November 29, 2022

12 Exercises

12.1 Question

Just as the return can be written recursively in terms of the first reward and itself one-step later (3.9), so can the λ-return. Derive the analogous recursive relationship from (12.2) and (12.1).

Answer

Equation 12.1:
$G_{t_t+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}(S_{t+n}, w_{t+n-1})$
Equation 12.2:
$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n}$

Let's start with expanding 12.2:
$G_t^\lambda = (1-\lambda)[\lambda^0 G_{t:t+1} + \lambda^1 G_{t:t+2} + \lambda^2 G_{t:t+3} + \ldots]$
$G_t^\lambda = (1-\lambda)G_{t:t+1} + (1-\lambda)[\lambda^1 G_{t:t+2} + \lambda^2 G_{t:t+3} + \ldots]$
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)[\lambda^0 G_{t:t+2} + \lambda^1 G_{t:t+3} + \ldots]$
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+1+n}$

Let's not touch the left expression for now and focus on the second expression. We have a return of form $G_{t:t+1+n}$ but a return of form $G_{t+1:t+1+n}$ would be more useful. Let's try to express one in the form of the other.
$G_{t+1:t+1+n} = R_{t+2} + \gamma R_{t+3} + \ldots + \gamma^{n-1} R_{t+1+n} + \gamma^n \hat{v}(s_{t+1+n}, w_{t+n})$ (12.1.1)
$G_{t:t+1+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^n R_{t+1+n} + \gamma^{n+1} \hat{v}(s_{t+1+n}, w_{t+n})$
$G_{t:t+1+n} = R_{t+1} + \gamma[R_{t+2} + \gamma R_{t+3} + \ldots + \gamma^{n-1} R_{t+1+n} + \gamma^n \hat{v}(s_{t+1+n}, w_{t+n})]$
$G_{t:t+1+n} = R_{t+1} + \gamma G_{t+1:t+1+n}$ (12.1.2)

Going back and using equation (12.1.2):
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1}(R_{t+1} + \gamma G_{t+1:t+1+n})$
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} R_{t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} \gamma G_{t+1:t+1+n}$

$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}R_{t+1} + \gamma\lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1:t+1+n}$

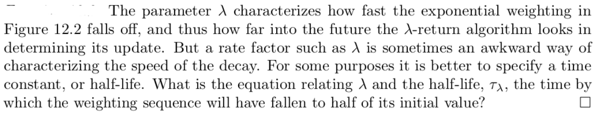
We have a geometric series of form ($\sum_{n=1}^{\text{inf}} r\lambda^{n-1} = \frac{r}{1-\lambda}$):
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda(1-\lambda)\frac{R_{t+1}}{(1-\lambda)} + \gamma\lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1:t+1+n}$
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda R_{t+1} + \gamma\lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1:t+1+n}$

The right most expression is a lambda return:
$G_t^\lambda = (1-\lambda)G_{t:t+1} + \lambda R_{t+1} + \gamma\lambda G_{t+1}^\lambda$

We can further simplify by expanding the 1-step return:
$G_t^\lambda = (1-\lambda)(R_{t+1} + \gamma G_{t+1:t+1}) + \lambda R_{t+1} + \gamma\lambda G_{t+1}^\lambda$
$G_t^\lambda = (1-\lambda)(R_{t+1} + \gamma\hat{v}(s_{t+1}, w_t)) + \lambda R_{t+1} + \gamma\lambda G_{t+1}^\lambda$
$G_t^\lambda = (1-\lambda)\gamma\hat{v}(s_{t+1}, w_t) + (1-\lambda)R_{t+1} + \lambda R_{t+1} + \gamma\lambda G_{t+1}^\lambda$
$G_t^\lambda = (1-\lambda)\gamma\hat{v}(s_{t+1}, w_t) + R_{t+1} + \gamma\lambda G_{t+1}^\lambda$
$G_t^\lambda = R_{t+1} + (1-\lambda)\gamma\hat{v}(s_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda$ where $t < T-1$

if $t \geq T-1$ then by equation 12.3:
$G_t^\lambda = G_t$

12.2 Question



The parameter λ characterizes how fast the exponential weighting in Figure 12.2 falls off, and thus how far into the future the λ-return algorithm looks in determining its update. But a rate factor such as λ is sometimes an awkward way of characterizing the speed of the decay. For some purposes it is better to specify a time constant, or half-life. What is the equation relating λ and the half-life, τ_λ, the time by which the weighting sequence will have fallen to half of its initial value? □

