Exercises - Chapter 7

Carl Fredriksson, c@msp.se

Exercise 7.1

In Chapter 6 we noted that the Monte Carlo error can be written as the sum of TD errors (6.6) if the value estimates don't change from step to step. Show that the n-step error used in (7.2) can also be written as a sum of TD errors (again if the value estimates don't change) generalizing the earlier result.

My answer:

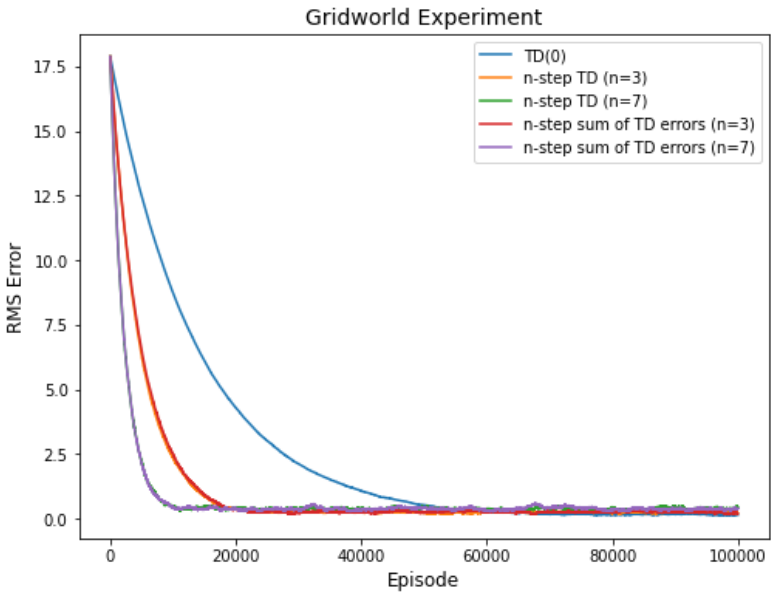
$$
\begin{aligned}
G_{t:t+n} - V(S_t) &= R_{t+1} + \gamma G_{t+1:t+n} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma \big[G_{t+1:t+n} - V(S_{t+1}) \big] \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \big[G_{t+2:t+n} - V(S_{t+2}) \big] \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \cdots + \gamma^{n-1} \big[G_{t+n-1:t+n} - V(S_{t+n-1}) \big] \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \cdots + \gamma^{n-1} \big[R_{t+n} + \gamma V(S_{t+n}) - V(S_{t+n-1}) \big] \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \cdots + \gamma^{n-1} \delta_{t+n-1} \\
&= \sum_{k=t}^{t+n-1} \gamma^{k-t} \delta_k
\end{aligned}
$$

Exercise 7.2 (programming)

With an n-step method, the value estimates do change from step to step, so an algorithm that used the sum of TD errors (see previous exercise) in place of the error in (7.2) would actually be a slightly different algorithm. Would it be a better algorithm or a worse one? Devise and program a small experiment to answer this question empirically.

My answer:

I programmed an experiment using the 4x4 gridworld in Example 4.1. There was a big difference in convergence speed when comparing the n-step methods to TD(0). However, I did not see any significant difference between n-step TD and n-step sum of TD errors. A constant step size of $\alpha = 0.001$ was used for all algorithms and the results were averaged over 10 runs per algorithm.



Exercise 7.3

Why do you think a larger random walk task (19 states instead of 5) was used in the examples of this chapter? Would a smaller walk have shifted the advantage to a different value of n? How about the change in left-side outcome from 0 to 1 made in the larger walk? Do you think that made any difference in the best value of n?

My answer:

I think 19 states was used instead of 5 in order to enable more n values that are reasonable. When $t + n \geq T$ for most time steps t, then n-step TD will resemble a Monte Carlo method, since:

> If $t+n \geq T$ (if the n-step return extends to or beyond termination), then all the missing terms are taken as zero, and the n-step return defined to be equal to the ordinary full return ($G_{t:t+n} = G_t$ if $t+n \geq T$).

