Chapter 10 solutions **Hanna Gábor**

1. *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? How would they perform on the Mountain Car task?*

   The Monte Carlo is basically n-step Sarsa with $n = T$. I expect it to perform poorly on the Mountain Car example: until it doesn't reach the goal, it doesn't learn anything, so the first episode might be really long. As opposed to this, n-step Sarsa tries new actions after it sees that the previous actions didn't lead to the end of the episode, so it will get to the goal line eventually.

2. *Give pseudocode for semi-gradient one-step Expected Sarsa for control.*

Algorithm 1 Semi-gradient expected Sarsa for control

Inputs: a differentiable action-value function parametrization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: steps size $\alpha > 0, \epsilon > 0$
Initialize value-function weights $w \in \mathbb{R}^d$ arbitrarily
for each episode **do**
  Initialize S_0 non-terminal state.
  for $t = 0, 1, 2 \ldots$ **do**
    Take action A_t according to the ϵ-greedy policy w.r.t. $\hat{q}(S_0, \cdot, w)$
    Observe reward R_{t+1} and next state S_{t+1}.
    if S_{t+1} is terminal **then**
      $w \leftarrow w + \alpha R_{t+1} \nabla \hat{q}(S_t, A_t, w)$
      go to next episode
    else
      $w \leftarrow w + \alpha \Big(R_{t+1} + \sum_a \pi(a|S_{t+1}) \hat{q}(S_{t+1}, a, w) - \hat{q}(S_t, A_t, w) \Big) \nabla \hat{q}(S_t, A_t, w)$, where
      π is the ϵ-greedy policy w.r.t. $\hat{q}(S_0, \cdot, w)$
      $S_t \leftarrow S_{t+1}$
    end if
  end for
end for

3. *Why do the results shown in Figure 10.4 have higher standard errors at large n than at small n?*

   If n is large, it's less predictable where we will end up in n steps. This causes our estimates to have a bigger standard deviation. That in turn, will cause the steps to be more varied among different runs.

4. *Give pseudocode for a differential version of semi-gradient Q-learning* See the algorithm on the next page. I used the average of the seen rewards as $\bar{R}$.

5. *What equations are needed (beyond 10.10) to specify the differential version of TD(0)?*

$$w_{t+1} = w_t + \alpha \delta_t \nabla \hat{v}(S_t, w_t)$$

Algorithm 2 Differential version of semi-gradient Q-learning

Inputs: a differentiable action-value function parametrization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: steps size $\alpha > 0, \epsilon > 0$
Initialize value-function weights $w \in \mathbb{R}^d$ arbitrarily.
$\bar{R} \leftarrow 0$
$visits \leftarrow 0$
for each episode **do**
  Initialize S_0 non-terminal state.
  for $t = 0, 1, 2 \ldots$ **do**
    Choose action A_t according to the ϵ-greedy policy w.r.t. $\hat{q}$
    Observe reward R_{t+1} and next state S_{t+1}.
    $w \leftarrow w + \alpha(R_{t+1} - \bar{R} + \max_a(\hat{q}(S_{t+1}, a, w)) - \hat{q}(S_t, A_t, w))\nabla\hat{q}(S_t, A_t, w)$
    $\bar{R} \leftarrow \frac{visits}{visits+1}(\bar{R} + R_{t+1})$
    $visits \leftarrow visits + 1$
    if S_{t+1} is terminal **then**
      Go to next episode.
    end if
  end for
end for

6. *Suppose there is an MDP that under any policy produces the deterministic sequence of rewards +1, 0, +1, 0, +1, 0,... going on forever. Technically, this violates ergodicity; there is no stationary limiting distribution μ_π and the limit (10.7) does not exist. Nevertheless, the average reward (10.6) is well defined. What is it? Now consider two states in this MDP. From A, the reward sequence is exactly as described above, starting with a +1, whereas, from B, the reward sequence starts with a 0 and then continues with +1, 0, +1, 0,.... We would like to compute the differential values of A and B. Unfortunately, the differential return (10.9) is not well defined when starting from these states as the implicit limit does not exist. To repair this, one could alternatively define the differential value of a state as*

$$v_\pi(s) = \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t (\mathbb{E}(R_{t+1} | S_0 = s) - r(\pi)).$$

Under this definition, what are the differential values of states A and B?

The average reward is 0.5.

$$\begin{aligned}
\lim_{h\to\infty}\sum_{t=0}^{h}\gamma^t(\mathbb{E}(R_{t+1}|S_0=A)-r(\pi)) &= \sum_{t=0}^{\infty}\gamma^{2t}-r(\pi)\sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma^2}-\frac{0.5}{1-\gamma} \\
v_\pi(A) &= \lim_{\gamma\to 1}\frac{1}{1-\gamma^2}-\frac{0.5}{1-\gamma} = \lim_{\gamma\to 1}\frac{1-0.5(1+\gamma)}{1-\gamma^2} \\
&= \lim_{\gamma\to 1}\frac{0.5(1-\gamma)}{1-\gamma^2} = \lim_{\gamma\to 1}0.5\frac{1}{1+\gamma} = 0.25 \\
\lim_{h\to\infty}\sum_{t=0}^{h}\gamma^t(\mathbb{E}(R_{t+1}|S_0=B)-r(\pi)) &= \sum_{t=0}^{\infty}\gamma^{2t+1}-r(\pi)\sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma^2}-\frac{0.5}{1-\gamma} \\
v_\pi(B) &= \lim_{\gamma\to 1}\frac{\gamma}{1-\gamma^2}-\frac{0.5}{1-\gamma} = \lim_{\gamma\to 1}\frac{\gamma-0.5(1+\gamma)}{1-\gamma^2} \\
&= \lim_{\gamma\to 1}\frac{0.5(\gamma-1)}{1-\gamma^2} = \lim_{\gamma\to 1}-0.5\frac{1}{1+\gamma} = -0.25
\end{aligned}$$