Answer

Half of initial weighting sequence: $\frac{(1-\lambda)}{2}$
Weighting sequence at time $t+m$: $(1-\lambda)\lambda^m$
Then by definition: $(1-\lambda)\lambda^m = \frac{(1-\lambda)}{2}$
$\lambda^m = \frac{1}{2}$
$m = \log_\lambda \frac{1}{2} = -\log_\lambda 2 = \frac{\ln 2}{\ln \lambda}$
The actual time step is:
$\tau_\lambda = t + m = t + \frac{\ln 2}{\ln \lambda}$

12.3 Question

Some insight into how $TD(\lambda)$ can closely approximate the offline λ-return algorithm can be gained by seeing that the latter's error term (in brackets in (12.4)) can be written as the sum of TD errors (12.6) for a single fixed w.

Show this, following the pattern of (6.6), and using the recursive relationship for the λ-return you obtained in Exercise 12.1.

Answer

Error term in (12.4) ($w_{t+1} = w_t + \alpha[G_t^\lambda - \hat{v}(S_t, w_t)]\Delta\hat{v}(S_t, w_t)$):
$G_t^\lambda - \hat{v}(S_t, w_t)$

TD error (12.6):
$\delta_t = R_{t+1} + \gamma\hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t)$

Recursive relationship for the λ-return obtained in Exercise 12.1:
$G_t^\lambda = R_{t+1} + (1-\lambda)\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda$

Following the pattern of (6.6):
$G_t^\lambda - \hat{v}(S_t, w_t) = R_{t+1} + (1-\lambda)\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda - \hat{v}(S_t, w_t)$
$G_t^\lambda - \hat{v}(S_t, w_t) = R_{t+1} + (1-\lambda)\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda - \hat{v}(S_t, w_t) - \gamma\hat{v}(S_{t+1}, w_t) + \gamma\hat{v}(S_{t+1}, w_t)$
$G_t^\lambda - \hat{v}(S_t, w_t) = \delta_t + (1-\lambda)\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda - \gamma\hat{v}(S_{t+1}, w_t)$
$G_t^\lambda - \hat{v}(S_t, w_t) = \delta_t - \lambda\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda G_{t+1}^\lambda$
$G_t^\lambda - \hat{v}(S_t, w_t) = \delta_t - \lambda\gamma\hat{v}(S_{t+1}, w_t) + \gamma\lambda[R_{t+2} + (1-\lambda)\gamma\hat{v}(S_{t+2}, w_{t+1}) + \gamma\lambda G_{t+2}^\lambda + \gamma\hat{v}(S_{t+2}, w_{t+1}) - \gamma\hat{v}(S_{t+2}, w_{t+1})]$
$G_t^\lambda - \hat{v}(S_t, w_t) = \delta_t + \gamma\lambda[\delta_{t+1} + (1-\lambda)\gamma\hat{v}(S_{t+2}, w_{t+1}) + \gamma\lambda G_{t+2}^\lambda - \gamma\hat{v}(S_{t+2}, w_{t+1})]$
$G_t^\lambda - \hat{v}(S_t, w_t) = \delta_t + \gamma\lambda[\delta_{t+1} - \lambda\gamma\hat{v}(S_{t+2}, w_{t+1}) + \gamma\lambda G_{t+2}^\lambda]$

for t = T-1, the inner-most expression will be:
$\delta_{T-2} - \gamma\lambda\hat{v}(S_{T-1}, w_{T-2}) + \gamma\lambda G_{T-1}^\lambda = \delta_{T-2} - \gamma\lambda\hat{v}(S_{T-1}, w_{T-2}) + \gamma\lambda G_t$
$= \delta_{T-2} - \gamma\lambda\hat{v}(S_{T-1}, w_{T-2}) + \gamma\lambda(R_T + \gamma\hat{v}(S_T, w_{T-1}))$
$= \delta_{T-2} + \gamma\lambda(R_T + \gamma\hat{v}(S_T, w_{T-1}) - \hat{v}(S_{T-1}, w_{T-2}))$
$= \delta_{T-2} + \gamma\lambda\delta_{T-1}$

