Solutions to Reinforcement Learning by Sutton Chapter 12

Yifan Wang

Jan 2020

Exercise 12.1 is too long to be contained in page 1, and is put in the next page. And this is one of the most important exercise in this Chapter as it supports 12.3 and 12.4.

Exercise 12.1

$$
\begin{aligned}
G_t^\lambda &= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t:t+n} \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big[\sum_{k=1}^{n}\gamma^{k-1}R_{t+k} + \gamma^n\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big[R_{t+1} + \sum_{k=2}^{n}\gamma^{k-1}R_{t+k} + \gamma^n\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\,R_{t+1} + (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big[\sum_{k=2}^{n}\gamma^{k-1}R_{t+k} + \gamma^n\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\,R_{t+1} + (1-\lambda)\sum_{n=1}^{1}\lambda^{n-1}\Big[\sum_{k=2}^{n}\gamma^{k-1}R_{t+k} + \gamma^n\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&\quad + \quad \gamma\,(1-\lambda)\sum_{n=2}^{\infty}\lambda^{n-1}\Big[\sum_{k=2}^{n}\gamma^{k-2}R_{t+k} + \gamma^{n-1}\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\,R_{t+1} + (1-\lambda)\sum_{n=1}^{1}\lambda^{n-1}\Big[\sum_{k=2}^{n}\gamma^{k-1}R_{t+k} + \gamma^n\hat{v}(S_{t+n},\mathrm{w}_{t+n-1})\Big] \\
&\quad + \quad \gamma\,(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big[\sum_{k=2}^{n}\gamma^{k-2}R_{t+k} + \gamma^n\hat{v}(S_{t+n+1},\mathrm{w}_{t+n}) \\
&\quad - \quad \gamma^n\hat{v}(S_{t+n+1},\mathrm{w}_{t+n}) + \gamma^{n-1}\hat{v}(S_{t+n},\mathrm{w}_{t+n-1}) + \gamma^{n-1}R_{t+n+1} - \gamma^{n-1}R_{t+n+1}\Big] - \gamma(1-\lambda)\hat{v}(S_{t+1},\mathrm{w}_t) \\
&= R_{t+1} + \gamma G_{t+1}^\lambda \\
&\quad + \quad \gamma(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\big[-\gamma^n\hat{v}(S_{t+n+1},\mathrm{w}_{t+n}) + \gamma^{n-1}\hat{v}(S_{t+n},\mathrm{w}_{t+n-1}) - \gamma^{n-1}R_{t+n+1}\big]
\end{aligned}
$$

If you expand the last term you will cancel out a lot of terms and remain the following:

$$
\begin{aligned}
&= R_{t+1} + \gamma G_{t+1}^\lambda + \gamma(1-\lambda)(-G_{t+1}^\lambda) + \gamma(1-\lambda)\hat{v}(S_{t+1},\mathrm{w}_t) \\
&= R_{t+1} + \lambda\gamma G_{t+1}^\lambda + \gamma(1-\lambda)\hat{v}(S_{t+1},\mathrm{w}_t)
\end{aligned}
$$

■

Exercise 12.2

By definition:

$$
\begin{aligned}
(1-\lambda)\lambda^{\tau_\lambda} &= \frac{1}{2}(1-\lambda) \\
\lambda^{\tau_\lambda} &= \frac{1}{2} \\
\tau_\lambda &= \log_\lambda \frac{1}{2}
\end{aligned}
$$

■

Exercise 12.3

Like (6.6), based on a fixed w we form equations as following:

$$
\begin{aligned}
G_t^\lambda - \hat{v}(S_t, \mathrm{w}) &= R_{t+1} + \gamma\lambda G_{t+1}^\lambda - \hat{v}(S_t, \mathrm{w}) + \gamma(1-\lambda)\hat{v}(S_{t+1}, \mathrm{w}) \\
&= R_{t+1} + \gamma\lambda G_{t+1}^\lambda - \hat{v}(S_t, \mathrm{w}) + \gamma\hat{v}(S_{t+1}, \mathrm{w}) - \gamma\hat{v}(S_{t+1}, \mathrm{w}) + \gamma(1-\lambda)\hat{v}(S_{t+1}, \mathrm{w}) \\
&= \delta_t + \gamma\lambda G_{t+1}^\lambda - \gamma\hat{v}(S_{t+1}, \mathrm{w}) + \gamma(1-\lambda)\hat{v}(S_{t+1}, \mathrm{w}) \\
&= \delta_t + \gamma\lambda\big(G_{t+1}^\lambda - \hat{v}(S_{t+1}, \mathrm{w})\big) - \gamma(1-\lambda)\hat{v}(S_{t+1}, \mathrm{w}) + \gamma(1-\lambda)\hat{v}(S_{t+1}, \mathrm{w}) \\
&= \delta_t + \gamma\lambda\Big[\delta_{t+1} + \gamma\lambda\big(G_{t+2}^\lambda - \hat{v}(S_{t+2}, \mathrm{w})\big)\Big] \\
&= \sum_{k=t}^{\infty} \gamma^{k-t}\lambda^{k-t}\delta_k
\end{aligned}
$$

