Chapter 10

Exercise 10.1:

Monte carlo methods code will have the full return instead of the bootstrapping return. It is reasonable not to give pseudocode for monte carlo because monte carlo is basically a specific case of N-Step TD algorithm (where N=T-t).
On the mountain car example monte carlo method might perform poorly because of several reasons. Firstly, monte carlo method will have to complete the first episode to perform any updates. To complete the first episode, it will have to rely on a random policy. The first episode itself may last for a very long time. Secondly, even if the first episode is complete, monte carlo methods will take a lot of time to converge to optimal policy because they have to wait till completion of every episode.

Exercise 10.2:

In the pseudocode on page 244, do not choose A^1 and replace the update with:

$$w \leftarrow w + \alpha[R + \gamma \sum_a \pi(a|S^1)q(S^1, a, w) - q(S, A, w)]\partial q(S, A, w)$$

Exercise 10.3:

In the initial episodes, the trajectory of each episode will be somewhat random and have many "bad" states. Large value of N will take more of these "bad" states into account and the initial values of many states will have a very high variance. This effect will eventually die down however since the result shown is after only 50 episodes, the effect is still prominent. On the other hand, lower values of N consider only a few next states of the trajectory. Even if these states are bad, the effect due to the updates will be small and die out quickly.

Exercise 10.4:

In the pseudocode for semi gradient SARSA, we need to make the following changes:
Choose action A according to $q(S_t, ., w)$
Take action and observe $R_{t+1} and\ S_{t+1}$
$w \leftarrow w + \alpha[R_{t+1} + \gamma max_a\ q(S_{t+1},\ a,\ w) - q(S_t,\ A_t,\ w)]\partial q(S_t, A_t, w)$

Here again it is obvious that we do not choose A_{t+1} at time t since the Q-Learning update involves a maximum overall actions.

Exercise 10.5:

Apart from the TD error and the update, we need the equation to calculate $\overline{R_t}$ which is an estimation of $r(\pi)$ at time t.

Remember that $r(\pi)$ is the average reward of the policy at steady state. However, since the environment might be stochastic and all transitions might not be known to us, we do not have values of $p(s^1, r|s, a)$ and $\mu_\pi(s)$.
Instead, we have the estimates of the above two $\mu_t(s)$ and $p_t(s^1, r|s, a)$ from the previous interactions with the environment.

Thus, $\overline{R_t} = \sum_s \mu_t(s) \sum_a \pi(a|s) \sum_{s^1,r} p(s^1, r|s, a)r$

Where s, s^1, r are from the set of states and reward which we have experienced so far.

Exercise 10.6:

It is easy to see that the average reward according to eq 10.6 is $r(\pi) = 1/2$.

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t (E[R_{t+1}|S_0 = A] - 1/2)$$

Reward at A starts with +1:

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0 (1 - 1/2) + \gamma^1 (0 - 1/2) + \gamma^2 (1 - 1/2)...$$

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0 * 1/2 - \gamma^1 * 1/2 + \gamma^2 * 1/2...$$

$$v(A) = (1/2) * lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h} (-1)^t \gamma^t$$

$$v(A) = (1/2) * lim_{\gamma \to 1} \sum_{t=0}^{\infty} (-1)^t \gamma^t$$

$$v(A) = (1/2) * lim_{\gamma \to 1} \frac{1}{1+\gamma}$$
$$v(A) = (1/2) * (1/2)$$
$$v(A) = 1/4$$

At B rewards start with 0:

$$v(B) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0(0 - 1/2) + \gamma^1(1 - 1/2) + \gamma^2(0 - 1/2)...$$
$$v(B) = lim_{\gamma \to 1} lim_{h \to \infty} -\gamma^0 * 1/2 + \gamma^1 * 1/2 - \gamma^2 * 1/2...$$
$$v(B) = (1/2) * lim_{\gamma \to 1} lim_{h \to \infty} - 1 * \sum_{t=0}^{h} (-1)^t \gamma^t$$
$$v(B) = (1/2) * lim_{\gamma \to 1} - 1 * \sum_{t=0}^{\infty} (-1)^t \gamma^t$$

$$v(B) = (1/2) * lim_{\gamma \to 1} \frac{-1}{1+\gamma}$$
$$v(B) = (1/2) * (-1/2)$$
$$v(B) = \ -1/4$$

Exercise 10.7:

We need to first compute $r(\pi)$, the average reward.
Here again, ergodicity is violated.
According to eq 10.6, $r(\pi) = 1/3$ (out of every 3 consecutive rewards, two will be 0 and one will be 1 thus the summation is basically h/3).