I think a smaller walk would have shifted the advantage to smaller values of n, since it seems like intermediate values of n work best on this task. As mentioned above, the larger n is compared to the size of the walk, the closer the method will be to the Monte Carlo extreme. Trajectories that start of to the right before eventually terminating to the left, and vice versa, are more likely to move some of the value estimates the wrong direction with larger values of n.

I think changing the left-side outcome from 0 to -1 was made to make terminating to the left immediately meaningful. Otherwise, with initializing all values to 0, initial trajectories terminating to the left would result in no update.

I think changing from 0 to -1 penalizes larger values of n, because:

- More updates will use ordinary full returns without bootstrapping - and ordinary full returns from trajectories terminating to the left would increase variance more with -1.
- As mentioned above, trajectories that start of to the right before eventually terminating to the left, and vice versa, are more likely to move some of the value estimates the wrong direction with larger values of n - this would be more consequential with -1.

Exercise 7.4

Prove that the n-step return of Sarsa (7.4) can be written exactly in terms of a novel TD error, as:

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{min(t+n,T)-1} \gamma^{k-t} \big[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k) \big]$$

My answer:

Let $\delta_t = R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)$ and $t+n < T$. Then we have:

$$\begin{aligned}
G_{t:t+n} &= R_{t+1} + \gamma G_{t+1:t+n} + Q_{t-1}(S_t, A_t) - Q_{t-1}(S_t, A_t) + \gamma Q_t(S_{t+1}, A_{t+1}) - \gamma Q_t(S_{t+1}, A_{t+1}) \\
&= Q_{t-1}(S_t, A_t) + \big[R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) \big] + \gamma \big[G_{t+1:t+n} - Q_t(S_{t+1}, A_{t+1}) \big] \\
&= Q_{t-1}(S_t, A_t) + \big[R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) \big] \\
&\quad + \gamma \big[R_{t+2} + \gamma Q_{t+1}(S_{t+2}, A_{t+2}) - Q_t(S_{t+1}, A_{t+1}) \big] + \gamma^2 \big[G_{t+2:t+n} - Q_{t+1}(S_{t+2}, A_{t+2}) \big] \\
&= Q_{t-1}(S_t, A_t) + \delta_t + \gamma \delta_{t+1} + \cdots + \gamma^{n-1} \delta_{t+n-1} + \gamma^n \big[G_{t+n:t+n} - Q_{t+n-1}(S_{t+n}, A_{t+n}) \big] \\
&= Q_{t-1}(S_t, A_t) + \delta_t + \gamma \delta_{t+1} + \cdots + \gamma^{n-1} \delta_{t+n-1} + \gamma^n \big[Q_{t+n-1}(S_{t+n}, A_{t+n}) - Q_{t+n-1}(S_{t+n}, A_{t+n}) \big] \\
&= Q_{t-1}(S_t, A_t) + \delta_t + \gamma \delta_{t+1} + \cdots + \gamma^{n-1} \delta_{t+n-1} \\
&= Q_{t-1}(S_t, A_t) + \sum_{k=t}^{t+n-1} \gamma^{k-t} \big[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k) \big]
\end{aligned}$$

In general, where $t+n \geq T$ is also possible, we get:

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{min(t+n,T)-1} \gamma^{k-t} \big[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k) \big]$$

Exercise 7.5

Write the pseudocode for the off-policy state-value prediction algorithm described above.

