Homework #7: Reinforcement learning
(NEURL-GA 3042, Fall 2022)
Due date: Wednesday November 30

Problem 1. V-learning

Recall from our lectures that we can think of a standard "V-learning" reinforcement learning computation as occurring in 2 steps. First we compute for the current trial, t:

$$\delta = r_t - V_{t-1}$$

where r is the just received reward from the current trial and V_{t-1} is the stored estimate of value from the previous trial. Next we compute the value for the current trial as:

$$V_t = V_{t-1} + \alpha\delta$$

where α is the learning rate. For this homework we would like you to implement this learning equation in python and plot V_t for the following set of rewards at t:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
r	2	1	2	1	2	1	2	1	2	1	2	1	3	3	3	3	3	3	3	3

You will notice that for the first 12 trials the value of the trial in expectation, $\mathbb{E}[V]$, is 1.5. Can you identify the value of α that converges to within 10% of $\mathbb{E}[V]$ as quickly as possible. Next, answer the same question for convergence the $\mathbb{E}[V]$ as quickly as possible, but now for the underlying value that arises after trial 13. **What does this tell you about the relationship between the learning rate and environmental stochasticity?**

Problem 2: Q-learning

Now let us consider a world in which there are two "signals", say a light (s_l) and a sound (s_s). We can represent them as two state variables:

$$Q_t(s_l, s_s) = Q_{t-1}(s_l, s_s) + \alpha\delta$$

but where the delta function takes the (standard) form:

$$\delta = r_t - [Q_{t-1}(1,0) + Q_{t-1}(0,1)]$$

In other words, δ is the SUM of all predictions for each of the possible states. That is, if it's a light trial, you subtract for light, if it's a sound trial you subtract for sound, and if it's both you subtract the sum of what you have for both (as shown above). Ok, now implement the equations above (using indicator variables on the state-dependent Q-values to turn them on and off as necessary) for the following trial matrix with an α set to 0.5.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
s_l	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
s_s	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
r	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Plot both Q-values as a function of trial number. What does this tell us about δ ? Oddly, exactly this predicted behavior has been widely observed, first by Kamin in 1969 [1]

Finally, contrast this with the graphs of the same system facing the following two matrices:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
s_l	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
s_s	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
r	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2

and

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
s_l	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
s_s	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
r	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2

Discuss what these simulations tell you about the learning rule, and learning based on prediction errors.

Note: *In all problems, start with the initial values (at $t = 0$) at 0.*