■

Exercise 12.4

$$\sum_{t=0}^{\infty} \alpha \left[(R_t + \gamma \hat{v}(S_{t+1}, \mathrm{w}) - \hat{v}(S_t, w) \right] \mathrm{z}_t \qquad \text{(from 12.7, sum of TD}(\lambda)\text{ updates)}$$

$$= \sum_{t=0}^{\infty} \alpha\, \delta_t \left[\gamma \lambda \mathrm{z}_{t-1} + \nabla \hat{v}(S_t, \mathrm{w}) \right]$$

$$= \sum_{t=0}^{\infty} \alpha\, \delta_t \Big[\gamma \lambda \big(\gamma \lambda z_{t-2} + \nabla \hat{v}(S_{t-1}, \mathrm{w}) \big) + \nabla \hat{v}(S_t, \mathrm{w}) \Big]$$

$$= \sum_{t=0}^{\infty} \alpha\, \delta_t \Big[\gamma^2 \lambda^2 z_{t-2} + \gamma \lambda \nabla \hat{v}(S_{t-1}, \mathrm{w}) + \nabla \hat{v}(S_t, \mathrm{w}) \Big]$$

$$= \sum_{t=0}^{\infty} \alpha\, \delta_t \Big[\sum_{k=0}^{t} \gamma^{t-k} \lambda^{t-k} \nabla \hat{v}(S_k, \mathrm{w}) \Big]$$

Consider for a given t, any $\nabla \hat{v}(S_k, \mathrm{w})$ will only have one occurrence of index in $\sum_{k=0}^{t} \gamma^{t-k} \lambda^{t-k}$ at each $t \geq k$.

Now we collect all those indices across different t's for unique state S_t:

$$= \sum_{t=0}^{\infty} \alpha\, \delta_t \Big[\sum_{k=t}^{\infty} \gamma^{k-t} \lambda^{k-t} \Big] \nabla \hat{v}(S_t, \mathrm{w})$$

Similarly, each δ_t will have indices from all $k \geq t$

$$= \sum_{t=0}^{\infty} \alpha \Big[\sum_{k=t}^{\infty} \gamma^{k-t} \lambda^{k-t} \delta_k \Big] \nabla \hat{v}(S_t, \mathrm{w})$$

$$= \sum_{t=0}^{\infty} \alpha \Big[G_t^{\lambda} - \hat{v}(S_t, \mathrm{w}) \Big] \nabla \hat{v}(S_t, \mathrm{w})$$

(By *exercise 12.3*, and sum of λ-return updates)

■

Exercise 12.5

$$
\begin{aligned}
G^{\lambda}_{t:t+k} &= \sum_{i=1}^{k-1} \lambda^{i-1} G_{t:t+i} - \sum_{i=1}^{k-1} \lambda^{i} G_{t:t+i} + \lambda^{k-1} G_{t:t+k} \qquad (12.9) \\
&= \sum_{i=0}^{k-2} \lambda^{i} G_{t:t+i+1} - \sum_{i=1}^{k-1} \lambda^{i} G_{t:t+i} + \lambda^{k-1} G_{t:t+k} \\
&= G_{t:t+1} - \lambda^{k-1} G_{t:t+k-1} + \sum_{i=1}^{k-2} \lambda^{i} G_{t:t+i+1} - \sum_{i=1}^{k-2} \lambda^{i} G_{t:t+i} + \lambda^{k-1} G_{t:t+k} \\
&= R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathrm{w}_t) + \sum_{i=1}^{k-1} \lambda^{i} \big[G_{t:t+i+1} - G_{t:t+i} \big] \\
&= R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathrm{w}_t) + \sum_{i=t+1}^{t+k-1} (\gamma\lambda)^{i-t} \big[R_{i+1} + \gamma \hat{v}(S_{i+1}, \mathrm{w}_i) - \hat{v}(S_i, \mathrm{w}_{i-1}) \big] \\
&= \hat{v}(S_t, \mathrm{w}_{t-1}) + \sum_{i=t}^{t+k-1} (\gamma\lambda)^{i-t} \big[R_{i+1} + \gamma \hat{v}(S_{i+1}, \mathrm{w}_i) - \hat{v}(S_i, \mathrm{w}_{i-1}) \big] \\
&= \hat{v}(S_t, \mathrm{w}_{t-1}) + \sum_{i=t}^{t+k-1} (\gamma\lambda)^{i-t} \delta'_i
\end{aligned}
$$

