M2 Biology **18/11/21**

TD7b: Learning III

Elie Oriol

TD material is available at:

`https://github.com/Elieoriol/2122_UlmM2_ThNeuro/tree/master/TD7b`

This last tutorial focuses on a model of reinforcement learning.

1 Model

We consider an agent in an environment. The environment can be in any state from a set $\{s\}$ and the agent can perform any action from a set $\{a\}$.

Actions change the state of the world: if at time t, the state of the world is s_t and the agent performs action a_t, then at time t+1, the state of the world is s_{t+1} where s_{t+1} is drawn from the probability distribution $p_{tr}(s_{t+1}|s_t, a_t)$.

Actions also allow the agent to obtain reward, depending on which state the world is in: every time the agent performs an action, it receives a reward $r_t = R(s_t, a_t)$.

1. We suppose that the agent can perceive which state s_t the world is in, but does not know the functions p_{tr} and R. How can it choose its actions?

2 Temporal Difference algorithm

We suppose that the agent has a policy which allows it to choose what action to perform given the state that it's in. This policy may be stochastic, in which case the action performed at time t is drawn from the distribution $\pi(a_t|s_t)$.

Given this policy, the subjective value of a state $V(s)$ is the expected total reward obtained by the agent supposing that it starts in this state and follows this policy.

2. Express $V(s_t)$ as a function of $V(s_{t+1})$.
3. Supposing that the agent starts with an estimate $\hat{V}$ of V, suggest an online update of $\hat{V}$ which can be implemented every time the agent performs an action.
4. Once the agent has found a good enough estimate of V, how should it modify its policy?

3 Maze learning

We consider a very simple maze: there is a single corridor, the agent starts in the middle and has to find the exit, which is at the right end of the maze. If it reaches the left end of the maze then it must stop.

5. How would you model this?
6. Supposing that the agent's policy is to go left or right with equal probability $1/2$ at each step, what is the value function?
7. Supposing that the agents initial estimate of V is a uniform function $\hat{V} = 1/2$, how does $\hat{V}$ evolve during the first episode? What is its value at the end of the episode? What is its value at the end of the second episode?