Exercises

Chapter 9

1. Implement acquisition and extinction as in figure 9.1 using the Rescorla-Wagner (delta) rule (equation 9.2).

2. Add a second stimulus and demonstrate that the delta rule can describe blocking, but that it fails to exhibit secondary conditioning.

3. Consider the case of partial reinforcement (studied in figure 9.1) in which reward $r = 1$ is provided randomly with probability p on any given trial. Assume that there is a single stimulus with $u = 1$, so that $\epsilon\delta u$, with $\delta = r - v = r - wu$, is equal to $\epsilon(r - w)$. By considering the expected value $\langle w + \epsilon(r - w)\rangle$ and the expected square value $\langle (w + \epsilon(r - w))^2 \rangle$ of the new weights, calculate the self-consistent equilibrium values of the mean and variance of the weight w. What happens to your expression for the variance if $\epsilon = 2$ or $\epsilon > 2$? To what features of the learning rule do these effects correspond?

4. The original application of temporal difference learning to conditioning (Sutton & Barto, 1990) considered the use of stimulus traces (as a preliminary to the linear filter of equation 9.5). That is, the prediction of sum future reward at time t is $v(t) = \mathbf{w} \cdot \mathbf{u}(t)$ where $u_i(t)$, with prediction weight is w_i, marks the presence (when $u_i(t) = 1$) or absence (when $u_i(t) = 0$) of stimulus i at time t. Also, the temporal difference learning rule of equation 9.10 is replaced by

$$w_i \to w_i + \epsilon\delta(t)\bar{u}_i(t)\,,$$

where

$$\bar{u}_i(t) = \lambda\bar{u}_i(t-1) + (1-\lambda)u_i(t)$$

is the stimulus trace for stimulus i, and $\delta(t)$ is as in equation 9.10. Here λ is the trace parameter which governs the length of the memory of the past occurrence of stimuli (see equation 9.30). Construct a trace learning model for a case similar to that of figure 9.2, but taking $r(t)$ to be the hat-function $r(t) = 1/5,\ 200 \le t \le 210$ and $r(t) = 0$ otherwise. Note that to match figure 9.2, you must use $\Delta t = 5$ for each time step rather than $\Delta t = 1$. Show the signals as in figure 9.2B for $\lambda = 0.5, 0.9, 0.99$, using $\epsilon = 0.2$. Could this model account for the data on the activity of the dopamine cells? Would it show secondary conditioning?

5. Use the prediction model of equation 9.5 and the standard temporal difference learning rule of equation 9.10 to reproduce figure 9.2. Take $r(t)$ to be the hat-function $r(t) = 1/5,\ 200 \le t \le 210$ and $r(t) = 0$ otherwise. In this figure, the increments of time are in steps of $\Delta t = 5$, and $\epsilon = 0.4$. Consider what happens if the time between the

stimulus and the reward is stochastic, drawn from a uniform distribution between 50 and 150. Show the average prediction error signal $\delta(t)$ time-locked to the stimulus and the reward. How does this differ from those in figure 9.2.

6. Implement a stochastic three-armed bandit using the indirect actor and the action choice softmax rule 9.12. Let arm a produce a reward of p_a, with $p_1 = 1/4$, $p_2 = 1/2$, $p_3 = 3/4$, and use a learning rate of $\epsilon = 0.01, 0.1, 0.5$ and $\beta = 1, 10, 100$. Consider what happens if after every 250 trials, the arms swap their reward probabilities at random. Averaging over a long run, explore to see which values of ϵ and β lead to the greatest cumulative reward. Can you account for this behavior?

7. Repeat exercise 6 using the direct actor (with learning rule 9.22). For $\bar{r}$, use a low-pass filtered version of the actual reward, which is obtained by using the update rule

$$\bar{r} \to \lambda\bar{r} + (1 - \lambda)r$$

with $\lambda = 0.95$. Study the effect of the different values of ϵ and β in controlling the average rate of rewards when the arms swap their reward probabilities at random every 250 trials.

8. Implement actor critic learning (equations 9.24 and 9.25) in the maze of figure 9.7, with learning rate $\epsilon = 0.5$ for both actor and critic, and $\beta = 1$ for the critic. Starting from zero weights for both the actor and critic, plot learning curves as in figures 9.8 and 9.9. Start instead from a policy in which the agent is biased to go left at both B and C, with initial probability 0.99. How does this affect learning at A?

9. Implement actor critic learning for the maze, as in exercise 8, except using vectorial state representations as in equations 9.26, 9.27, and 9.28. If $\mathbf{u}(\mathrm{A}) = (1, 0, 0)$, $\mathbf{u}(\mathrm{B}) = (0, 1, 0)$ and $\mathbf{u}(\mathrm{C}) = (0, 0, 1)$, then the result should be exactly as in exercise 8. What happens to the speed of leaning if $\mathbf{u}(\mathrm{A}) = (1, a, a)$ (while retaining $\mathbf{u}(\mathrm{B}) = (0, 1, 0)$ and $\mathbf{u}(\mathrm{C}) = (0, 0, 1)$) for $a = +0.5$ and $a = -0.5$, and why?