■

Exercise 12.6

Replace the loop in $\mathcal{F}(S, A)$ with following:

Loop for i **in** $\mathcal{F}(\mathcal{S}, \mathcal{A})$**:**
| $s \leftarrow s + z_i$
Loop for i **in** $\mathcal{F}(\mathcal{S}, \mathcal{A})$**:**
| $\delta \leftarrow \delta - w_i$
| $z_i \leftarrow \gamma\lambda z_i + 1 - \alpha\gamma\lambda\, s$

Delete the z part in last loop. ■

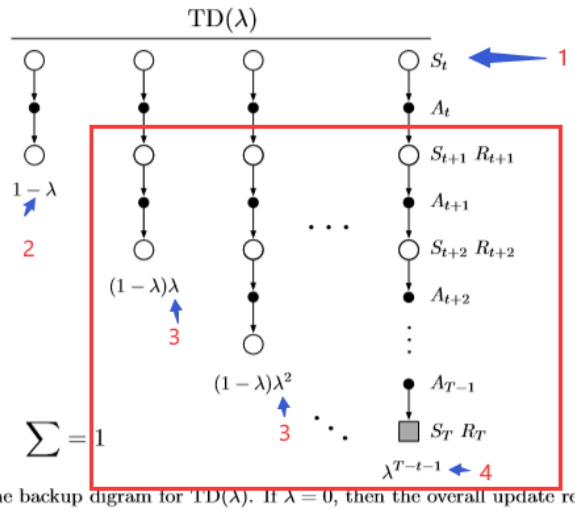

Figure 12.1: The backup digram for TD(λ). If $\lambda = 0$, then the overall update reduces to its first component, the one-step TD update, whereas if $\lambda = 1$, then the overall update reduces to its last component, the Monte Carlo update.

Figure 1: Exercise 12.7

Exercise 12.7

The book does not explain well about the new definition. We need to understand how (12.18) comes from. It is actually not a new definition but rather derived from former definition of λ return.

To illustrate, one should take a look at figure 12.1. I made a copy with annotation to illustrate points. See Figure 1. From this figure, our goal is to figure out how we obtain G_t^λ from G_{t+1}^λ. From now on, please take attention to the small number I have written in the figure.

Point 0. There is no 0 on figure. Just make sure to have a correct picture of G_{t+1}^λ in your mind. Keep in mind which variables should be changed.

Point 1. We need to delete or ignore the R_{t+1} which is unfortunately in the box I have drawn for G_{t+1}^λ. Indeed, If we have G_{t+1}^λ(small box) and we want G_t^λ(whole picture) we need to add a row of R_{t+1} on the top during our calculation. Besides, we need to multiply γ, or γ_{t+1} in

our new variable settings to prepare the discounting.

Point 2. We need to add additional column for $G_{t:t+1}$ which is clearly outside of the box. And multiplied by $(1-\lambda)$

Point 3. Of course remaining columns from G^{λ}_{t+1} should be multiplied with a new λ. Here and in steps above, λ should be written as λ_{t+1} for our purpose.

Point 4. Recall there is a tailing item. What is it? It is $\lambda^{T-t-2}G_{t+1}$ if we write it out in the formula of G^{λ}_{t+1}. What is G_{t+1}? It is Monte Carlo Final Return, a sum of bunches of rewards. Namely, $\sum_{n=t+1}^{T}\gamma^{n-t-1}R_n$.We do not need this term, we need instead a G_t. So what's the difference between G_t and G_{t+1}? We need add a R_{t+1} and multiplied additional γ to all other terms. Because we have multiplied γ in step 1, we only need to take out the R_{t+1} term. However, this term will also be affected by the λ. Thus we need to add λR_{t+1} to all columns to cancel out effect in the tailing term.

