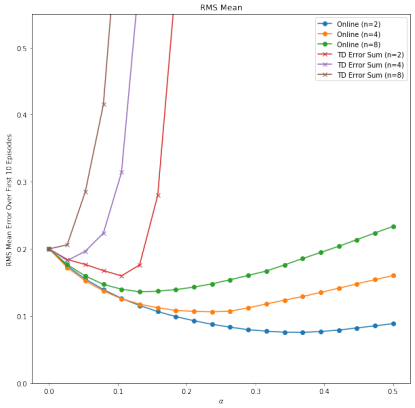
Exercise 7.1 In Chapter 6 we noted that the Monte Carlo error can be written as the sum of TD errors (6.6) if the value estimates don't change from step to step. Show that the n-step error used in (7.2) can also be written as a sum TD errors (again if the value estimates don't change) generalizing the earlier result. □

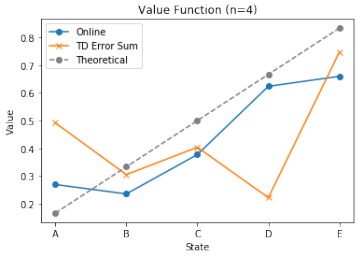
$$G_{t:t+n} - V(S_t) = \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right] + \gamma\left(G_{t+1:t+n} - V(S_{t+1})\right) =$$

$$= \delta_t + \gamma\underbrace{\left[R_{t+2} + \gamma V(S_{t+2}) - V(S_{t+1})\right]}_{\delta_{t+1}} + \gamma^2\underbrace{\left[\ldots\right]}_{\delta_{t+2}} + \ldots + \gamma^{n-1}\underbrace{\left[R_{t+n} + \gamma V(S_{t+n}) - V(S_{t+n-1})\right]}_{\delta_{t+n-1}} =$$

$$= \sum_{k=t}^{t+n-1} \gamma^{k-t}\delta_k$$

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. □





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? □

Expected distance of the agent from the starting position grows with $\sqrt{t}$. It takes 3 steps from C to terminate via state A or E so the expected duration of the episode is $\mathbb{E}[T] = 3^2 = 9$ steps.
Such short episodes would effectively truncate the update rule with $n > 9$ to $n = 9$ and smaller n methods would have an advantage. 19 state version is expected to terminate in $9^2 = 81$ steps (but with significant variance). Therefore it allows the n-step methods to propagate the terminal reward to all n previous steps.

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

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t} \left[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k) \right]. \tag{7.6}$$

□

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

$$= R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) + Q_{t-1}(S_t, A_t) + \gamma \left[R_{t+2} + \gamma Q_{t+1}(S_{t+2}, A_{t+2}) - Q_t(S_{t+1}, A_{t+1}) \right] +$$

$$+ \gamma^2 \left[R_{t+3} + \gamma Q_{t+2}(S_{t+3}, A_{t+3}) - Q_{t+1}(S_{t+2}, A_{t+2}) \right] + \dots + \gamma^{n-1} \left[R_{t+n} + \gamma Q_{t+n-1}(S_{t+n}, A_{t+n}) - Q_{t+n-2}(S_{t+n-1}, A_{t+n-1}) \right] =$$

This can be repeated for each R_t or until R_{T-1}

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

Exercise 7.5 Write the pseudocode for the off-policy state-value prediction algorithm described above. □

n-step TD for estimating $V \approx v_\pi$

Input: a 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 $S_0 \neq$ terminal
  $T \leftarrow \infty$
  Loop for $t = 0, 1, 2, \dots$:
  | If $t < T$, then:
  | Take an action according to $\pi(\cdot|S_t)$; $\rho_t \leftarrow \pi(A_t|S_t)/b(A_t|S_t)$
  | Observe and store the next reward as R_{t+1} and the next state as S_{t+1}
  | If S_{t+1} is terminal, then $T \leftarrow t + 1$
  | $\tau \leftarrow t - n + 1$ (τ is the time whose state's estimate is being
  | If $\tau \geq 0$:
  | $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$ $\leftarrow \bar{G}(\tau, \tau+n)$
  | If $\tau + n < T$, then: $G \leftarrow G + \gamma^n V(S_{\tau+n})$
  | $V(S_\tau) \leftarrow V(S_\tau) + \alpha \left[G - V(S_\tau) \right]$
  Until $\tau = T - 1$

