = 0.2$ for one stimulus vs. $\epsilon = 0.1$ for the other)?

Background Literature: Dayan & Abbott, Chapter 9.1-9.2

PROBLEM 2 *Simple decision strategy for flower sampling by bees.*

We assume a bee is collecting nectar from yellow and blue flowers. During one day the bee can sample 100 flowers. Let us observe its behavior during two days. During the first day, blue flowers carry a reward of $r_b = 8$, and yellow flowers a reward of $r_y = 2$. During the second day, rewards are reversed, so that $r_b = 2$ and $r_y = 8$.

We assume that in each trial the bee decides to land either on a blue or yellow flower. The bee bases its decision on estimates of the nectar reward that it receives on blue or yellow flowers. Let us denote the bee's internal estimate of these nectar rewards as m_b (blue flowers) and m_y (yellow flowers). Then the bee will choose a blue flower with probability

$$p_b = \frac{1}{1 + \exp\big(\beta(m_y - m_b)\big)} \tag{3}$$

where we call β the "exploitation-exploration trade-off" parameter.

(a) Where does the name "exploitation-exploration trade-off" come from? Plot p_b as a function β (with $\beta > 0$) and fixed difference $m_y - m_b$, then plot p_b as a function of $m_y - m_b$ and fixed β. I am not giving you any values for the parameters, because you are supposed to find interesting values yourself! (Note: the above decision-making strategy is sometimes called a "softmax"-strategy.)

(b) **Dumb bee.** Using this decision-making strategy, generate a sequence of choices that a bee will make during the two days described above. We assume that the bee is "dumb" and cannot learn from experience. During both days, the bee thinks that yellow flowers carry a reward $m_y = 5$ and blue flowers carry no reward, $m_b = 0$. Look at two different exploration-exploitation trade-offs: How does the bee behave if $\beta = 0$? How does it behave if $\beta = 0.8$?
Programming Help: *The difficulty here is to draw choices with probabilities determined by p_b. The trick is to resort to the function RAND (random.rand for numpy/scipy, rand for matlab). A number drawn with this function is uniformly distributed within the interval $[0, 1]$. Use this number and turn it into choice "blue" with probability p_b or choice "yellow" with probability $1 - p_b$.*

(c) **Smart bee.** Now let us assume that the bee is "smart" and that it can learn from its experiences. Each time it visits a blue flower, it will update the estimated reward according to the online update rule,

$$m_b \to m_b + \epsilon(r_b - m_b) \tag{4}$$

and similarly for a yellow flower,

$$m_y \to m_y + \epsilon(r_y - m_y) \tag{5}$$

Given a learning parameter $\epsilon = 0.2$ and the initial assumptions about flower reward from above ($m_y = 5$, $m_b = 0$), simulate the bees sequence of choices during the two days. How do the reward estimates change over time? Explore the case of purely explorative behavior ($\beta = 0$) and the case of strongly exploitative behavior ($\beta = 1$). What do you observe?

Background Literature: Dayan & Abbott, Chapter 9.3

PROBLEM 3 *The drift diffusion model of decision-making.*

In a two-alternative forced choice task (2AFC-task), subjects are asked to choose between two alternative actions. We will here consider the case where a subject receives a visual motion stimulus (a set of points on a screen that are moving in different directions) and then needs to indicate whether the points were moving upwards or downwards. If such a motion stimulus is ambiguous or "noisy", the task can be quite difficult. We will assume that the motion stimulus continues until the subject has made a choice. This scenario is well described by the "drift-diffusion-model", in which the subject compares the firing rate m_A of an upward-motion sensitive neuron with the firing rate m_B of a downward-motion sensitive neuron and integrates the difference between the two,

$$\dot{x} = m_A - m_B + \sigma\eta(t) \quad .$$

where $\eta(t)$ is a noise term (Gaussian white noise with unit standard deviation) that simulates the noisiness of real neurons. If the integration variable x surpasses a threshold μ, then the subject decides for outcome A; if x decreases below a threshold $-\mu$, then the subject decides for outcome B.