Final step, lots of λ and γ above, change it to corresponding form. Let's do it again in Math.

$$
\begin{aligned}
G^{\lambda s}_{t} &= (1-\lambda_{t+1})G_t + \gamma_{t+1}\lambda_{t+1}G^{\lambda s}_{t+1} + \lambda_{t+1}R_{t+1} && \text{(By all points above)} \\
&= (1-\lambda_{t+1})\big[R_{t+1} + \gamma_{t+1}\,\hat{v}(S_{t+1}, w_t)\big] + \gamma_{t+1}\lambda_{t+1}G^{\lambda s}_{t+1} + \lambda_{t+1}R_{t+1} \\
& && \text{(write out bootstrap formula of } G_t\text{)} \\
&= R_{t+1} + \gamma_{t+1}\big((1-\lambda_{t+1})\hat{v}(S_{t+1}, w_t) + \lambda_{t+1}G^{\lambda s}_{t+1}\big) && (12.18)
\end{aligned}
$$

Now let's look at our problem. Author wants us to find similar equations in truncated TD. Looking at Figure 12.7 in the book, it is very familiar with our manipulation above. When $t < h$, there will be some slight difference. But one can easily see that from $G_{t+1:h}$ to $G_{t:h}$, the method we need to apply is exactly same. We need to apply discounting and add a new reward. Thus, I claim the redefinition should be in exactly same form as follows.

$$G^{\lambda s}_{t:h} \doteq R_{t+1} + \gamma_{t+1}\Big((1-\lambda_{t+1})\hat{v}(S_{t+1}, \mathrm{w}_t) + \lambda_{t+1} G^{\lambda s}_{t+1:h}\Big),$$
$$G^{\lambda a}_{t:h} \doteq R_{t+1} + \gamma_{t+1}\Big((1-\lambda_{t+1})\hat{q}(S_{t+1}, A_{t+1}, \mathrm{w}_t) + \lambda_{t+1} G^{\lambda s}_{t+1:h}\Big),$$
$$G^{\lambda s}_{t:h} \doteq R_{t+1} + \gamma_{t+1}\Big((1-\lambda_{t+1})\bar{V}_t(S_{t+1}) + \lambda_{t+1} G^{\lambda s}_{t+1:h}\Big).$$

■

Exercise 12.8

This type of questions have been appearing many times. You can write out and cancelling out or put it back to the formula and write out the recursion. Let me do the later type.

$$G^{\lambda s}_t \doteq \rho_t\Big(R_{t+1} + \gamma_{t+1}\big((1-\lambda_{t+1})V_{t+1} + \lambda_{t+1}G^{\lambda s}_{t+1}\big)\Big) + (1-\rho_t)V_t$$

(12.22 rewritten)

$$= \rho_t\Big(R_{t+1} + \gamma_{t+1}V_{t+1} - V_t + \gamma_{t+1}\big(-\lambda_{t+1}V_{t+1} + \lambda_{t+1}G^{\lambda s}_{t+1}\big)\Big) + V_t$$
$$= \rho_t\Big(\delta^s_t + \gamma_{t+1}\big(\lambda_{t+1}(G^{\lambda s}_{t+1} - V_{t+1})\big)\Big) + V_t$$
$$= \rho_t\Bigg(\delta^s_t + \gamma_{t+1}\Big(\lambda_{t+1}\big(\rho_{t+1}\big(R_{t+2} + \gamma_{t+2}\big((1-\lambda_{t+2})V_{t+2} + \lambda_{t+2}G^{\lambda s}_{t+2}\big)\big) + (1-\rho_{t+1})V_{t+1} - V_{t+1}\big)\Big)\Bigg) + V_t$$
$$= \rho_t\Bigg(\delta^s_t + \gamma_{t+1}\Big(\lambda_{t+1}\big(\rho_{t+1}\big(R_{t+2} + \gamma_{t+2}V_{t+2} - V_{t+1}\gamma_{t+2}\big(\lambda_{t+2}(G^{\lambda s}_{t+2} - V_{t+2})\big)\big)\big)\Big)\Bigg) + V_t$$
$$= \rho_t\Bigg(\delta^s_t + \gamma_{t+1}\Big(\lambda_{t+1}\big(\rho_{t+1}\big(\delta^s_{t+1} + \gamma_{t+2}\big(\lambda_{t+2}(G^{\lambda s}_{t+2} - V_{t+2})\big)\big)\big)\Big)\Bigg) + V_t$$
$$= \ldots \qquad \text{(Repeat until } \lim_{n\to\infty}(G^{\lambda s}_n - V_n) = 0 - 0 = 0\text{)}$$
$$= \rho_t \sum_{k=t}^{\infty} \delta^s_k \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i + V_t \qquad (12.24)$$

