October 7, 2022

7 Exercises

7.1 Question

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.

Answer

Value estimates are assumed not to change thus we can omit value estimate subscripts such that $V_t(S_t) = V_{t+1}(S_t)$.

n-step TD Error used in 7.2 is:

$G_{t:t+n} - V(S_t) = R_{t+1} + \gamma G_{t+1:t+n} - V(S_t)$
$G_{t:t+n} - V(S_t) = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) + \gamma G_{t+1:t+n} - \gamma V(S_{t+1})$
$G_{t:t+n} - V(S_t) = \delta_t + \gamma(G_{t+1:t+n} - V(S_{t+1}))$
$G_{t:t+n} - V(S_t) = \delta_t + \gamma\delta_{t+1} + \gamma^2(G_{t+2:t+n} - V(S_{t+2}))$
$G_{t:t+n} - V(S_t) = \sum_{k=t}^{t+n-1} \gamma^{k-t}\delta_k$

7.2 Question

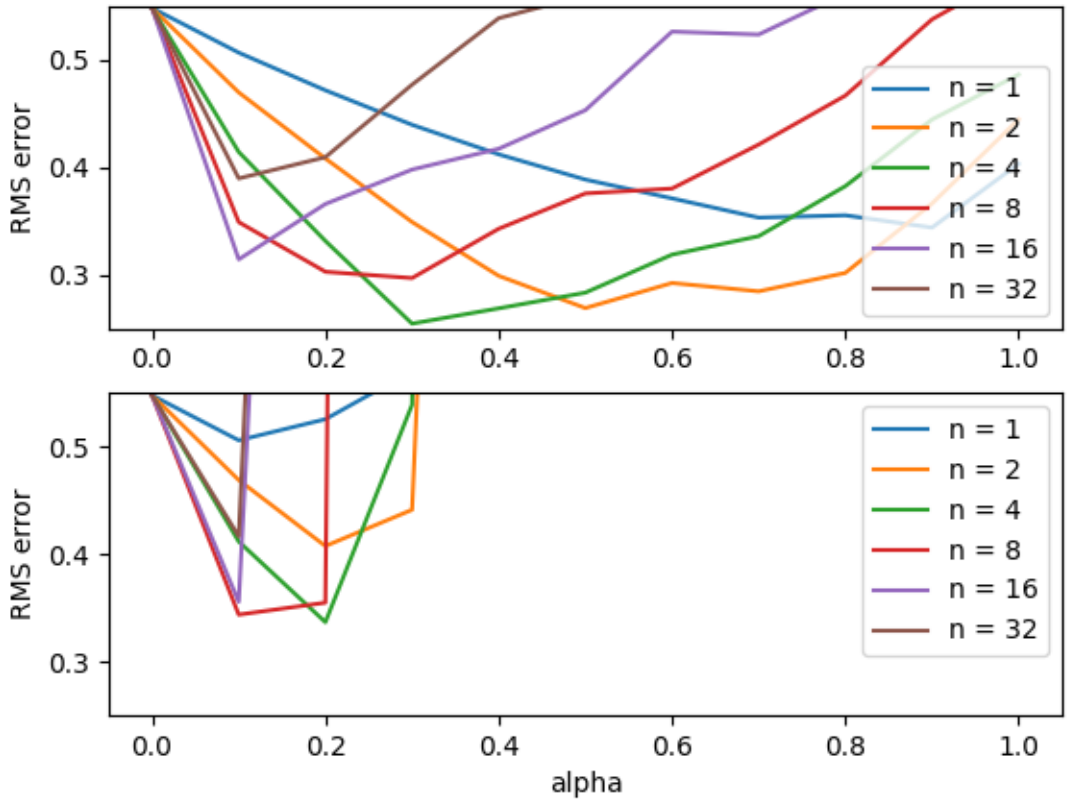
(programming) With an n-step method, the value estimates do change from step to step, so an algorithm that used the sum of TD errors (see previous exercise) in place of the error in (7.2) would actually be a slightly different algorithm. Would it be a better algorithm or a worse one? Devise and program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

Answer

The chart above shows regular n-step TD with different n parameters. The chart below shows the same configuration with unchanged value functions.

Value function updates are applied only after an episode terminates.

Using sum of TD errors as in place of the error in 7.2 performs worse in all n and α values.



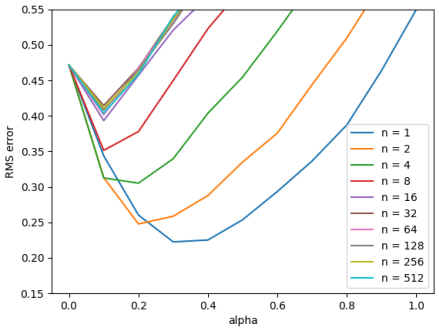
7.3 Question

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?

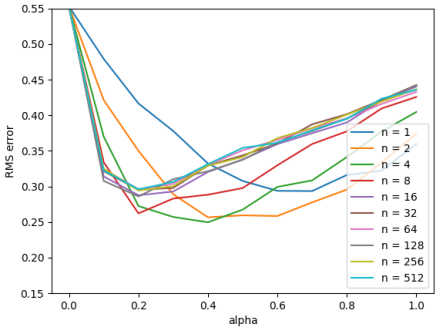
Answer

If n-step size is close to or bigger than the average number steps to complete an episode then the algorithm approaches to MC which involves variance. Using 19 states increases average number of steps to complete an episode thus helps to show how n-step size effects the algorithm.

If number of states was 5, optimum n-step size would be smaller. An empiric study shows that if return value -1 is used with 5 states, most optimum n value would be 1.



Randomwalk with 5 states and reward of 0 on the left, results are found to be different from the -1 case. We can interpolate and conclude that changing the return value may affect the result.



7.4 Question

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

Answer

Given expression:

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

Can be expanded for n:

$$\begin{aligned} G_{t:t+n} = Q_{t-1}(S_t, A_t) &+ \gamma^0[R_{t+1} + \gamma Q_t \quad (S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)] \\ &+ \gamma^1[R_{t+2} + \gamma Q_{t+1}(S_{t+2}, A_{t+2}) - Q_t \quad (S_{t+1}, A_{t+1})] \\ &+ \gamma^2[R_{t+3} + \gamma Q_{t+2}(S_{t+3}, A_{t+3}) - Q_{t+1}(S_{t+2}, A_{t+2})] \\ &\dots \\ &+ \gamma^{n-1}[R_{t+n} + \gamma Q_{t+n-1}(S_{t+n}, A_{t+n}) - Q_{t+n-2}(S_{t+n-2}, A_{t+n-2})] \end{aligned}$$

γ distributed:

$$\begin{aligned} G_{t:t+n} = Q_{t-1}(S_t, A_t) &+ \quad R_{t+1} + \gamma Q_t \quad (S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) \\ &+ \gamma \; R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1})] \\ &+ \gamma^2 R_{t+3} + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2})] \\ &\dots \\ &+ \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) - \gamma^{n-1} Q_{t+n-2}(S_{t+n-2}, A_{t+n-2})] \end{aligned}$$

After γ distribution diagonal $\gamma Q_k(S_{k+1}, A_{k+1})$ and $Q_{k-1}(S_k, A_k)$ terms cancel out.

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

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

Finally we obtain n-step return of Sarsa (7.4), hence proved.