Exercise 7

[Avik Banerjee (3374885), Soumyadeep Bhattacharjee (3375428)]

Text in italics are notes taken during the tutorial

1 Planning and Learning

a) In linear function approximation, the value of a state is approximated as a linear combination of a feature vector $\boldsymbol{x}(s)$ and a weight vector $\boldsymbol{w}$, such that $\hat{v}(s, \boldsymbol{w}) = \boldsymbol{x}(s) \cdot \boldsymbol{w}$.

   In the tabular case, we simply store the dervied value function for each state. The feature vector for each state can be constructed as a one-hot indicator vector with $x_i(s) = 1$ only for the present state and 0 for all other states. Then the weight vector $\boldsymbol{w}$ will consist of values corresponding to individual states such that $\boldsymbol{x}(s) \cdot \boldsymbol{w}$ will give the value of one particular state.

b) Update rules for Sarsa(λ) [while updating the state action values]:

   - In the tabular case: we need an eligibility trace for each action value pair:

     $$E_0(s, a) = 0$$
     $$E_t(s, a) = \gamma\lambda E_{t-1}(s, a) + \mathbf{1}(S_t = s, A_t = a)$$

     Then we update $Q(s, a)$ for every (S, A) proportionally to TD-error δ_t:

     $$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$
     $$Q(S, A) \leftarrow Q(S, A) + \alpha\delta_t E_t(S, A)$$

   - With function approximation: The state action value function is parameterized by the weights $\mathbf{w}$. Hence the weights need to be updated in each step:

     $$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[Q_\pi(S_t, A_t) - \hat{Q}(S_t, A_t, \mathbf{w}_t)]\nabla\hat{Q}(S_t, A_t, \mathbf{w})$$

     where $\hat{Q}$ is the approximated Q function using weights $\mathbf{w}$. The update uses stochastic gradient descent to find the local minimum. Q_π is the true value function which is used to find the error in each step.

   - With linear function approximation: Using linear function approximation, each state-action pair is represented by a feature vector $\mathbf{x}(s, a)$. In this case, the derivative

     $$\nabla\hat{Q}(S_t, A_t, \mathbf{w}) = \mathbf{x}(s, a)$$

     Hence, the update step is

     $$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[Q_\pi(S_t, A_t) - \hat{Q}(S_t, A_t, \mathbf{w}_t)]\mathbf{x}(s, a)$$