■

Exercise 12.9

'Guess', Eh, interesting.
OK, fine, I just guess that is the following. Convince yourself.
(See final case in the previous question and replace it with h, open an issue if you still don't believe the answer.)

$$G_{t:h}^{\lambda s} \approx \hat{v}(S_t, \mathrm{w}_t) + \rho_t \sum_{k=t}^{h} \delta_k^s \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$$

■

Exercise 12.10

I will use hint and the simplest version to solve it not like the one in 12.8.

$$\begin{aligned} \delta_0^a &= R_1 + \gamma_1 \bar{V}(S_1) - Q_0 \\ G_0^{\lambda a} &= R_1 + \gamma_1 (\bar{V}(S_1) + \lambda_1 \rho_1 [G_1^{\lambda a} - Q_1] \end{aligned}$$

$$\begin{aligned} G_0^{\lambda a} - \delta_0^a &= R_1 + \gamma_1 (\bar{V}(S_1) + \lambda_1 \rho_1 [G_1^{\lambda a} - Q_1] - \left[R_1 + \gamma_1 \bar{V}(S_1) - Q_0\right] \\ &= \lambda_1 \rho_1 [G_1^{\lambda a} - Q_1] + Q_0 \\ &= \dots \qquad \text{(Repeat writing out } G_n^{\lambda a} - Q_n \text{ for all } n) \\ &= Q_t + \sum_{t}^{\infty} \delta_k^a \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i \end{aligned}$$

■

Exercise 12.11

Another Guess.

$$G_{t:h}^{\lambda a} \approx \hat{q}(S_t, A_t, \mathrm{w}_t) + \sum_{k=t}^{h} \delta_k^a \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$$

■

Exercise 12.12

Let's follow the hints.

$$\mathrm{w}_{t+1} \doteq \mathrm{w}_t + \alpha\big[G_t^{\lambda} - \hat{q}(S_t, A_t, \mathrm{w}_t)\big]\nabla\hat{q}(S_t, A_t, \mathrm{w}_t) \tag{12.15}$$

$$\begin{aligned}\mathrm{w}_{t+1} &\doteq \mathrm{w}_t + \alpha\big[G_t^{\lambda a} - \hat{q}(S_t, A_t, \mathrm{w}_t)\big]\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\\ &\approx \mathrm{w}_t + \alpha\big[\sum_{k=t}^{\infty}\delta_k^a \prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i\big]\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\end{aligned}$$

(replace from 12.27)

$$\begin{aligned}\sum_{t=1}^{\infty}(\mathrm{w}_{t+1} - \mathrm{w}_t) &\approx \sum_{t=1}^{\infty}\sum_{k=t}^{\infty}\alpha\delta_k^a\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i\\ &= \sum_{k=1}^{\infty}\sum_{t=1}^{k}\alpha\delta_k^a\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i \quad \text{(summation trick)}\\ &= \sum_{k=1}^{\infty}\alpha\delta_k^a\sum_{t=1}^{k}\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i\end{aligned}$$

$$\begin{aligned}\mathrm{z}_k &= \sum_{t=1}^{k}\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i\\ &= \sum_{t=1}^{k-1}\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k}\gamma_i\lambda_i\rho_i + \nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\\ &\qquad\qquad \text{(product disappeared due to } k+1 > k\text{)}\\ &= \gamma_k\lambda_k\rho_k\sum_{t=1}^{k-1}\nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\prod_{i=t+1}^{k-1}\gamma_i\lambda_i\rho_i + \nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\\ &= \gamma_k\lambda_k\rho_k\mathrm{z}_{k-1} + \nabla\hat{q}(S_t, A_t, \mathrm{w}_t)\\ &= \gamma_t\lambda_t\rho_t\mathrm{z}_{t-1} + \nabla\hat{q}(S_t, A_t, \mathrm{w}_t) \qquad \text{(Changing } k \text{ to } t\text{)}\end{aligned}$$

■

Exercise 12.13

Not yet complete. Please share your opinions by opening new issues.

Exercise 12.14

Not yet complete. Please share your opinions by opening new issues.