ELEN E6885: Introduction to Reinforcement Learning
Homework #3

Chenye Yang cy2540@columbia.edu

November 14, 2019

P1

Ans:
The lift-hand side of the equation can be written as

$$
\begin{aligned}
&\max_s \Big| \mathbb{E}_\pi \left[G_{t:t+n} \mid S_t = s \right] - v_\pi(s) \Big| \\
=&\max_s \Big| \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n}) \mid S_t = s \right] - v_\pi(s) \Big| \\
=&\max_s \Big| \mathbb{E}_\pi \left[R_{t:t+n} + \gamma^n V_{t+n-1}(S_{t+n}) \mid S_t = s \right] - v_\pi(s) \Big|,
\end{aligned}
$$

where $R_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n}$
Because

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right] \\
&= \mathbb{E}_\pi \left[R_{t:t+n} + \gamma^n \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \mid S_t = s \right] \\
&= \mathbb{E}_\pi \left[R_{t:t+n} \mid S_t = s \right] + \gamma^n \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \mid S_t = s \right]
\end{aligned}
$$

The lift-hand side of the equation is

$$
\begin{aligned}
&\max_s \Big| \mathbb{E}_\pi \left[R_{t:t+n} + \gamma^n V_{t+n-1}(S_{t+n}) \mid S_t = s \right] - v_\pi(s) \Big| \\
=&\max_s \Big| \mathbb{E}_\pi \left[\gamma^n V_{t+n-1}(S_{t+n}) \mid S_t = s \right] - \gamma^n \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \mid S_t = s \right] \Big| \\
=&\gamma^n \max_s \Big| \mathbb{E}_\pi \left[V_{t+n-1}(S_{t+n}) - \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \mid S_t = s \right] \Big|
\end{aligned}
$$

Considering $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$, we get

$$
\Big| \mathbb{E}_\pi \left[V_{t+n-1}(S_{t+n}) - \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \mid S_t = s \right] \Big| \leq \mathbb{E}_\pi \left[\Big| V_{t+n-1}(S_{t+n}) - \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \Big| \mid S_t = s \right]
$$

Thus the left-hand side of the equation has

$$\max_s |\mathbb{E}_\pi [G_{t:t+n}|S_t = s] - v_\pi(s)| \leq \gamma^n \max_s \mathbb{E}_\pi \left[\left| V_{t+n-1}(S_{t+n}) - \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \right| \mid S_t = s \right]$$

Given a completely free, unrestricted choice for S_t, the set of possible states S_{t+n} is a subset of the set of possible states S_t. Thus the following inequality is true.

$$\max_s \mathbb{E}_\pi \left[\left| V_{t+n-1}(S_{t+n}) - \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1} \right| \mid S_t = s \right] \leq \max_s \mathbb{E}_\pi \left[\left| V_{t+n-1}(S_t) - \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \right| \mid S_t = s \right]$$

Therefore, the left-hand side of the equation

$$\begin{aligned} \max_s |\mathbb{E}_\pi [G_{t:t+n}|S_t = s] - v_\pi(s)| &\leq \gamma^n \max_s \mathbb{E}_\pi \left[\left| V_{t+n-1}(S_t) - \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \right| \mid S_t = s \right] \\ &= \gamma^n \max_s \mathbb{E}_\pi \left[\left| V_{t+n-1}(S_t) - v_\pi(s) \right| \mid S_t = s \right] \\ &= \gamma^n \max_s \left| V_{t+n-1}(s) - v_\pi(s) \right| \end{aligned}$$

The error reduction property of n-step returns is proved.

P2

1.

Ans:
On-line updating methods update value function after each step in the episode.
In off-line updating methods, updates are accumulated within episode and applied in batch at the end of episode.

2.

Ans:
$\alpha = 0.1$
On-line every-visit constant-α Monte Carlo method:

$$\begin{aligned} &V_0(A) = 0 \\ &V_1(A) = V_0(A) + \alpha \times (1 + 2 + 1 - V_0(A)) = 0.4 \\ &V_2(A) = V_1(A) + \alpha \times (1 - V_1(A)) = 0.46 \end{aligned}$$

Off-line every-visit constant-α Monte Carlo method:

$$\begin{aligned} &\alpha \times (1 + 2 + 1 - V_0(A)) = 0.4 \\ &\alpha \times (1 - V_0(A)) = 0.1 \end{aligned}$$

$$V(A) = V_0(A) + 0.4 + 0.1 = 0.5$$

3.

Ans:
On-line $TD(0)$ method, $\lambda = 0$:

$$\begin{aligned} V_1(A) &= V_0(A) + \alpha\,(1 - V_0(A)) = 0.1 \\ V_2(A) &= V_1(A) + \alpha\,(1 - V_1(A)) = 0.19 \end{aligned}$$

Off-line $TD(0)$ method, $\lambda = 0$:

$$\begin{aligned} \alpha\,(1 - V_0(A)) = 0.1 \\ \alpha\,(1 - V_0(A)) = 0.1 \end{aligned}$$

$$V(A) = V_0(A) + 0.1 + 0.1 = 0.2$$

4.