Now we can write whole expression as a proper sum:
$G_t^\lambda - \hat{v}(S_t, w_t) = \sum_{k=t}^{T-1}(\lambda\gamma)^{(k-t)}\delta_k$

12.4 Question

Use your result from the preceding exercise to show that, if the weight updates over an episode were computed on each step but not actually used to change the weights (w remained fixed), then the sum of TD(λ)'s weight updates would be the same as the sum of the offline λ-return algorithm's updates.

Answer

Equation 12.5:
$z_{-1} = 0$
$z_t = \gamma\lambda z_{t-1} + \Delta\hat{v}(S_t, w_t)$ where $0 \leq t \leq T$

Let's try to write as a sum:
$z_t = \gamma\lambda z_{t-1} + \Delta\hat{v}(S_t, w)$
$z_t = \gamma\lambda(\gamma\lambda z_{t-2} + \Delta\hat{v}(S_{t-1}, w)) + \Delta\hat{v}(S_t, w)$
$z_t = \sum_{k=0}^{t}(\gamma\lambda)^{k-t}\Delta\hat{v}(S_k, w)$

Sum of TD(λ)'s weight updates:
$\alpha\sum_{t=0}^{\text{inf}}\delta_t z_t = \alpha\sum_{t=0}^{\text{inf}}\delta_t\sum_{k=0}^{t}(\gamma\lambda)^{t-k}\Delta\hat{v}(S_k, w)$

Expanding the sum:
t=0 $\alpha\delta_0[(\gamma\lambda)^0\Delta\hat{v}(S_0, w)]$
t=1 $\alpha\delta_1[(\gamma\lambda)^1\Delta\hat{v}(S_0, w) + (\gamma\lambda)^0\Delta\hat{v}(S_1, w)]$
t=2 $\alpha\delta_2[(\gamma\lambda)^2\Delta\hat{v}(S_0, w) + (\gamma\lambda)^1\Delta\hat{v}(S_1, w) + (\gamma\lambda)^0\Delta\hat{v}(S_2, w)]$
t=3 $\alpha\delta_3[(\gamma\lambda)^3\Delta\hat{v}(S_0, w)+(\gamma\lambda)^2\Delta\hat{v}(S_1, w)+(\gamma\lambda)^1\Delta\hat{v}(S_2, w)+(\gamma\lambda)^0\Delta\hat{v}(S_3, w)]$

Let's sum vertically:
$s = S_0\ \alpha\Delta\hat{v}(S_0, w)\sum_{k=0}^{\text{inf}}(\gamma\lambda)^k\delta_k$
$s = S_1\ \alpha\Delta\hat{v}(S_1, w)\sum_{k=0}^{\text{inf}}(\gamma\lambda)^k\delta_{k+1}$
$s = S_2\ \alpha\Delta\hat{v}(S_2, w)\sum_{k=0}^{\text{inf}}(\gamma\lambda)^k\delta_{k+2}$
$s = S_3\ \alpha\Delta\hat{v}(S_3, w)\sum_{k=0}^{\text{inf}}(\gamma\lambda)^k\delta_{k+3}$

$\alpha\sum_{t=0}^{\text{inf}}\delta_t z_t = \sum_{t=0}^{\text{inf}}\Delta\hat{v}(S_t, w)\sum_{k=0}^{\text{inf}}(\lambda\gamma)^k\delta_{k+t}$

Start inner index k from t so that it is similar to result of question 12.3:
$\alpha\sum_{t=0}^{\text{inf}}\delta_t z_t = \sum_{t=0}^{\text{inf}}\Delta\hat{v}(S_t, w)\sum_{k=t}^{\text{inf}}(\lambda\gamma)^{k-t}\delta_t$

Now we can use result of question 12.3
$\alpha\sum_{t=0}^{\text{inf}}\delta_t z_t = \sum_{t=0}^{\text{inf}}\Delta\hat{v}(S_t, w)(G_t^\lambda - \hat{v}(S_t, w_t))$

12.5 Question

Several times in this book (often in exercises) we have established that returns can be written as sums of TD errors if the value function is held constant. Why is (12.10) another instance of this? Prove (12.10).

