Solutions to Reinforcement Learning by Sutton
Chapter 7

Yifan Wang

Aug 2019

Exercise 7.1

Recall, $\delta_t \doteq R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)$**.**
Update of n-step TD is $V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha[G_{t:t+n} - V_{t+n-1}(S_t)]$**.**
n-step errors used in (7.2) is $G_{t:t+n} - V_{t+n-1}(S_t)$**.**
Derivation is as follows:

$$
\begin{aligned}
G_{t:t+n} - V_{t+n-1}(S_t) &= R_{t+1} + \gamma R_{t+2} + ... + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t) \\
&= R_{t+1} + \gamma R_{t+2} + ... + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n}) - V(S_t) \\
&\quad \text{(because we assume V will not change)} \\
&= \delta_t - \gamma V(S_{t+1}) + V(S_t) + \gamma R_{t+2} + ... + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n}) - V(S_t) \\
&= \gamma^0[\delta_t - \gamma V(S_{t+1}) + V(S_t)] + \gamma[\delta_{t+1} - \gamma V(S_{t+2}) + V(S_{t+1})] + \\
&... + \gamma^{n-1}[\delta_{t+n-1} - \gamma V(S_{t+n}) + V(S_{t+n-1})] + \gamma^n V(S_{t+n}) - V(S_t) \\
&= \sum_{k=t}^{t+n-1} \left[\gamma^{k-t}\delta_k - \gamma^{k-t+1} V(S_{k+1}) + \gamma^{k-t} V(S_k)\right] + \gamma^n V(S_{t+n}) - V(S_t) \\
&= \sum_{k=t}^{t+n-1} \left[\gamma^{k-t}\delta_k\right] - \sum_{k=t}^{t+n-2} \left[\gamma^{k-t+1} V(S_{k+1})\right] + \sum_{k=t+1}^{t+n-1} \left[\gamma^{k-t} V(S_k)\right] \\
&= \sum_{k=t}^{t+n-1} \left[\gamma^{k-t}\delta_k\right] - \sum_{k=t+1}^{t+n-1} \left[\gamma^{k-t} V(S_k)\right] + \sum_{k=t+1}^{t+n-1} \left[\gamma^{k-t} V(S_k)\right] \\
&= \sum_{k=t}^{t+n-1} \gamma^{k-t}\delta_k
\end{aligned}
$$

■

Exercise 7.2

Programming problem will be finished later with collaboration.

■

Exercise 7.3

Because we are using n-steps and we want to avoid early finish for our random walk. If we have a small problem, the expected value of number of states of travelling will be smaller than n, making the performance test less fair and accurate for large n. If we have a very small problem, I do believe smaller n should be better. And while the problem is very big, like 100 states random walk, I believe the best value of n will be much bigger, too.

■

Exercise 7.4

Recall 7.4 of Sarsa is the following:

$G_{t:t+n} \doteq R_{t+1}+\gamma R_{t+2}+...+\gamma^{n-1}R_{t+n}+\gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$, $n \geq 1, 0 \leq t < T-n$

With little bit manipulation, we have the following:

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

The cancellation always happens between the second Q term for each R_k and the fourth Q term for the next R_{k+1}. As special case of first R

and last R: the underlined one must be taken out since R_{t-1} does not exist. The cancellation of the last term $\gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$ is with the second Q term of R_{t+n}. And this final result completes the proof. ■

Exercise 7.5

Replace the original update process of the one in page 149 with (7.13), and change Q to V with appropriate modification of memory. It is tedious to write out and the complete algorithm will appear in later chapters officially. The importance of 7.13 is not for the application only. As the book will talk later about the unified Q method, it provides a toolbox to unify the importance sampling and the n-step tree. ■

Exercise 7.6

Recall the off-policy control variate is the following:

$$\begin{aligned} G_{t:h} &\doteq R_{t+1} + \gamma\big(\rho_{t+1}G_{t+1:h} + \bar{V}_{h-1}(S_{t+1}) - \rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})\big) \\ &= R_{t+1} + \gamma\rho_{t+1}\big(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}), \quad t < h \leq T \end{aligned}$$