$\bar{G}(t, h)$:
if $t = h$
  return $V(S_h)$
if $t = T$:
  return 0
return $\rho_t \left(R_{t+1} + \gamma \bar{G}(t+1, h) \right) + (1 - \rho_t) V(S_t)$

Exercise 7.6 Prove that the control variate in the above equations does not change the expected value of the return. □

The control variate return should be equal to the expectation of return when following policy π. Thaks to the linearity of expectation we have:

$$\mathbb{E}_b\left[\rho_t(R_{t+1}+\gamma G_{t+1:h})+(1-\rho_t)V_{h-1}(S_t)\right] = \underbrace{\mathbb{E}_b\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right]}\left(R_{t+1}+\gamma G_{t+1:h}\right)\Big] +$$

$$\mathbb{E}_b\left[V_{h-1}(S_t)\right] - \mathbb{E}_b\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}V_{h-1}(S_t)\right] = \mathbb{E}_\pi\left[G_{t:h}\right] + \overbrace{V_{h-1}(S_t)-V_{h-1}(S_t)}^{=0}$$

Similarly for action-value function:

$$\mathbb{E}_b\left[R_{t+1}+\gamma\rho_{t+1}\left(G_{t+1:h}-Q_{h-1}(S_{t+1},A_{t+1})\right)+\gamma\bar{V}_{h-1}(S_{t+1})\right] = \mathbb{E}\left[R_{t+1}\,|\,A_t\right] +$$

$$+\gamma\,\mathbb{E}_b\left[\frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})}G_{t+1:h}\right]-\gamma\,\mathbb{E}_b\left[\frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})}Q_{h-1}(S_{t+1},A_{t+1})\right]+\gamma\bar{V}_{h-1}(S_{t+1}) =$$

$$= \mathbb{E}\left[R_{t+1}|A_t\right]+\gamma\,\mathbb{E}_\pi\left[G_{t+1:h}\right]-\gamma\Big(\underbrace{\mathbb{E}_\pi\left[Q_{h-1}(S_{t+1},A_{t+1})\right]}_{=\bar{V}_{h-1}(S_{t+1})}+\bar{V}_{h-1}(S_{t+1})\Big)$$

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. □

Off-policy n-step Sarsa for estimating $Q \approx q_*$ or q_π

Input: an arbitrary behavior policy b such that $b(a|s) > 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}$
Initialize $Q(s,a)$ arbitrarily, for all $s \in \mathcal{S}, a \in \mathcal{A}$
Initialize π to be greedy with respect to Q, or as a fixed given policy
Algorithm parameters: step size $\alpha \in (0,1]$, a positive integer n
All store and access operations (for S_t, A_t, and R_t) can take their index

Loop for each episode:
  Initialize and store $S_0 \neq$ terminal
  Select and store an action $A_0 \sim b(\cdot|S_0)$
  $T \leftarrow \infty$
  Loop for $t = 0, 1, 2, \ldots$:
    If $t < T$, then:
      Take action A_t ; $\rho_t \leftarrow \pi(A_t|S_t)/b(A_t|S_t)$
      Observe and store the next reward as R_{t+1} and the next state as
      If S_{t+1} is terminal, then:
        $T \leftarrow t+1$
      else:
        Select and store an action $A_{t+1} \sim b(\cdot|S_{t+1})$
    $\tau \leftarrow t-n+1$ (τ is the time whose estimate is being updated)
    If $\tau \geq 0$:
      ~~$\rho \leftarrow \prod_{i=\tau+1}^{\min(\tau+n-1,T-1)} \frac{\pi(A_i|S_i)}{b(A_i|S_i)}$~~
      ~~$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$~~ $\leftarrow \bar{G}(\tau, \tau+n)$
      ~~If $\tau+n < T$, then: $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$~~
      $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha[G - Q(S_\tau, A_\tau)]$
      If π is being learned, then ensure that $\pi(\cdot|S_\tau)$ is greedy wrt Q
  Until $\tau = T-1$