Any ordinary differential equation can be solved numerically using the Euler method, i.e., using the approximation

$$x(t + \Delta t) = x(t) + \dot{x}\Delta t \tag{6}$$

For stochastic differential equations, i.e., those that have a noise-term, the random part grows with the square root of the time step, a technical issue that you can ignore for now. It leads to the following discrete approximation of the drift-diffusion-model:

$$x(t + \Delta t) = x(t) + (m_A - m_B)\Delta t + \sigma\eta(t)\sqrt{\Delta t} \quad . \tag{7}$$

(a) Assume $m_A = 1$ and $m_B = 0.95$. Plot several runs of the drift-diffusion-model, always starting with $x(0) = 0$. Choose a stepwidth of $\Delta t = 0.1$ ms, a noise level $\sigma = 0.5\,\mathrm{ms}^{-\frac{1}{2}}$ and iterate the Euler method over 10000 time steps up to time $t = 1$ s. You should choose a reasonable value for the threshold μ.

(b) Run the model 1000 times and store the outcome (A vs. B) and the time of threshold crossing, t_i in each run. Assume that the reaction time of the subject is given by $\mathrm{RT}_i = 100ms + t_i$. Plot the distribution of reaction times for outcome A and the distribution for outcome B.

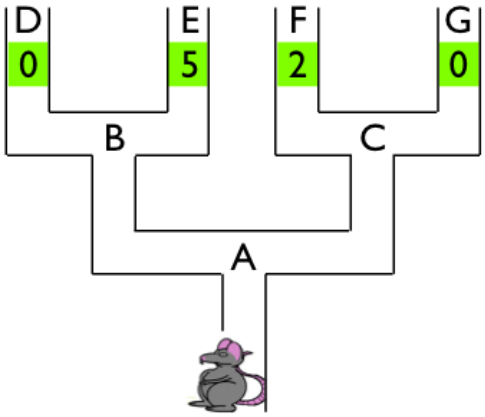
Programming Help: *Run the program with a smaller amount of trials first; once it's working, go to the number of trials asked for!*

(c) We will denote the evidence for outcome A versus outcome B as $m_E = m_A - m_B$. Plot the probability of outcome A for values of m_E ranging from -0.2 to 0.2. Compare the results of your simulation with the analytical formula, which states that the probability of choosing B is given by

$$p_b = \frac{1}{1 + \exp\big(\beta(m_A - m_B)\big)} \tag{8}$$

where $\beta = 2\mu/\sigma^2$.

Background Literature: Glimcher PW (2001). Making Choices: The Neurophysiology of Visual-Saccadic Decision Making. Trends in Neurocience. 24: 654-659.



PROBLEM 4 *Advanced and optional exercise: Reinforcement learning in a maze.*

Imagine a rat going through the maze shown below. The rat enters the maze at state A, then moves on to either state B or C, then to D, E, F, or G, where it can potentially collect a reward (as given by the numbers in the figure), and finally, the rat is taken out of the maze by the experimenter, and thereby moves into the "terminal" state H (not shown).

(a) Assume the rat follows a random decision-making strategy ("policy"), i.e., at each junction, it moves left or right with 50% probability. How often does the rat visit each state s (where $s = \{A, B, C, D, E, F, G, H\}$)? Give the theoretical number (what you expect given the policy), then perform a numerical simulation. Generate $N = 100$ trials of the rat's behavior and count how often it visits each state. (Note: each trial consists of a succession of four states: s_1, s_2, s_3, s_4, the first state is always $s_1 = A$ and the last state is always the terminal state $s_4 = H$. We assume that the rat never turns back and always moves forward.)

(b) In reinforcement learning theory, each state has a value, which is the expected sum of all possible future rewards. These values can be learned through experience using a method called "temporal difference learning". Initially, the rat assumes that no state carries any value ($V(s) = 0$ for all s). After each trial, the state values are updated according to the temporal difference learning rule:

$$V(s_t) \to V(s_t) + \epsilon\big[r(s_t) + V(s_{t+1}) - V(s_t)\big] \tag{9}$$

where s_t with $t = \{1, 2, 3, 4\}$ denotes the sequence of states in a trial, and $r(s)$ is the reward obtained in state s. Use the trial sequences generated in (a) to update the values $V(s)$. Plot the values $V(s)$ as a function of the trial number.

(c) A smart rat may want to use the information it is collecting about the values of the different states (i.e., the expected future reward!) to adapt its decision-making strategy. Use the estimated values $V(s)$ from above to change the rat's policy: assume that at each junction, the rat compares the values of the two choices using the "softmax"-decision rule used in bee-learning. What happens to the learning process if the rat is "greedy", i.e., usually goes for the side with the larger value, what if the rat is very explorative?

Background Literature: Dayan & Abbott, Chapter 9.4