Computational Neuroscience Homework 7

Zihan Zhang (Steven)

December 5, 2022

0.1 Question 1

In this question, we would implement the standard V-learning reinforcement learning computation. We compute the current trial t:

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

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

$$V_t = V_{t-1} + \alpha\delta. \tag{2}$$

where α is the learning rate. In all problems, we start with the initial values (at $t = 0$) at 0. Figure 1 describes that different sampled α would lead to various convergence speeds to the expectation value $\mathbb{E}[V] = 1.5$. The interval between two dashed lines is the $10\% \cdot \mathbb{E}[V]$ variation. Also, we could qualitatively deduce that the optimal α choice is approximately 0.5 with the shortest time. Similarly, we repeat the calculations for trails after 13, and Figure 2 shows the result. Here the optimal α choice is 1. Conclusively speaking, a larger (smaller) environmental stochasticity should be trained with a smaller (larger) learning rate that would lead to a quicker convergence speed. That is intuitively correct since we should slowly learn and conclude when the data is oscillating or changeable. When the data is stable (or constant, in some extreme cases), we could increase the learning rate, even taking the result as granted.

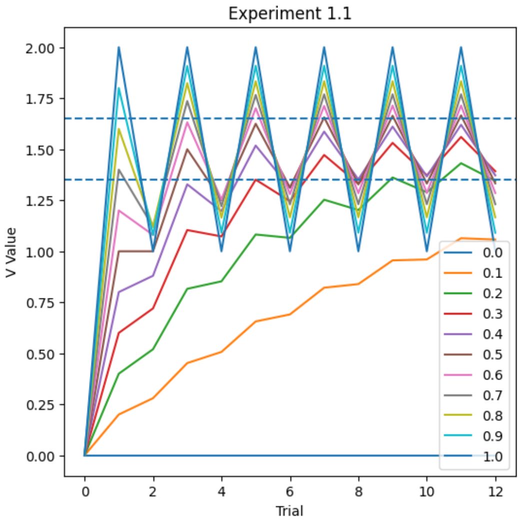

Figure 1: Learning process of V_t rewards at first 12 trials with different α.

0.2 Question 2

In this question, we would implement Q-learning with two signals, respectively the light s_l and sound s_s. We represent them as two state variables:

$$Q_t(s_l, s_s) = Q_{t-1}(s_l, s_s) + \alpha_\delta, \tag{3}$$

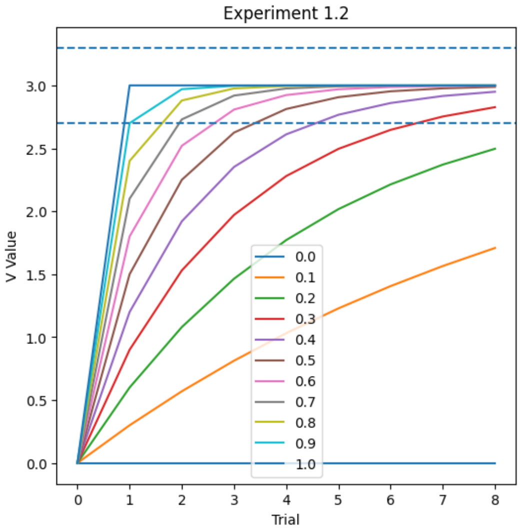

Figure 2: Learning process of V_t rewards after trail 13 with different α.

where the delta function takes the standard form:

$$\delta = r_t - [Q_{t-1}(1,0) + Q_{t-1}(0,1)]. \tag{4}$$

which is the sum of all predictions for each of the possible states. In other words, we would update Q_{sl} when $sl = 1$ (light trial), update Q_{ss} when $ss = 1$ (sound trial), and update both when $sl = ss = 1$ (both trails). Figure 3 describes the Q-values as a function of the trial number for experiment 2.1. Notice that δ gradually converges. We repeat the calculation in another two datasets, and Figure 4 and 5 show the results. Comparing the results, we could conclude that when the prediction error is large, the learning speed is faster; the speed gradually decreases when the environment is stable.

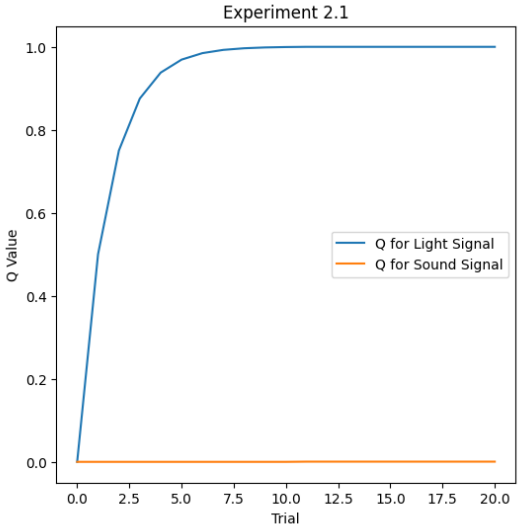

Figure 3: Learning process of Q_t rewards for experiment 2.1.

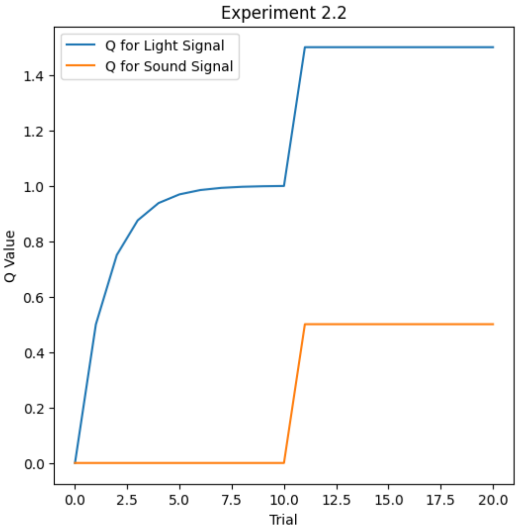

Figure 4: Learning process of Q_t rewards for experiment 2.2.

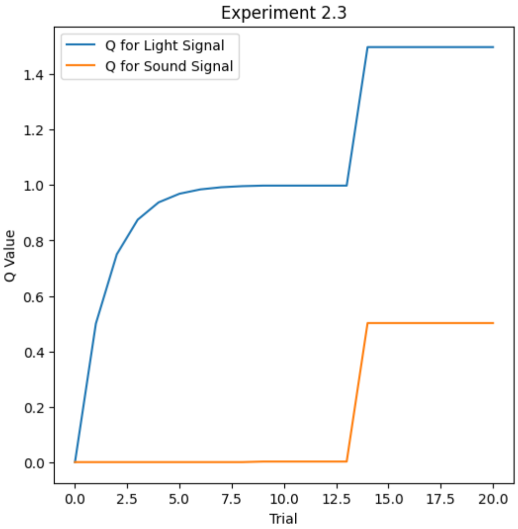

Figure 5: Learning process of Q_t rewards for experiment 2.3.