Exercises - Chapter 10

Carl Fredriksson, c@msp.se

Exercise 10.1

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?

My answer:

> What would they be like?

Generate episodes in the same way, but update value estimates only in between episodes using full returns from the previous episode.

> Why is it reasonable not to give pseudocode for them?

Because they are special cases of n-step Sarsa (by setting n to be the length of the the episodes).

> How would they perform on the Mountain Car task?

Poorly. The performance seems to be best for intermediate levels of bootstrapping (peaking at $n = 4$). Larger n leads to worse performance, and as mentioned above, Monte Carlo (MC) methods are equivalent to the extreme case with the maximum n. We can also reason about how the episodes would play out. With MC methods and $\mathbf{w}$ initialized to $\mathbf{0}$, we would have very long early episodes. Since updates are only made after reaching the terminal state, action selection would be completely random for the entirety of the first episode, which would be incredibly long for the majority of runs.

Exercise 10.2

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

My answer:

- Input: a differentiable action-value function parameterization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
- Input: a policy π (if estimating q_π)
- Algorithm parameters: step size $\alpha > 0$, small $\epsilon > 0$
- Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)
- Loop for each episode:
    - Initialize $S \neq terminal$
    - Loop for each step of episode$
        - Select $A \sim \pi(\cdot|S)$ or ϵ-greedy wrt $\hat{q}(S, \cdot, \mathbf{w})$
        - Take action A, observe R, S'
        - $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[R + \sum_a \pi(a|S')\hat{q}(S', a, \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\big]\nabla\hat{q}(S, A, \mathbf{w})$
        - $S \leftarrow S'$
        - If S is terminal:
            - Go to next episode

Exercise 10.3

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

My answer:

I think it's due to larger n resulting in more possible trajectories per update, thus increasing the variance.

Exercise 10.4

Give pseudocode for a differential version of semi-gradient Q-learning.

My answer:

- Input: a differentiable action-value function parametrization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
- Algorithm parameters: step sizes $\alpha, \beta > 0$, small $\epsilon > 0$
- Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)
- Initialize average reward estimate $\bar{R} \in \mathbb{R}$ arbitrarily (e.g., $\bar{R} = 0$)
- Initialize state S
- Loop for each step:
    - Choose A as a function of $\hat{q}(S, \cdot, \mathbf{w})$ (e.g., ϵ-greedy)
    - Take action A, observe R, S'
    - $\delta \leftarrow R - \bar{R} + \max_a \hat{q}(S', a, \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
    - $\bar{R} \leftarrow \bar{R} + \beta\delta$
    - $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$
    - $S \leftarrow S'$

Exercise 10.5

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

My answer:

$\bar{R}$-update:

$$\bar{R}_{t+1} \leftarrow \bar{R}_t + \beta\delta_t$$

$\mathbf{w}$-update:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha\delta_t\nabla\hat{v}(S_t, \mathbf{w}_t)$$

Exercise 10.6

Suppose there is an MDP that under any policy produces the deterministic sequence of rewards $+1, 0, +1, 0, +1, 0, \ldots$ 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 $+1, 0, +1, 0, \ldots$ 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

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

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

My answer:

Nevertheless, the average reward (10.6) is well defined. What is it?

$$
\begin{aligned}
r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_{t+1} | S_0, A_{0:t-1} \sim \pi] \\
&= \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h/2} 1 + 0 \\
&= \frac{1}{2}
\end{aligned}
$$

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

$$
\begin{aligned}
v_\pi(A) &= \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = A] - r(\pi) \Big) \\
&= \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = A] - \frac{1}{2} \Big) \\
&= \lim_{\gamma\to 1} \Big[\gamma^0 (1 - \frac{1}{2}) + \gamma^1 (0 - \frac{1}{2}) + \gamma^2 (1 - \frac{1}{2}) + \gamma^3 (0 - \frac{1}{2}) + \dots \Big] \\
&= \frac{1}{2} \lim_{\gamma\to 1} \Big[1 - \gamma + \gamma^2 - \gamma^3 + \dots \Big] \\
&= \frac{1}{2} \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} (-1)^t \gamma^t \\
&= \frac{1}{2} \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} (-\gamma)^t \\
&= \frac{1}{2} \lim_{\gamma\to 1} \frac{1}{1 - (-\gamma)} \\
&= \frac{1}{2} \lim_{\gamma\to 1} \frac{1}{1 + \gamma} \\
&= \frac{1}{4}
\end{aligned}
$$