My answer:

One version:

- Input: a behavior policy b and a target policy π
- Algorithm parameters: step size $\alpha \in (0, 1]$, a positive integer n
- Initialize $V(s)$ arbitrarily, for all $s \in \mathcal{S}$
- All store and access operations (for S_t and R_t) can take their index mod $n+1$
- Loop for each episode:
    - Initialize and store $S_0 \neq \text{terminal}$
    - $T \leftarrow \infty$
    - Loop for $t = 0, 1, 2, \cdots$:
        - If $t < T$, then:
            - Take an action according to $b(\cdot|S_t)$
            - Observe and store the next reward R_{t+1} and the next state S_{t+1}
            - If S_{t+1} is terminal, then $T \leftarrow t+1$

        - $\rho_t = \frac{\pi(S_t|A_t)}{b(S_t|A_t)}$
      - $\tau \leftarrow t - n + 1$ (τ is the time whose state's estimate is being updated)
      - If $\tau \geq 0$:
        - $G \leftarrow 0$
        - $h \leftarrow T$
        - If $\tau + n < T$, then:
          - $G \leftarrow V(S_{\tau+n})$
          - $h \leftarrow \tau + n$
        - Loop for $i = h-1, h-2, \ldots, \tau$:
          - $G \leftarrow \rho_i(R_{i+1} + \gamma G) + (1 - \rho_i)V(S_i)$
        - $V(S_\tau) \leftarrow V(S_\tau) + \alpha\big[G - V(S_\tau)\big]$
    - Until $\tau = T - 1$

Another version:

- Input: a behavior policy b and a target policy π
- Algorithm parameters: step size $\alpha \in (0, 1]$, a positive integer n
- Initialize $V(s)$ arbitrarily, for all $s \in \mathcal{S}$
- All store and access operations (for S_t and R_t) can take their index mod $n + 1$
- Loop for each episode:
  - Initialize and store $S_0 \neq \text{terminal}$
  - $T \leftarrow \infty$
  - Loop for $t = 0, 1, 2, \cdots$:
    - If $t < T$, then:
      - Take an action according to $b(\cdot|S_t)$
      - Observe and store the next reward R_{t+1} and the next state S_{t+1}
      - If S_{t+1} is terminal, then $T \leftarrow t + 1$
      - $\rho_t = \frac{\pi(S_t|A_t)}{b(S_t|A_t)}$
    - $\tau \leftarrow t - n + 1$ (τ is the time whose state's estimate is being updated)
    - If $\tau \geq 0$:
      - $G \leftarrow \left(\sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i \prod_{j=\tau}^{i-1} \rho_j \right) + (1 - \rho_\tau)V(S_\tau) + \left(\sum_{i=\tau+1}^{\min(\tau+n,T)-1} \gamma^{i-\tau}(1 - \rho_i)V(S_i) \prod_{j=\tau}^{i-2} \rho_j \right)$
      - If $\tau + n < T$, then: $G \leftarrow G + \gamma^n V(S_{\tau+n}) \prod_{j=\tau}^{\tau+n-1} \rho_j$
      - $V(S_\tau) \leftarrow V(S_\tau) + \alpha\big[G - V(S_\tau)\big]$
    - Until $\tau = T - 1$

Exercise 7.6

Prove that the control variate in the above equations does not change the expected value of the return.

My answer:

Let $G_{t:h}^{cv}$ denote the return with control variates. Let's assume that $h < T$, then we have:

$$
\begin{aligned}
G_{t:h}^{cv} &= R_{t+1} + \gamma\rho_{t+1}\big[G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\big] + \gamma\overline{V}_{h-1}(S_{t+1}) \\
&= R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) - \gamma\rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1}) + \\
&\quad + \gamma\rho_{t+1}\Big(R_{t+2} + \gamma\rho_{t+2}\big[G_{t+2:h} - Q_{h-1}(S_{t+2}, A_{t+2})\big] + \gamma\overline{V}_{h-1}(S_{t+2})\Big) \\
&= R_{t+1} + \gamma\rho_{t+1}R_{t+2} + \gamma^2\rho_{t+1:t+2}R_{t+3} + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}R_h + \\
&\quad + \gamma\overline{V}_{h-1}(S_{t+1}) + \gamma^2\rho_{t+1}\overline{V}_{h-1}(S_{t+2}) + \cdots + \gamma^{h-t}\rho_{t+1:h-1}\overline{V}_{h-1}(S_h) - \\
&\quad - \big[\gamma\rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1}) + \gamma^2\rho_{t+1:t+2}Q_{h-1}(S_{t+2}, A_{t+2}) + \cdots + \gamma^{h-t}\rho_{t+1:h}Q_{h-1}(S_h, A_h)\big] + \\
&\quad + \gamma^{h-t}\rho_{t+1:h}G_{h:h} \\
&= R_{t+1} + \gamma\rho_{t+1}R_{t+2} + \gamma^2\rho_{t+1:t+2}R_{t+3} + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}R_h + \\
&\quad + \gamma\overline{V}_{h-1}(S_{t+1}) + \gamma^2\rho_{t+1}\overline{V}_{h-1}(S_{t+2}) + \cdots + \gamma^{h-t}\rho_{t+1:h-1}\overline{V}_{h-1}(S_h) - \\
&\quad - \big[\gamma\rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1}) + \gamma^2\rho_{t+1:t+2}Q_{h-1}(S_{t+2}, A_{t+2}) + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}Q_{h-1}(S_{h-1}, A_{h-1})\big] \\
&= R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) + \sum_{i=t+1}^{h-1}\gamma^{i-t}\rho_{t+1:i}R_{i+1} + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\rho_{t+1:i}\overline{V}_{h-1}(S_{i+1}) - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\rho_{t+1:i}Q_{h-1}(S_i, A_i)
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
\mathbb{E}[G^{cv}_{t:h}] &= \mathbb{E}\Bigg[R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) + \sum_{i=t+1}^{h-1}\gamma^{i-t}\rho_{t+1:i}R_{i+1} + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\rho_{t+1:i}\overline{V}_{h-1}(S_{i+1}) - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\rho_{t+1:i}Q_{h-1}(S_i, A_i)\Bigg] \\
&= \mathbb{E}\big[R_{t+1}\big] + \gamma\mathbb{E}\big[\overline{V}_{h-1}(S_{t+1})\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}R_{i+1}\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\mathbb{E}\big[\rho_{t+1:i}\overline{V}_{h-1}(S_{i+1})\big] - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}Q_{h-1}(S_i, A_i)\big] \\
&= \mathbb{E}\big[R_{t+1}\big] + \gamma\mathbb{E}\big[\overline{V}_{h-1}(S_{t+1})\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}R_{i+1}\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\mathbb{E}\big[\rho_{t+1:i}\big]\mathbb{E}\big[\overline{V}_{h-1}(S_{i+1})\big] - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}\big]\mathbb{E}\big[Q_{h-1}(S_i, A_i)\big] \qquad (\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \text{ when } X, Y \text{ are independent}) \\
&= \mathbb{E}\big[R_{t+1}\big] + \gamma\mathbb{E}\big[\overline{V}_{h-1}(S_{t+1})\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}R_{i+1}\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\mathbb{E}\big[\overline{V}_{h-1}(S_{i+1})\big] - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[Q_{h-1}(S_i, A_i)\big] \qquad (\mathbb{E}[\rho_{t+1:i}] = 1) \\
&= \mathbb{E}\big[R_{t+1}\big] + \gamma\overline{V}_{h-1}(S_{t+1}) + \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}R_{i+1}\big] + \sum_{i=t+1}^{h-1}\gamma^{i-t+1}\overline{V}_{h-1}(S_{i+1}) - \\
&\quad - \sum_{i=t+1}^{h-1}\gamma^{i-t}\overline{V}_{h-1}(S_i) \\
&= \mathbb{E}\big[R_{t+1}\big] + \gamma^{h-t}\overline{V}_{h-1}(S_h) + \sum_{i=t+1}^{h-1}\gamma^{i-t}\mathbb{E}\big[\rho_{t+1:i}R_{i+1}\big] \\
&= \mathbb{E}\big[R_{t+1} + \gamma\rho_{t+1:t+1}R_{t+2} + \gamma^2\rho_{t+1:t+2}R_{t+3} + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}R_h + \gamma^{h-t}\rho_{t+1:h}\overline{V}_{h-1}(S_h)\big]
\end{aligned}
$$