Eq 10.13:

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

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t (E[R_{t+1} | S_0 = A] - 1/3)$$

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0 (0 - 1/3) + \gamma^1 (0 - 1/3) + \gamma^2 (1 - 1/3) \ldots$$

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h/3} - \gamma^{3t} * 1/3 - \gamma^{3t+1} * 1/3 + \gamma^{3t+2} * 2/3$$

$$v(A) = lim_{\gamma \to 1} lim_{h \to \infty} [\sum_{t=0}^{h/3} - \gamma^{3t} * 1/3 - \sum_{t=0}^{h/3} \gamma^{3t+1} * 1/3 + \sum_{t=0}^{h/3} \gamma^{3t+2} * 2/3]$$

$$v(A) = lim_{\gamma \to 1} [- \frac{1}{3(1-\gamma^3)} - \frac{\gamma}{3(1-\gamma^3)} + \frac{2\gamma^2}{3(1-\gamma^3)}]$$

$$v(A) = lim_{\gamma \to 1} [\frac{2\gamma^2 - \gamma - 1}{3(1-\gamma^3)}]$$

$$v(A) = lim_{\gamma \to 1} - [\frac{4\gamma - 1}{9\gamma^2}]$$

$$v(A) = \ - 1/3$$

Similarly:

$$v(B) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0 (0 - 1/3) + \gamma^1 (1 - 1/3) + \gamma^2 (0 - 1/3) \ldots$$

$$v(B) = lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h/3} - \gamma^{3t} * 1/3 + \gamma^{3t+1} * 2/3 - \gamma^{3t+2} * 1/3$$

$$v(B) = lim_{\gamma \to 1} lim_{h \to \infty} [\sum_{t=0}^{h/3} - \gamma^{3t} * 1/3 + \sum_{t=0}^{h/3} \gamma^{3t+1} * 2/3 - \sum_{t=0}^{h/3} \gamma^{3t+2} * 1/3]$$

$$v(B) = lim_{\gamma \to 1} [- \frac{1}{3(1-\gamma^3)} + \frac{2\gamma}{3(1-\gamma^3)} - \frac{\gamma^2}{3(1-\gamma^3)}]$$

$$v(B) = lim_{\gamma \to 1} [\frac{-\gamma^2 + 2\gamma - 1}{3(1-\gamma^3)}]$$

$$v(B) = lim_{\gamma \to 1} - [\frac{2 - 2\gamma}{9\gamma^2}]$$

$$v(B) = \ 0$$

Similarly:

$$v(C) = lim_{\gamma \to 1} lim_{h \to \infty} \gamma^0(1 - 1/3) + \gamma^1(0 - 1/3) + \gamma^2(0 - 1/3) \ldots$$

$$v(C) = lim_{\gamma \to 1} lim_{h \to \infty} \sum_{t=0}^{h/3} \gamma^{3t} * 2/3 - \gamma^{3t+1} * 1/3 - \gamma^{3t+2} * 1/3$$

$$v(C) = lim_{\gamma \to 1} lim_{h \to \infty} [\sum_{t=0}^{h/3} \gamma^{3t} * 2/3 - \sum_{t=0}^{h/3} \gamma^{3t+1} * 1/3 - \sum_{t=0}^{h/3} \gamma^{3t+2} * 1/3]$$

$$v(C) = lim_{\gamma \to 1} [\frac{2}{3(1-\gamma^3)} - \frac{\gamma}{3(1-\gamma^3)} - \frac{\gamma^2}{3(1-\gamma^3)}]$$

$$v(C) = lim_{\gamma \to 1} [\frac{-\gamma^2-\gamma+2}{3(1-\gamma^3)}]$$

$$v(C) = lim_{\gamma \to 1} [\frac{1+2\gamma}{9\gamma^2}]$$

$$v(C) = 1/3$$

Exercise 10.8:

If the value of $\overline{R}_t$ is fixed at 1/3
Assuming start from A:
The error sequence $R_{t+1} - \overline{R}_t$ will be $-1/3,\ -1/3,\ 2/3,\ -1/3,\ -1/3,\ -1/3$....
The δ_t sequence will be:
$-1/3 + 0 + 1/3 = 0,\ -1/3 + 1/3 - 0 = 0,\ 2/3 - 1/3 - 1/3 = 0, \ldots$

Hence we see that all δ_t are 0 and hence no updates will take place.
Given that we were already at steady state with $\overline{R}_t = 1/3$ using the reward error would still cause $\overline{R}_t$ to fluctuate. However using δ_t nothing will change. Therefore using δ_t produces a more stable estimate.

Exercise 10.9:

In the pseudocode on page 255, Instead of updating $\overline{R}$ after calculating δ, we need to update it immediately after observing the reward R_{t+1}.
We also need to maintain o_t
The following lines need to be added after observing reward:

$$o_t = o_{t-1} + \alpha^1(1 - o_{t-1})$$

$$\beta = \alpha^1 / o_t$$

$$\overline{R} \leftarrow \overline{R} + \beta(R_{t+1} - \overline{R})$$

Where α^1 is the constant step size parameter for $\overline{R}$