Answer

Equation 12.9 (h replaced with t+n):
$G^{\lambda}_{t:t+n} = (1-\lambda)\sum_{k=1}^{n-1} \lambda^{k-1} G_{t:t+k} + \lambda^{n-1} G_{t:t+n}$

Get rid of $(1-\lambda)$:
$G^{\lambda}_{t:t+n} = \sum_{k=1}^{n-1} \lambda^{k-1} G_{t:t+k} - \sum_{k=1}^{n-1} \lambda^{k} G_{t:t+k} + \lambda^{n-1} G_{t:t+n}$

Make the sums similar :
$G^{\lambda}_{t:t+n} = \sum_{k=1}^{n} \lambda^{k-1} G_{t:t+k} - \sum_{k=1}^{n-1} \lambda^{k} G_{t:t+k}$
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} G_{t:t+k+1} - \sum_{k=1}^{n-1} \lambda^{k} G_{t:t+k}$
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} G_{t:t+k+1} - \sum_{k=0}^{n-1} \lambda^{k} G_{t:t+k} - (-G_{t:t+0})$
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} G_{t:t+k+1} - \sum_{k=0}^{n-1} \lambda^{k} G_{t:t+k} + \hat{v}(S_t, w)$
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} [G_{t:t+k+1} - G_{t:t+k}] + \hat{v}(S_t, w)$

The returns can be written in the form of the other. One can expand both returns to show that:
$G_{t:t+k+1} = G_{t:t+k} + \gamma^{k} R_{k+1} + \gamma^{k+1} \hat{v}(S_{k+1}, w) - \gamma^{k} \hat{v}(S_k, w)$
$G_{t:t+k+1} = G_{t:t+k} + \gamma^{k} (R_{k+1} + \gamma \hat{v}(S_{k+1}, w) - \hat{v}(S_k, w))$
$G_{t:t+k+1} = G_{t:t+k} + \gamma^{k} \delta_{t+k}$

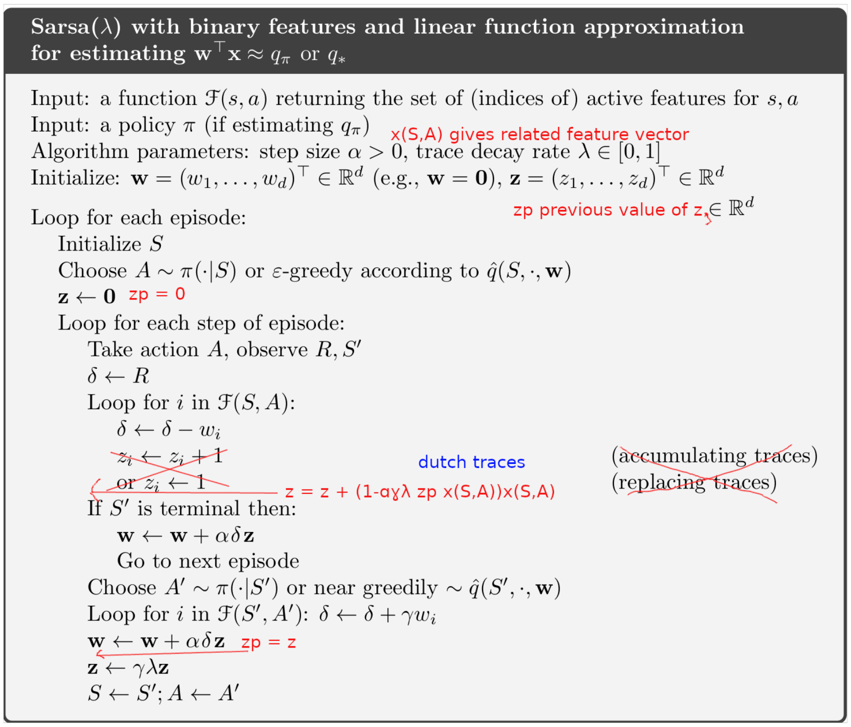
Apply the last finding to the equation:
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} [G_{t:t+k} + \gamma^{k} \delta_{t+k} - G_{t:t+k}] + \hat{v}(S_t, w)$
$G^{\lambda}_{t:t+n} = \sum_{k=0}^{n-1} \lambda^{k} \gamma^{k} \delta_{t+k} + \hat{v}(S_t, w)$