Recall the natural algorithm of the n-step Sarsa is:

$$Q_{t+n}(S_t, A_t) \doteq Q_{t+n-1}(S_t, A_t) + \alpha[G_{t:t+n} - Q_{t+n-1}(S_t, A_t)], \quad 0 \leq t < T$$

Let's try to combine both.

$$\begin{aligned} \mathbb{E}\Big[G_{t:h}\Big] &= \mathbb{E}\Big[R_{t+1} + \gamma\rho_{t+1}\big(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1})\Big] \\ &= \mathbb{E}\Big[R_{t+1}\Big] + \mathbb{E}\Big[\gamma\rho_{t+1}\big(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\big)\Big] + \mathbb{E}\Big[\gamma\bar{V}_{h-1}(S_{t+1})\Big] \\ &\quad \textbf{Because } \rho \textbf{ is unrelated with any estimate and } \mathbb{E}[\rho] = 1 \\ &= R_{t+1} + \mathbb{E}\Big[\gamma G_{t+1:h} - \gamma Q_{h-1}(S_{t+1}, A_{t+1})\Big] + \gamma\bar{V}_{h-1}(S_{t+1}) \\ &= R_{t+1} + \mathbb{E}\Big[G_{t:h} - R_{t+1} - \gamma Q_{h-1}(S_{t+1}, A_{t+1})\Big] \\ &= \mathbb{E}\Big[G_{t:h}\Big] + \gamma\Big[\mathbb{E}\big[-Q_{h-1}(S_{t+1}, A_{t+1})\big] + \bar{V}_{h-1}(S_{t+1})\Big] \\ &\quad \textbf{By Definition of } \bar{V} \textbf{ (7.8)} \\ &= \mathbb{E}\Big[G_{t:h}\Big] \end{aligned}$$

■

Exercise 7.7

Treat the termination at the horizon as $G_{h:h} \doteq \bar{V}_{h-1}(S_h)$ and treat the non-terminated condition as $G_{h:h} \doteq Q_{h-1}(S_h, A_h)$.
Specific pseudo-code will be skipped. This answer may be updated later.

■

Exercise 7.8

Recall the off-policy version of the n-step return (7.13) is:

$$G_{t:h} \doteq \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1-\rho_t)V_{h-1}(S_t), \quad t < h < T$$

and, $\delta_t \doteq R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)$**.**
Let's try to expand (7.13).

$$\begin{aligned}
G_{t:h} &\doteq \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1-\rho_t)V(S_t) \\
&= \rho_t(R_{t+1} + \gamma\big[\rho_{t+1}(R_{t+2} + \gamma G_{t+2:h}) + (1-\rho_{t+1})V(S_{t+1})\big]) + (1-\rho_t)V(S_t) \\
&= \rho_t(R_{t+1} + \underline{\gamma V(S_{t+1})} - \underline{V(S_t)} + \gamma\big[\rho_{t+1}(R_{t+2} + \gamma G_{t+2:h}) + (\underline{0} - \rho_{t+1})V(S_{t+1})\big]) + (1-\underline{0})V(S_t) \\
&= \rho_t(\delta_t + \gamma\big[\rho_{t+1}(R_{t+2} + \gamma G_{t+2:h}) - \rho_{t+1}V(S_{t+1})\big]) + V(S_t) \\
&= \rho_t\Bigg(\delta_t + \gamma\Bigg[\rho_{t+1}\Big(R_{t+2} + \gamma\Big[\rho_{t+2}\big(\delta_{t+2} + \gamma\big[\rho_{t+3}(R_{t+4} + \gamma G_{t+4:h}) - \rho_{t+3}V(S_{t+3})\big]\big) \\
&\quad + V(S_{t+2})\Big]\Big) - \rho_{t+1}V(S_{t+1})\Bigg]\Bigg) + V(S_t) \\
&= \rho_t\Bigg(\delta_t + \gamma\Bigg[\rho_{t+1}\Big(R_{t+2} + \gamma V(S_{t+2}) - V(S_{t+1}) + \gamma\Big[\rho_{t+2}\big(\delta_{t+2} + \gamma\big[\rho_{t+3}(R_{t+4} + \gamma G_{t+4:h}) \\
&\quad - \rho_{t+3}V(S_{t+3})\big]\big) + 0\Big]\Big) - 0\Bigg]\Bigg) + V(S_t) \\
&= \rho_t\Bigg(\delta_t + \gamma\Bigg[\rho_{t+1}\Big(\delta_{t+1} + \gamma\Big[\rho_{t+2}\big(\delta_{t+2} + \gamma\big[\rho_{t+3}(R_{t+4} + \gamma G_{t+4:h}) - \rho_{t+3}V(S_{t+3})\big]\big)\Big]\Big)\Bigg]\Bigg) + V(S_t) \\
&= \sum_{k=t}^{min(T-t,h)} \left(\prod_{l=t}^{l=k} \rho_l \gamma^{k-t}\right)\delta_t + V(S_t)
\end{aligned}$$