Let's now assume $h \geq T$, then we have (I left out some steps for brevity):

$$
\begin{aligned}
G^{cv}_{t:h} &= R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) + \sum_{i=t+1}^{T-1}\gamma^{i-t}\rho_{t+1:i}R_{i+1} + \sum_{i=t+1}^{T-2}\gamma^{i-t+1}\rho_{t+1:i}\overline{V}_{h-1}(S_{i+1}) - \\
&\quad - \sum_{i=t+1}^{T-1}\gamma^{i-t}\rho_{t+1:i}Q_{h-1}(S_i, A_i)
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
\mathbb{E}[G^{cv}_{t:h}] &= \Bigg[R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) + \sum_{i=t+1}^{T-1}\gamma^{i-t}\rho_{t+1:i}R_{i+1} + \sum_{i=t+1}^{T-2}\gamma^{i-t+1}\rho_{t+1:i}\overline{V}_{h-1}(S_{i+1}) - \\
&\quad - \sum_{i=t+1}^{T-1}\gamma^{i-t}\rho_{t+1:i}Q_{h-1}(S_i, A_i)\Bigg] \\
&= \mathbb{E}\Bigg[R_{t+1} + \sum_{i=t+1}^{T-1}\gamma^{i-t}\rho_{t+1:i}R_{i+1}\Bigg] \\
&= \mathbb{E}\big[R_{t+1} + \gamma\rho_{t+1:t+1}R_{t+2} + \gamma^2\rho_{t+1:t+2}R_{t+3} + \cdots + \gamma^{T-t-1}\rho_{t+1:T-1}R_T\big]
\end{aligned}
$$

Thus we have shown that the control variate does not change the expected value of the return.

Exercise 7.7

Write the pseudocode for the off-policy action-value prediction algorithm described immediately above. Pay particular attention to the termination conditions for the recursion upon hitting the horizon or the end of episode.

My answer:

- Input: a behavior policy b and a target policy π
- Algorithm parameters: step size $\alpha \in (0, 1]$, a positive integer n

- Initialize $Q(s,a)$ arbitrarily, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
- All store and access operations (for S_t and R_t) can take their index mod $n+1$
- Loop for each episode:
  - Initialize and store $S_0 \neq \text{terminal}$
  - $T \leftarrow \infty$
  - Loop for $t = 0, 1, 2, \cdots$:
    - If $t < T$, then:
      - Take an action according to $b(\cdot|S_t)$
      - Observe and store the next reward R_{t+1} and the next state S_{t+1}
      - If S_{t+1} is terminal, then $T \leftarrow t+1$
      - $\rho_t = \frac{\pi(S_t|A_t)}{b(S_t|A_t)}$
    - $\tau \leftarrow t - n + 1$ (τ is the time whose state's estimate is being updated)
    - If $\tau \geq 0$:
      - $G \leftarrow R_T$
      - $h \leftarrow T - 1$
      - If $\tau + n < T$, then:
        - $G \leftarrow Q(S_{\tau+n}, A_{\tau+n})$
        - $h \leftarrow \tau + n$
      - Loop for $i = h-1, h-2, \ldots, \tau$:
        - $G \leftarrow R_{i+1} + \gamma\rho_i\big[G - Q(S_{i+1}, A_{i+1})\big] + \gamma\overline{V}(S_{i+1})$
      - $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha\big[G - Q(S_\tau, A_\tau)\big]$
  - Until $\tau = T - 1$

Exercise 7.8

Show that the general (off-policy) version of the n-step return (7.13) can still be written exactly and compactly as the sum of state-based TD errors (6.5) if the approximate state value function does not change.

