Worksheet C3M3

1. (*Exercise 10.2 S&B*) Give pseudocode for semi-gradient one-step Expected Sarsa for control. You can build on the semi-gradient Sarsa code for this question.

Episodic Semi-gradient Sarsa for Estimating $\hat{q} \approx q_*$

```
Input: a differentiable action-value function parameterization q̂ : S × A × R^d → R
Algorithm parameters: step size α > 0, small ε > 0
Initialize value-function weights w ∈ R^d arbitrarily (e.g., w = 0)

Loop for each episode:
    S, A ← initial state and action of episode (e.g., ε-greedy)
    Loop for each step of episode:
        Take action A, observe R, S'
        If S' is terminal:
            w ← w + α[R − q̂(S, A, w)]∇q̂(S, A, w)
            Go to next episode
        Choose A' as a function of q̂(S', ·, w) (e.g., ε-greedy)
        w ← w + α[R + γq̂(S', A', w) − q̂(S, A, w)]∇q̂(S, A, w)
        S ← S'
        A ← A'
```

2. (*Exercise 10.1 S&B*) We have not explicitly considered or given pseudocode for any Monte Carlo methods in this chapter. What would they be like? Why is it reasonable not to give pseudocode for them?

3. How would you use optimistic initial values, for Sarsa with a tile coding function approximator? Assume you have a two dimensional input, and you use m tilings, and n tiles, to give m grids of size $n \times n$ resulting in $m \times n \times n$ features. What size is your weight vector? And how do you initialize your weights to ensure you have optimistic initial values? Assume the maximum reward is $R_{\max}$ and we use a $\gamma < 1$.

4. (*Exercise 10.8 S&B*) The pseudocode in the box on page 251 updates $\bar{R}_t$ using δ_t as an error rather than simply $R_{t+1} - \bar{R}_t$. Both errors work, but using δ_t is better. To see why, consider the ring MRP of three states from Exercise 10.7. The estimate of the average reward should tend towards its true value of $\frac{1}{3}$. Suppose you fix $\bar{R}_t = \frac{1}{3}$ and fix $v_\pi(A) = \frac{-1}{3}, v_\pi(B) = 0, v_\pi(C) = \frac{1}{3}$, which are the true values. What is the sequence of $R_{t+1} - \bar{R}_t$ errors, when going from A to B, B to C and then C to A? Correspondingly, what is the sequence of TD errors? Here, since we use the true values, we have $\delta_t = R_{t+1} - \bar{R}_t + v_\pi(S_{t+1}) - v_\pi(S_t)$. What does this tell us about which error sequence would produce a more stable estimate of the average reward if the estimates were allowed to change in response to the errors? Why?