Make δ indexed with i in the same way as in 12.10:
$G^{\lambda}_{t:t+n} = \sum_{k=t}^{t+n-1} \lambda^{k-t} \gamma^{k-t} \delta_{k} + \hat{v}(S_t, w)$
$G^{\lambda}_{t:t+n} = \sum_{i=t}^{t+n-1} \lambda^{i-t} \gamma^{i-t} \delta_{i} + \hat{v}(S_t, w)$

12.6 Question

Modify the pseudocode for Sarsa(λ) to use dutch traces (12.11) without the other distinctive features of a true online algorithm. Assume linear function approximation and binary features.

Answer



Sarsa(λ) with binary features and linear function approximation for estimating $\mathbf{w}^\top \mathbf{x} \approx q_\pi$ or q_*

```
Input: a function $\mathcal{F}(s,a)$ returning the set of (indices of) active features for $s, a$
Input: a policy $\pi$ (if estimating $q_\pi$)   x(S,A) gives related feature vector
Algorithm parameters: step size $\alpha > 0$, trace decay rate $\lambda \in [0,1]$
Initialize: $\mathbf{w} = (w_1, \dots, w_d)^\top \in \mathbb{R}^d$ (e.g., $\mathbf{w} = \mathbf{0}$), $\mathbf{z} = (z_1, \dots, z_d)^\top \in \mathbb{R}^d$
                                                   zp previous value of z $\in \mathbb{R}^d$
Loop for each episode:
    Initialize $S$
    Choose $A \sim \pi(\cdot|S)$ or $\varepsilon$-greedy according to $\hat{q}(S, \cdot, \mathbf{w})$
    $\mathbf{z} \leftarrow \mathbf{0}$  zp = 0
    Loop for each step of episode:
        Take action $A$, observe $R, S'$
        $\delta \leftarrow R$
        Loop for $i$ in $\mathcal{F}(S,A)$:
            $\delta \leftarrow \delta - w_i$
            ~~$z_i \leftarrow z_i + 1$~~                dutch traces     ~~(accumulating traces)~~
            ~~or $z_i \leftarrow 1$~~  z = z + (1-αγλ zp x(S,A))x(S,A)   ~~(replacing traces)~~
        If $S'$ is terminal then:
            $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\mathbf{z}$
            Go to next episode
        Choose $A' \sim \pi(\cdot|S')$ or near greedily $\sim \hat{q}(S', \cdot, \mathbf{w})$
        Loop for $i$ in $\mathcal{F}(S',A')$: $\delta \leftarrow \delta + \gamma w_i$
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\mathbf{z}$  zp = z
        $\mathbf{z} \leftarrow \gamma\lambda\mathbf{z}$
        $S \leftarrow S'$; $A \leftarrow A'$
```

12.7 Question

Generalize the three recursive equations above to their truncated versions.

Answer

Truncated versions of the recursive equations should be similar. However, recursion conditions have to change. Normally all components up to the end of episode are accumulating after fading as it is the case in figure 12.1. In the truncated version, only components up to min(t+n,T) should be accumulated as it the case in figure 12.7. Thus, recursive functions should terminate after returning:

$$\lambda^{\begin{cases} n-1 & \text{if } t+n<T \\ T-t-1 & \text{otherwise} \end{cases}} G^{\lambda}_{t:\min(t+n,T)}$$

12.8 Question

Prove that (12.24) becomes exact if the value function does not change. To save writing, consider the case of t = 0, and use the notation