My answer:

Let $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$, then we have:

$$
\begin{aligned}
G_{t:h} &= \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1-\rho_t)V(S_t) \\
&= \rho_t R_{t+1} + \gamma\rho_t G_{t+1:h} + V(S_t) - \rho_t V(S_t) + \gamma\rho_t V(S_{t+1}) - \gamma\rho_t V(S_{t+1}) \\
&= V(S_t) + \rho_t\big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\big] + \gamma\rho_t\big[G_{t+1:h} - V(S_{t+1})\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_t\big[G_{t+1:h} - V(S_{t+1})\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_t\big[\rho_{t+1}R_{t+2} + \gamma\rho_{t+1}G_{t+2:h} + V(S_{t+1}) - \rho_{t+1}V(S_{t+1}) + \gamma\rho_{t+1}V(S_{t+2}) - \gamma\rho_{t+1}V(S_{t+2}) - V(S_{t+1})\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_{t:t+1}\big[R_{t+2} + \gamma V(S_{t+2}) - V(S_{t+1})\big] + \gamma^2\rho_{t:t+1}\big[G_{t+2:h} - V(S_{t+2})\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_{t:t+1}\delta_{t+1} + \gamma^2\rho_{t:t+1}\big[G_{t+2:h} - V(S_{t+2})\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_{t:t+1}\delta_{t+1} + \cdots + \gamma^{h-t-1}\rho_{t:h-1}\delta_{h-1} + \gamma^{h-t}\rho_{t:h-1}\big[G_{h:h} - V(S_h)\big] \\
&= V(S_t) + \rho_t\delta_t + \gamma\rho_{t:t+1}\delta_{t+1} + \cdots + \gamma^{h-t-1}\rho_{t:h-1}\delta_{h-1} + \gamma^{h-t}\rho_{t:h-1}\big[V(S_h) - V(S_h)\big] \\
&= V(S_t) + \sum_{k=t}^{h-1}\gamma^{k-t}\rho_{t:k}\delta_k
\end{aligned}
$$

Exercise 7.9

Repeat the above exercise for the action version of the off-policy n-step return (7.14) and the Expected Sarsa TD error (the quantity in brackets in Equation 6.9).

My answer:

Let

$$
\begin{aligned}
\delta_t &= R_{t+1} + \gamma\sum_a \pi(a|S_{t+1})Q(S_{t+1}, a) - Q(S_t, A_t) \\
&= R_{t+1} + \gamma\overline{V}(S_{t+1}) - Q(S_t, A_t)
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
G_{t:h} &= R_{t+1} + \gamma\rho_{t+1}\big[G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big] + \gamma\overline{V}(S_{t+1}) \\
&= R_{t+1} + \gamma\rho_{t+1}\big[G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big] + \gamma\overline{V}(S_{t+1}) + Q(S_t, A_t) - Q(S_t, A_t) \\
&= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\big[G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\Big(R_{t+2} + \gamma\rho_{t+2}\big[G_{t+2:h} - Q(S_{t+2}, A_{t+2})\big] + \gamma\overline{V}(S_{t+2}) - Q(S_{t+1}, A_{t+1})\Big) \\
&= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1:t+2}\delta_{t+2} + \gamma^3\rho_{t+1:t+3}\big[G_{t+3:h} - Q(S_{t+3}, A_{t+3})\big]
\end{aligned}
$$

For $t < T$ we have:

$$
\begin{aligned}
G_{t:h} &= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1:t+2}\delta_{t+2} + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}\delta_{h-1} + \gamma^{h-t}\rho_{t+1:h}\big[G_{h:h} - Q(S_h, A_h)\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1:t+2}\delta_{t+2} + \cdots + \gamma^{h-t-1}\rho_{t+1:h-1}\delta_{h-1} + \gamma^{h-t}\rho_{t+1:h}\big[Q(S_h, A_h) - Q(S_h, A_h)\big] \\
&= Q(S_t, A_t) + \delta_t + \sum_{k=t+1}^{h-1}\gamma^{k-t}\rho_{t+1:k}\delta_k
\end{aligned}
$$

