1 n-step TD Prediction

1.1 Exercise 7.1

Q

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 TD errors (again if the value estimates don't change) generalizing the earlier result.

A

Write

$$G_{t:t+n} = \sum_{i=1}^{n} \gamma^{i-1} R_{t-i} + \gamma^n V(S_{t+n}),$$

then have the TD error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t).$$

So the n-step error can be written as

$$\begin{aligned} G_{t:t+n} - V(S_t) &= R_{t+1} + \gamma \sum_{i=1}^{n-1} \gamma^{i-1} R_{t+1+i} + \gamma^n V(S_{t+n}) - V(S_t) \\ &= \delta_t + \gamma \left(G_{t+1:t+n} - V(S_{t+1}) \right) \\ &\vdots \\ &= \sum_{k=t}^{t+n-1} \gamma^{k-t} \delta_k \end{aligned}$$

1.2 Exercise 7.2 (programming)

Q

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.

A

...

1.3 Exercise 7.3

Q

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?

A

- Smaller walk would have shifted advantage to smaller n because when $n \geq \frac{\#\text{states}-1}{2}$ (#states is odd) the algorithm updates all states visited by the terminal reward. This means that the algorithm only makes value changes of size α, since the values are no longer bootstrapped or backed up.
- The addition of the -1 reward on the left favours smaller values of n, because in longer episodes the larger values of n will have to update many states by the terminal reward (now -1 rather than 0) thus increasing variance

1.4 Exercise 7.4

Q

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

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

A

Denote

$$G_{t:t+n} \doteq \sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

for $n \geq 1$ and $0 \leq t < T - n$ and with $G_{t:t+n} = G_t$ if $t + n > T$.

Set $\tau = \min(t+n, T) - 1$ and observe that

$$\begin{aligned}
&\sum_{k=t}^{\tau} \gamma^{k-t}[R_{t+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)] \\
&= \sum_{k=t}^{\tau} \gamma^{k-t} R_{t+1} + \gamma \sum_{k=t}^{\tau} \gamma^{k-t} Q_k(S_{k+1}, A_{k+1}) - \sum_{k=t}^{\tau} \gamma^{k-t} Q_{k-1}(S_k, A_k) \\
&= G_{t:t+n} - \mathbb{1}\{t+n < T\}\gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) + \gamma^\tau Q_\tau(S_\tau, A_\tau) - Q_{t-1(S_t, A_t)} \\
&= G_{t:t+n} - Q_{t-1}(S_t, A_t).
\end{aligned}$$

1.5 Exercise 7.5

Q

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

A

The update that we use is the same as the n-step TD update, only multiplied by the importance sampling ratio and with the control variate added.

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

The algorithm is therefore the same, but with these steps replacing the old returns calculations, using the latest available value function.

If you take the recursion relation literally then we should get a control variate for each of the intermediary states $(t+1, \ldots, t+n)$, but I don't think that this is what is intended as this would just increase variance further.

1.6 Exercise 7.6

Q

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

A

In (7.13) we have

$$\begin{aligned}\mathbb{E}[(1-\rho_t)V_{h-1}(S_t)] &= \mathbb{E}_b[(1-\rho_t)V_{h-1}(S_t)]\\ &= \mathbb{E}_b[(1-\rho_t)]\mathbb{E}_b[V_{h-1}(S_t)]\\ &= 0.\end{aligned}$$

In the second case (7.14) we have

$$\begin{aligned}&\mathbb{E}_b[\bar{V}_{h-1}(S_{t+1}) - \rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})|S_{t+1}]\\ &= \sum_a \pi(a|S_{t+1})Q_{h-1}(S_{t+1}, a) - \sum_a b(a|S_{t+1})\frac{\pi(a|S_{t+1})}{b(a|S_{t+1})}Q_{h-1}(S_{t+1}, a)\\ &= 0\end{aligned}$$

1.7 *Exercise 7.7

Q

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.

A

...

1.8 Exercise 7.8

Q

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.

A

Update target is

$$G_{t:h} = \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1-\rho_t)V_{h-1}(S_t).$$

Assume state-value function does not change and introduce the TD error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t).$$

Then

$$\begin{aligned}
G_{t:h} - V(S_t) &= \rho_t \left(R_{t+1} + \gamma G_{t+1:h} - V(S_t)\right) \\
&= \rho_t \left(R_{t+1} + \gamma[G_{t+1:h} - V(S_{t+1})] + \gamma V(S_{t+1}) - V(S_t)\right) \\
&= \rho_t \delta_t + \rho_t \gamma [G_{t+1:h} - V(S_{t+1})] \\
&\vdots \\
&= \sum_{i=t}^{\min(h,T)-1} \rho_{t:i} \gamma^{i-t} \delta_i
\end{aligned}$$

1.9 Exercise 7.9

Q

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).

A

Action-value update is

$$G_{t:h} = R_{t+1} + \gamma \rho_{t+1} \left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1}) - \gamma \bar{V}_{h-1}(S_h)\right)$$

where

$$\bar{V}_h \doteq \sum_a \pi(a|S_h) Q(S_h, a)$$

and we assume that the action value function does not change between iterations. Define

$$\delta_t \doteq R_{t+1} + \gamma \bar{V}(S_{t+1}) - Q(S_t, A_t).$$

Then

$$\begin{aligned}
G_{t:h} - Q(S_t, A_t) &= \delta_t + \gamma \rho_{t+1} (G_{t+1:h} - Q(S_{t+1}, A_{t+1})) \\
&\vdots \\
&= \sum_{i=t}^{\min(h,T)-1} \gamma^{i-t} \rho_{t+1:i} \delta_i
\end{aligned}$$

where we enforce the convention that $\rho_{a:b} = 1$ if $a > b$.

1.10 Exercise 7.10 (programming)

Q

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).

A

...

1.11 Exercise 7.11

Q

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

A

Assume action-values unchanging. The recursion formula for tree backup is

$$G_{t:t+n} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1})Q(S_{t+1}, a) + \gamma\pi(A_{t+1}, S_{t+1})G_{t+1:t+n}.$$

Define

$$\delta_t \doteq R_{t+1} + \gamma\bar{V}(S_{t+1}) - Q(S_t, A_t)$$

where

$$\bar{V}_h \doteq \sum_a \pi(a|S_h)Q(S_h, a).$$

Then

$$\begin{aligned} G_{t:t+n} - Q(S_t, A_t) &= R_{t+1} + \gamma\bar{V}(S_{t+1}) - \gamma\pi(A_{t+1}, S_{t+1})Q(S_{t+1}, A_{t+1}) \\ &\quad - Q(S_t, A_t) + \gamma\pi(A_{t+1}, S_{t+1})G_{t+1:t+n} \\ &= \delta_t - \gamma\pi(A_{t+1}, S_{t+1})[G_{t+1:t+n} - Q(S_{t+1}, A_{t+1})] \\ &\vdots \\ &= \sum_{i=1}^{\min(t+n,T)-1} \delta_i \prod_{j=t+1}^{i} \gamma\pi(A_j, S_j) \end{aligned}$$

where we define the product operator to have the behaviour $\prod_a^b[\cdot] = 1$ for $a > b$.