$$
\begin{aligned}
v_\pi(B) &= \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = B] - r(\pi) \Big) \\
&= \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = B] - \frac{1}{2} \Big) \\
&= \lim_{\gamma\to 1} \Big[\gamma^0 (0 - \frac{1}{2}) + \gamma^1 (1 - \frac{1}{2}) + \gamma^2 (0 - \frac{1}{2}) + \gamma^3 (1 - \frac{1}{2}) + \dots \Big] \\
&= \frac{1}{2} \lim_{\gamma\to 1} \Big[-1 + \gamma - \gamma^2 + \gamma^3 - \dots \Big] \\
&= \frac{1}{2} \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} (-1)^{t+1} \gamma^t \\
&= -\frac{1}{2} \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} (-1)^t \gamma^t \\
&= -\frac{1}{4}
\end{aligned}
$$

Exercise 10.7

Consider a Markov reward process consisting of a ring of three states A, B, and C, with state transitions going deterministically around the ring. A reward of $+1$ is received upon arrival in A and otherwise the reward is 0. What are the differential values of the three states, using (10.13)?

My answer:

$$r(\pi) = \frac{1}{3}$$

$$\begin{aligned}
v_\pi(A) &= \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = A] - r(\pi) \Big) \\
&= \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \Big(\mathbb{E}_\pi[R_{t+1} | S_0 = A] - \frac{1}{3} \Big) \\
&= \lim_{\gamma \to 1} \Big[\gamma^0 (0 - \frac{1}{3}) + \gamma^1 (0 - \frac{1}{3}) + \gamma^2 (1 - \frac{1}{3}) + \gamma^3 (0 - \frac{1}{3}) + \gamma^4 (0 - \frac{1}{3}) + \gamma^5 (1 - \frac{1}{3}) + \dots \Big] \\
&= \frac{1}{3} \lim_{\gamma \to 1} \Big[-1 - \gamma + 2\gamma^2 - \gamma^3 - \gamma^4 + 2\gamma^5 - \dots \Big] \\
&= \frac{1}{3} \lim_{\gamma \to 1} \Big[(-1 - \gamma - \gamma^3 - \dots) + 2(\gamma^2 + \gamma^5 + \gamma^8 + \dots) \Big] \\
&= \lim_{\gamma \to 1} \Big[\frac{1}{3}(-1 - \gamma - \gamma^2 - \dots) + \gamma^2 (1 + \gamma^3 + \gamma^6 + \dots) \Big] \\
&= \lim_{\gamma \to 1} \Big[-\frac{1}{3} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t + \gamma^2 \lim_{h \to \infty} \sum_{t=0}^{h} (\gamma^3)^t \Big] \\
&= \lim_{\gamma \to 1} \Big[-\frac{1}{3(1-\gamma)} + \frac{\gamma^2}{(1-\gamma^3)} \Big] \\
&= \lim_{\gamma \to 1} \frac{3\gamma^2(1-\gamma) - (1-\gamma^3)}{3(1-\gamma)(1-\gamma^3)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{(\gamma-1)^2(-2\gamma-1)}{(\gamma-1)^2(\gamma^2+\gamma+1)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{-2\gamma-1}{\gamma^2+\gamma+1} \\
&= -\frac{1}{3}
\end{aligned}$$