For $t \geq T$ we have:

$$
\begin{aligned}
G_{t:h} &= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1:t+2}\delta_{t+2} + \cdots + \gamma^{T-t-2}\rho_{t+1:T-2}\delta_{T-2} + \gamma^{T-t-1}\rho_{t+1:T-1}\big[G_{T-1:h} - Q(S_{T-1}, A_{T-1})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1:t+2}\delta_{t+2} + \cdots + \gamma^{T-t-2}\rho_{t+1:T-2}\delta_{T-2} + \gamma^{T-t-1}\rho_{t+1:T-1}\big[R_T - Q(S_{T-1}, A_{T-1})\big] \\
&= Q(S_t, A_t) + \delta_t + \left(\sum_{k=t+1}^{T-2}\gamma^{k-t}\rho_{t+1:k}\delta_k\right) + \gamma^{T-t-1}\rho_{t+1:T-1}\big[R_T - Q(S_{T-1}, A_{T-1})\big]
\end{aligned}
$$

Exercise 7.10 (programming)

Devise a small off-policy prediction problem and use it to show that the off-policy learning algorithm using (7.13) and (7.2) is more data efficient than the simpler algorithm using (7.1) and (7.9).

My answer:

TODO

Exercise 7.11

Show that if the approximate action values are unchanging, then the tree-backup return (7.16) can be written as a sum of expectation-based TD errors:

$$
G_{t:t+n} = Q(S_t, A_t) + \sum_{k=t}^{min(t+n,T)-1}\delta_k\prod_{i=t+1}^{k}\gamma\pi(A_i|S_i)
$$

where $\delta_t = R_{t+1} + \gamma\overline{V}_t(S_{t+1}) - Q(S_t, A_t)$ and $\overline{V}_t$ is given by (7.8).

My answer:

Let $h = min(t+n, T) - 1$, then we have:

$$
\begin{aligned}
G_{t:t+n} &= R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q(S_{t+1}, a) + \gamma \pi(A_{t+1}|S_{t+1}) G_{t+1:h} \\
&= R_{t+1} + \gamma \sum_{a} \pi(a|S_{t+1}) Q(S_{t+1}, a) - \gamma \pi(A_{t+1}|S_{t+1}) Q(S_{t+1}, A_{t+1}) + \gamma \pi(A_{t+1}|S_{t+1}) G_{t+1:h} + Q(S_t, A_t) - Q(S_t, A_t) \\
&= Q(S_t, A_t) + \big[R_{t+1} + \gamma \overline{V}(S_{t+1}) - Q(S_t, A_t)\big] + \gamma \pi(A_{t+1}|S_{t+1}) \big[G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \big[G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \big[R_{t+2} + \gamma \sum_{a} \pi(a|S_{t+2}) Q(S_{t+2}, a) - \\
&\quad - \gamma \pi(A_{t+2}|S_{t+2}) Q(S_{t+2}, A_{t+2}) + \gamma \pi(A_{t+2}|S_{t+2}) G_{t+2:h} - Q(S_{t+1}, A_{t+1})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \delta_{t+1} + \gamma^2 \pi(A_{t+1}|S_{t+1}) \pi(A_{t+2}|S_{t+2}) \big[G_{t+2:h} - Q(S_{t+2}, A_{t+2})\big] \\
&= Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \delta_{t+1} + \cdots + \delta_{h-1} \prod_{i=t+1}^{h-1} \gamma \pi(A_i|S_i) + \big[G_{h:h} - Q(S_h, A_h)\big] \prod_{k=t+1}^{h} \gamma \pi(A_i|S_i) \\
&= Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \delta_{t+1} + \cdots + \delta_{h-1} \prod_{i=t+1}^{h-1} \gamma \pi(A_i|S_i) \\
&= Q(S_t, A_t) + \sum_{k=t}^{h-1} \delta_k \prod_{i=t+1}^{k} \gamma \pi(A_i|S_i) \\
&= Q(S_t, A_t) + \sum_{k=t}^{\min(t+n-1, T-1)} \delta_k \prod_{i=t+1}^{k} \gamma \pi(A_i|S_i)
\end{aligned}
$$