Answer

Equation 12.22:
$G_t^{\lambda s} = \rho_t[R_{t+1}\gamma_{t+1}((1-\lambda_{t+1})\hat{v}_{t+1} + \lambda_{t+1}G_{t+1}^{\lambda s})] + (1-\rho_t)\hat{v}_t$

Start with removing the parenthesis:
$G_t^{\lambda s} = \rho_t[R_{t+1}\gamma_{t+1}(\hat{v}_{t+1} - \lambda_{t+1}\hat{v}_{t+1} + \lambda_{t+1}G_{t+1}^{\lambda s}) - \hat{v}_t] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[R_{t+1} + \gamma_{t+1}\hat{v}_{t+1} - \hat{v}_t + \gamma_{t+1}\lambda_{t+1}(-\hat{v}_{t+1} + G_{t+1}^{\lambda s})] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[\delta_t + \gamma_{t+1}\lambda_{t+1}(-\hat{v}_{t+1} + G_{t+1}^{\lambda s})] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[\delta_t + \gamma_{t+1}\lambda_{t+1}(-\hat{v}_{t+1} + \rho_{t+1}[\delta_{t+1} + \gamma_{t+2}\lambda_{t+2}(-\hat{v}_{t+2} + G_{t+2}^{\lambda s})] + \hat{v}_{t+1})] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[\delta_t + \gamma_{t+1}\lambda_{t+1}\rho_{t+1}[\delta_{t+1} + \gamma_{t+2}\lambda_{t+2}(-\hat{v}_{t+2} + G_{t+2}^{\lambda s})]] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[\delta_t + \gamma_{t+1}\lambda_{t+1}\rho_{t+1}[\delta_{t+1} + \gamma_{t+2}\lambda_{t+2}(-\hat{v}_{t+2} + \rho_{t+2}[\delta_{t+2} + \gamma_{t+3}\lambda_{t+3}(-\hat{v}_{t+3} +$
$G_{t+3}^{\lambda s})] + \hat{v}_{t+2})]] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t[\delta_t + \gamma_{t+1}\lambda_{t+1}\rho_{t+1}[\delta_{t+1} + \gamma_{t+2}\lambda_{t+2}\rho_{t+2}[\delta_{t+2} + \gamma_{t+3}\lambda_{t+3}(-\hat{v}_{t+3} +$
$G_{t+3}^{\lambda s})]]] + \hat{v}_t$

Distribute the leading coefficients to see the pattern:
$G_t^{\lambda s} = \rho_t\delta_t + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}[\delta_{t+1} + \gamma_{t+2}\lambda_{t+2}\rho_{t+2}[\delta_{t+2} + \gamma_{t+3}\lambda_{t+3}(-\hat{v}_{t+3} +$
$G_{t+3}^{\lambda s})]] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t\delta_t + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}\delta_{t+1} + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}\gamma_{t+2}\lambda_{t+2}\rho_{t+2}[\delta_{t+2} + \gamma_{t+3}\lambda_{t+3}(-\hat{v}_{t+3} +$
$G_{t+3}^{\lambda s})] + \hat{v}_t$
$G_t^{\lambda s} = \rho_t\delta_t + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}\delta_{t+1} + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}\gamma_{t+2}\lambda_{t+2}\rho_{t+2}\delta_{t+2} + \rho_t\gamma_{t+1}\lambda_{t+1}\rho_{t+1}\gamma_{t+2}\lambda_{t+2}\rho_{t+2}\gamma_{t+3}\lambda_{t+3}$
$G_{t+3}^{\lambda s}) + \hat{v}_t$
$G_t^{\lambda s} = \delta_t\rho_t + \delta_{t+1}\rho_t\rho_{t+1}\gamma_{t+1}\lambda_{t+1} + \delta_{t+2}\rho_t\rho_{t+1}\rho_{t+2}\gamma_{t+1}\gamma_{t+2}\lambda_{t+1}\lambda_{t+2} + \rho_t\rho_{t+1}\rho_{t+2}\gamma_{t+1}\gamma_{t+2}\gamma_{t+3}\lambda_{t+1}\lambda_{t+2}\lambda_{t+3}$
$G_{t+3}^{\lambda s}) + \hat{v}_t$
$G_t^{\lambda s} = \rho_t \sum_{k=t}^{\text{inf}} \delta_k \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i + \hat{v}_t$

12.9 Question

The truncated version of the general off-policy return is denoted G t:h . Guess the correct equation, based on (12.24).