$$
\begin{aligned}
v_\pi(B) &= \lim_{\gamma \to 1} \left[\gamma^0 (0 - \frac{1}{3}) + \gamma^1 (1 - \frac{1}{3}) + \gamma^2 (0 - \frac{1}{3}) + \dots \right] \\
&= \frac{1}{3} \lim_{\gamma \to 1} \left[-1 + 2\gamma - \gamma^2 - \dots \right] \\
&= \frac{1}{3} \lim_{\gamma \to 1} \left[(-1 - \gamma^2 - \gamma^3 - \dots) + 2(\gamma + \gamma^4 + \gamma^7 + \dots) \right] \\
&= \lim_{\gamma \to 1} \left[\frac{1}{3} (-1 - \gamma - \gamma^2 - \dots) + \gamma (1 + \gamma^3 + \gamma^6 + \dots) \right] \\
&= \lim_{\gamma \to 1} \left[-\frac{1}{3(1-\gamma)} + \frac{\gamma}{(1-\gamma^3)} \right] \\
&= \lim_{\gamma \to 1} \frac{3\gamma(1-\gamma) - (1-\gamma^3)}{3(1-\gamma)(1-\gamma^3)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{(\gamma-1)^3}{(\gamma-1)^2(\gamma^2+\gamma+1)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{\gamma-1}{\gamma^2+\gamma+1} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
v_\pi(C) &= \lim_{\gamma \to 1} \left[\gamma^0 (1 - \frac{1}{3}) + \gamma^1 (0 - \frac{1}{3}) + \gamma^2 (0 - \frac{1}{3}) + \dots \right] \\
&= \frac{1}{3} \lim_{\gamma \to 1} \left[2 - \gamma - \gamma^2 - \dots \right] \\
&= \lim_{\gamma \to 1} \left[\frac{1}{3} (-1 - \gamma - \gamma^2 - \dots) + (1 + \gamma^3 + \gamma^6 + \dots) \right] \\
&= \lim_{\gamma \to 1} \left[-\frac{1}{3(1-\gamma)} + \frac{1}{(1-\gamma^3)} \right] \\
&= \lim_{\gamma \to 1} \frac{3(1-\gamma) - (1-\gamma^3)}{3(1-\gamma)(1-\gamma^3)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{(\gamma-1)^2(\gamma+2)}{(\gamma-1)^2(\gamma^2+\gamma+1)} \\
&= \frac{1}{3} \lim_{\gamma \to 1} \frac{\gamma+2}{\gamma^2+\gamma+1} \\
&= \frac{1}{3}
\end{aligned}
$$

Exercise 10.8

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 it was already there and was held stuck there. What would the sequence of $R_{t+1} - \bar{R}_t$ errors be? What would the sequence of δ_t errors be (using Equation 10.10)? Which error sequence would produce a more stable estimate of the average reward if the estimate were allowed to change in response to the errors? Why?

My answer:

> Suppose it was already there and was held stuck there. What would the sequence of $R_{t+1} - \bar{R}_t$ errors be?

$$A \to B : 0 - \frac{1}{3} = -\frac{1}{3}$$

$$B \to C : 0 - \frac{1}{3} = -\frac{1}{3}$$

$$C \to A : 1 - \frac{1}{3} = \frac{2}{3}$$

> What would the sequence of δ_t errors be (using Equation 10.10)?

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{v}(S_{t+1}) - \hat{v}(S_t)$$

$$A \to B : 0 - \frac{1}{3} + \hat{v}(B) - \hat{v}(A) = -\frac{1}{3} + 0 - (-\frac{1}{3}) = 0$$

$$B \to C : 0 - \frac{1}{3} + \hat{v}(C) - \hat{v}(B) = -\frac{1}{3} + \frac{1}{3} - 0 = 0$$

$$C \to A : 1 - \frac{1}{3} + \hat{v}(A) - \hat{v}(C) = \frac{2}{3} + (-\frac{1}{3}) - \frac{1}{3} = 0$$

> Which error sequence would produce a more stable estimate of the average reward if the estimate were allowed to change in response to the errors?

The sequence using δ_t errors.

> Why?

The δ_t errors are all 0 and thus $\bar{R}_t$ would not change. The $R_{t+1} - \bar{R}_t$ errors are not 0, and if used in updates, $\bar{R}_t$ would change after every update assuming $\beta > 0$.

Exercise 10.9

In the differential semi-gradient n-step Sarsa algorithm, the step-size parameter on the average reward, β, needs to be quite small so that $\bar{R}$ becomes a good long-term estimate of the average reward. Unfortunately, $\bar{R}$ will then be biased by its initial value for many steps, which may make learning inefficient. Alternatively, one could use a sample average of the observed rewards for $\bar{R}$. That would initially adapt rapidly but in the long run would also adapt slowly. As the policy slowly changed, $\bar{R}$ would also change; the potential for such long-term nonstationarity makes sample-average methods ill-suited. In fact, the step-size parameter on the average reward is a perfect place to use the unbiased constant-step-size trick from Exercise 2.7. Describe the specific changes needed to the boxed algorithm for differential semi-gradient n-step Sarsa to use this trick.

My answer:

- Add to initialization:
  - $\bar{o} = 0$
- Add before updating $\bar{R}$:
  - $\bar{o} \leftarrow \bar{o} + \beta(1 - \bar{o})$
- Change $\bar{R} \leftarrow \bar{R} + \beta\delta$ to:
  - $\bar{R} \leftarrow \bar{R} + (\beta / \bar{o})\delta$