$\bar{G}(t,h)$:
if $t = h$
  return $Q(S_h, A_h)$
if $t = T-1$
  return R_T
return R_{t+1} +
$+\gamma\rho_{t+1}\left(\bar{G}(t+1,h) - Q(S_{t+1},A_{t+1})\right)$
$+\gamma\,\bar{V}(t+1)$

$\bar{V}(t)$:
return $\sum_a \pi(a|S_t)Q(S_t,a)$

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. □

$G_{n:n} = V(S_n)$
$t < h < T$

$$G_{t:n} = \rho_t(R_{t+1} + \gamma G_{t+1:n}) + (1-\rho_t)V(S_t) = \rho_t(R_{t+1} + \gamma G_{t+1:n} - V(S_t)) + V(S_t) =$$

$$= V(S_t) + \rho_t\underbrace{[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]}_{\delta_t} + \gamma\rho_{t:t+1}\underbrace{(R_{t+2} + \gamma G_{t+2:n} - V(S_{t+1}))}_{\delta_{t+1} - \gamma V(S_{t+2})} =$$

$$= V(S_t) + \rho_t\delta_t + \rho_{t:t+1}\gamma\delta_{t+1} + \dots + \gamma^{h-t}\rho_{t:h}\underbrace{[R_h + \gamma V(S_h) - V(S_{h-1})]}_{\delta_{h-1}} =$$

$$= V(S_t) + \sum_{k=t}^{h-1}\rho_{t:k}\gamma^{k-t}\delta_k$$

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). □

$G_{n:n} = Q(S_n, A_n)$
$t < h \leq T$

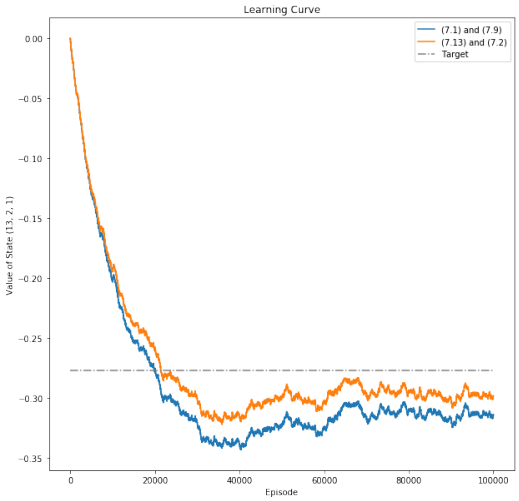
$$G_{t:h} = R_{t+1} + \gamma\rho_{t+1}(G_{t+1:h} - Q(S_{t+1}, A_{t+1})) + \gamma\bar{V}(S_{t+1}) = \left[R_{t+1} + \gamma\bar{V}(S_{t+1}) - Q(S_t, A_t)\right] +$$

$$Q(S_t, A_t) + \gamma\rho_{t+1}\left[R_{t+2} + \gamma V(S_{t+2}) - Q(S_{t+1}, A_{t+1})\right] + \gamma\rho_{t+1:t+2}(G_{t+2:h} - Q(S_{t+2}, A_{t+2})) =$$

$$= Q(S_t, A_t) + \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \dots + \gamma^{h-2-t}\rho_{t+1:h-2}\underbrace{\left[R_{h-1} + \gamma\bar{V}(S_{h-1}) - Q(S_h, A_h)\right]}_{\delta_{h-2}} =$$

$$= Q(S_t, A_t) + \sum_{k=t}^{h-2}\gamma^{k-t-1}\rho_{t+1:k}\delta_k$$

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). □



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:

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

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). □

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

$$+ \gamma \pi(A_{t+1}|S_{t+1}) \left(G_{t+1:t+n} - Q(S_{t+1}, A_{t+1})\right) = Q(S_t, A_t) + \delta_t + \gamma \pi(A_{t+1}|S_{t+1}) \delta_{t+1} +$$

$$+ \gamma^2 \pi(A_{t+1}|S_{t+1}) \pi(A_{t+2}|S_{t+2}) \left(\underbrace{G_{t+2:t+n}}_{\text{Recursion ends at } t=t+n-1 \text{ with } Q(A_{t+1}, S_{t+n}) \text{ or when the episode ends } G_{T-1:t+n} = R_T} - Q(S_{t+2}, A_{t+2})\right) =$$

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