Answer

The truncated version of 12.24 should sum up to h instead of infinity.
$G_{t:h}^{\lambda s} = \rho_t \sum_{k=t}^{h} \delta_k \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i + \hat{v}_t$

12.10 Question

Prove that (12.27) becomes exact if the value function does not change. To save writing, consider the case of t = 0, and use the notation Q k = q(S k , A k , w). Hint: Start by writing out 0 a and G 0 a , then G 0 a Q 0 .

Answer

Equation 12.26:
$G_t^{\lambda a} = R_{t+1} + \gamma_{t+1}(\bar{V}_t(S_{t+1}) + \lambda_{t+1}\rho_{t+1}[G_{t+1}^{\lambda a} - \hat{q}(S_{t+1}, A_{t+1}, w_t)])$

Equation 12.28:
$\delta_t^a = R_{t+1} + \gamma_{t+1}\bar{V}_t(S_{t+1}) - \hat{q}(S_t, A_t, w_t)]$

Following the hints:
$\delta_0^a = R_1 + \gamma_1\bar{V}(S_1) - Q_0$
$G_0^{\lambda a} = R_1 + \gamma_1(\bar{V}(S_1) + \lambda_1\rho_1[G_1^{\lambda a} - Q_1])$
$G_0^{\lambda a} - Q_0 = R_1 + \gamma_1(\bar{V}(S_1) + \lambda_1\rho_1[G_1^{\lambda a} - Q_1]) - Q_0$
$G_0^{\lambda a} - Q_0 = \delta_0^a + \gamma_1\lambda_1\rho_1[G_1^{\lambda a} - Q_1]$
$G_0^{\lambda a} - Q_0 = \delta_0^a + \gamma_1\lambda_1\rho_1[\delta_1^a + \gamma_2\lambda_2\rho_2[G_2^{\lambda a} - Q_2]]$

The pattern is visible:
$G_0^{\lambda a} - Q_0 = \sum_{k=0}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i$
$G_t^{\lambda a} - Q_t = \sum_{k=t}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i$
$G_t^{\lambda a} = Q_t + \sum_{k=t}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i$

12.11 Question

The truncated version of the general off-policy return is denoted G t:h. Guess the correct equation for it, based on (12.27).

Answer

$G_{t:h}^{\lambda a} = Q_t + \sum_{k=t}^{h} \delta_k^a \prod_{i=t+1}^{k} \gamma_i\lambda_i\rho_i$

12.12 Question

Show in detail the steps outlined above for deriving (12.29) from (12.27). Start with the update (12.15), substitute $G_t^{\lambda a}$ a from (12.26) for G_t^{λ} , then follow similar steps as led to (12.25).

Answer

Equation 12.27:
$G_t^{\lambda a} = Q_t + \sum_{k=t}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$

Following the hints:
$w_{t+1} = w_t + \alpha[G_t^{\lambda} - Q_t]\Delta Q_t$ t=0,1,. . . ,T-1 Equation 12.15
$w_{t+1} = w_t + \alpha[G_t^{\lambda a} - Q_t]\Delta Q_t$ t=0,1,. . . ,T-1 Equation 12.15
$w_{t+1} = w_t + \alpha[Q_t + \sum_{k=t}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i - Q_t]\Delta Q_t$ t=0,1,. . . ,T-1 Equation 12.15
$w_{t+1} = w_t + \alpha[\sum_{k=t}^{\inf} \delta_k^a \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i]\Delta Q_t$ t=0,1,. . . ,T-1 Equation 12.15