■

Exercise 7.9

Recall off-policy n-step return (7.14) is the following:

$$\begin{aligned}
G_{t:h} &\doteq R_{t+1} + \gamma\big(\rho_{t+1}G_{t+1:h} + \bar{V}_{h-1}(S_{t+1}) - \rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})\big) \\
&= R_{t+1} + \gamma\rho_{t+1}\big(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}), \quad t < h \leq T
\end{aligned}$$

Recall Expected Sarsa TD error is the following:

$$\delta_t = R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)$$

Let's do this:

$$\begin{aligned}
G_{t:h} &= R_{t+1} + \gamma\rho_{t+1}\big(G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}) \\
&= \delta_t - \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) + Q(S_t, A_t) + \gamma\rho_{t+1}\big(G_{t+1:h} - Q(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}) \\
&= \delta_t - \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) + Q(S_t, A_t) + \gamma\rho_{t+1}\big(- Q(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}) \\
&\quad + \gamma\rho_{t+1} G_{t+1:h} \\
&= \delta_t - \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) + Q(S_t, A_t) + \gamma\rho_{t+1}\big(- Q(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}) \\
&\quad + \gamma\rho_{t+1}\Big[R_{t+2} + \gamma\rho_{t+2}\big(G_{t+2:h} - Q(S_{t+2}, A_{t+2})\big) + \gamma\bar{V}_{h-1}(S_{t+2})\Big] \\
&= \delta_t - \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) + Q(S_t, A_t) + \gamma\rho_{t+1}\big(- Q(S_{t+1}, A_{t+1})\big) + \gamma\bar{V}_{h-1}(S_{t+1}) \\
&\quad + \gamma\rho_{t+1}\Big[\delta_{t+1} - \gamma \sum_a \pi(a|S_{t+2}) Q(S_{t+2}, a) + Q(S_{t+1}, A_{t+1}) + \gamma\rho_{t+2}\big(G_{t+2:h} - Q(S_{t+2}, A_{t+2})\big) \\
&\quad + \gamma\bar{V}_{h-1}(S_{t+2})\Big] \\
&= \sum_{k=t}^{h-1} \Big[\gamma^{k-t}\big[\prod_{i=t+1}^{k} \rho_i\big]\big[\delta_k - \gamma \sum_a \pi(a|S_{k+1}) Q(S_{k+1}, a) + Q(S_k, A_k) - \gamma\rho_{k+1} Q(S_{k+1}, A_{k+1}) \\
&\quad + \gamma\bar{V}_{h-1}(S_{k+1})\big]\Big] \\
&= \text{Remain for help here.}
\end{aligned}$$

■

Exercise 7.10

Skip the programming part until cooperation.