The sum of the forward-view update over time is:
$\sum_{t=1}^{\infty}(w_{t+1} - w_t) = \sum_{t=1}^{\infty} \sum_{k=t}^{\infty} \alpha \delta_k^a \Delta Q_t \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$
$\sum_{t=1}^{\infty}(w_{t+1} - w_t) = \sum_{k=1}^{\infty} \sum_{t=1}^{k} \alpha \delta_k^a \Delta Q_t \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$ by summation rule $\sum_{t=x}^{y} \sum_{k=t}^{y} = \sum_{k=x}^{y} \sum_{t=x}^{k}$
$\sum_{t=1}^{\infty}(w_{t+1} - w_t) = \sum_{k=1}^{\infty} \alpha \delta_k^a \sum_{t=1}^{k} \Delta Q_t \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$
$z_k = \sum_{t=1}^{k} \Delta Q_t \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i$
$z_k = \sum_{t=1}^{k-1} \Delta Q_t \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i + \Delta Q_k$
$z_k = \gamma_k \lambda_k \rho_k [\sum_{t=1}^{k-1} \Delta Q_t \prod_{i=t+1}^{k-1} \gamma_i \lambda_i \rho_i] + \Delta Q_k$
$z_k = \gamma_k \lambda_k \rho_k z_{k-1} + \Delta Q_k$
$z_t = \gamma_t \lambda_t \rho_t z_{t-1} + \Delta Q_t$

12.13 Question

What are the dutch-trace and replacing-trace versions of off-policy eligibility traces for state-value and action-value methods?

Answer

My dutch-trace guesses are:
$z_t = \rho_t(\gamma_t \lambda_t z_{t-1} + (1 - \alpha \gamma_t \lambda_t z_{t-1} x_t) x_t)$
$z_t = \gamma_t \lambda_t \rho_t z_{t-1} + (1 - \alpha \gamma_t \lambda_t z_{t-1} \rho_t x_t) x_t$

My replacing-trace guesses are (the replacing traces are defined for binary features):
$z_{t,i} = \rho_t(\gamma_t \lambda_t z_{t-1}(1 - x_{t,i}) + x_{t,i})$
$z_{t,i} = \gamma_t \lambda_t \rho_t z_{t-1}(1 - x_{t,i}) + \rho_t x_{t,i}$

12.14 Question

What are the dutch-trace and replacing-trace versions of off-policy eligibility traces for state-value and action-value methods?

Answer

Semi gradient parameter update rule (12.7)
$w_{t+1} = w_t + \alpha \delta_t z_t$

Dutch trace (12.11):
$z_t = \lambda \gamma z_{t-1} + (1 - \alpha \gamma \lambda z_{t-1}^T x_t) x_t$

Expectation based TD error (12.28)
$\delta_t^a = R_{t+1} + \gamma_{t+1} \bar{V}_t(S_{t+1}) - \hat{q}(S_t, A_t, w_t)$

Expectation approximate value function (12.21)
$\bar{V}_t(s) = \sum_a \pi(a|s) \hat{q}(s, a, w)$

Slightly modify 12.21:
$\bar{V}_{k,t}(s) = \sum_a \pi_k(a|s) \hat{q}(s, a, w_k)$

In double expected SARSA there will be two policy to select actions π_1 and π_2 with their respective action value functions Q_1 and Q_2. (From exercise 6.12)
$Q_1(S_t, A_t) = Q_1(S_t, A_t) + \alpha[R_{t+1} + \gamma E_{\pi_2}[Q_2(S_{t+1}, A_{t+1})|S_{t+1}] - Q_1(S_t, A_t)]$
$Q_2(S_t, A_t) = Q_2(S_t, A_t) + \alpha[R_{t+1} + \gamma E_{\pi_1}[Q_1(S_{t+1}, A_{t+1})|S_{t+1}] - Q_2(S_t, A_t)]$

After combining all the available data, the update rules should look like:
$w_{1,t+1} = w_{1,t} + \alpha(R_{t+1} + \gamma_{t+1} \bar{V}_{2,t}(S_{t+1}) - \hat{q}(S_t, A_t, w_{1,t})) z_t$
$w_{2,t+1} = w_{2,t} + \alpha(R_{t+1} + \gamma_{t+1} \bar{V}_{1,t}(S_{t+1}) - \hat{q}(S_t, A_t, w_{2,t})) z_t$