Ans:
$\lambda = 0.5$
On-line forward-view $TD(\lambda)$ method:

$$\begin{aligned} V_1(A) &= V_0(A) + \alpha\left(G_t^\lambda - V_0(A)\right) \\ &= 0.225 \\ V(A) &= V_1(A) + \alpha\left(G_t^\lambda - V_1(A)\right) \\ &= 0.225 + 0.1 \times (1 - 0.225) \\ &= 0.3025 \end{aligned}$$

Off-line forward-view $TD(\lambda)$ method:

$$\alpha\left(G_{t_1}^{\lambda}-V_0(A)\right)=0.225$$
$$\alpha\left(G_{t_2}^{\lambda}-V_0(A)\right)=0.1$$

$$V(A)=V_0(A)+0.225+0.1=0.325$$

5.

Ans:
On-line backward-view $TD(\lambda)$ method, $\gamma=1,\lambda=0.5$:
Because

$$E_0(A)=E_0(B)=E_0(T)=0$$
$$E_1(A)=\gamma\lambda E_0(A)+\mathbf{1}_{S_t=s}=1$$
$$E_1(B)=\gamma\lambda E_0(B)+\mathbf{1}_{S_t=s}=1$$

$$\delta_A=R+\gamma V_0(B)-V_0(A)=1$$

Thus

$$V_1(A)=V_0(A)+0.1\times 1\times 1=0.1$$

Then

$$\delta_B=R+\gamma V_1(A)-V_0(B)=2+0.1=2.1$$

$$V_1(B)=V_0(B)+0.1\times 2.1\times 1=0.21$$

Then

$$E_2(A)=(\gamma\lambda)^2E_1(A)+\mathbf{1}_{S_t=s}=1.25$$

$$\begin{aligned}\delta_A&=R+\gamma V_0(T)-V_1(A)\\&=1+0-0.1\\&=0.9\end{aligned}$$

Therefore

$$V(A)=V_1(A)+0.1\times 0.9\times 1.25=0.2125$$

Off-line backward-view $TD(\lambda)$ method:

$$\begin{aligned}\sum_{t=1}^{T}\alpha\delta_tE_{t(s)}&=\sum_{t=1}^{T}\alpha\left(G_t^{\lambda}-V\left(s_t\right)\right)\mathbf{1}_{S_t=s}\\&=0.1\times(0.5+0.75+1)+0.1\times 1\\&=0.1\times 2.25+0.1\\&=0.325\end{aligned}$$

$$V(A)=V_0(A)+0.325=0.325$$

P3

1.a

Ans:
Left-hand side of the equation is backward view (Off-line $TD(1)$) while right-hand side is $\lambda = 1$ forward view (Off-line every-visit constant-α Monte Carlo method). According to the original statement, they are equivalent.

1.b

Ans:
Eligibility trace

$$E_0(s) = 0$$
$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbf{1}\left(S_t = s\right)$$

Assume all the k-s, which let $S_k = s$, are in order $\{k_1, k_2, ..., k_n\}$, then we have

$$E_{k_1}(s) = (\gamma\lambda)^i E_0(s) + \mathbf{1}\left(S_t = s\right) = 1$$

Thus

$$\begin{aligned}
E_{k_n}(s) &= (\gamma\lambda)^{k_n - k_{n-1}} E_{k_{n-1}}(s) + 1 \\
&= (\gamma\lambda)^{k_n - k_{n-1}} [(\gamma\lambda)^{k_{n-1} - k_{n-2}} E_{k_{n-2}}(s) + 1] + 1 \\
&= (\gamma\lambda)^{k_n - k_{n-1}} [(\gamma\lambda)^{k_{n-1} - k_{n-2}} ((\gamma\lambda)^{k_{n-2} - k_{n-3}} E_{k_{n-3}}(s) + 1) + 1] + 1 \\
&= \cdots \\
&= (\gamma\lambda)^{k_n - k_1} E_{k_1}(s) + (\gamma\lambda)^{k_n - k_2} + ... + (\gamma\lambda)^{k_n - k_{n-1}} + 1 \\
&= (\gamma\lambda)^{k_n - k_1} + (\gamma\lambda)^{k_n - k_2} + ... + (\gamma\lambda)^{k_n - k_{n-1}} + (\gamma\lambda)^{k_n - k_n}
\end{aligned}$$

Therefore,

$$E_t(s) = \sum_{k=0}^{t} \gamma^{t-k} \cdot \mathbf{1}\left(S_k = s\right)$$

1.c

Ans:
Accumulating eligibility trace can be written as

$$E_t(s) = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \mathbf{1}(S_k = s)$$

Thus, the left-hand side of the equation

$$\begin{aligned}
\sum_{t=0}^{T-1} \alpha \delta_t E_t(s) &= \sum_{t=0}^{T-1} \alpha \delta_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \mathbf{1}(S_k = s) \\
&= \sum_{k=0}^{T-1} \alpha \delta_k \sum_{t=0}^{k} (\gamma\lambda)^{k-t} \mathbf{1}(S_k = s) \\
&= \sum_{k=t}^{T-1} \alpha \delta_k \sum_{t=0}^{T-1} (\gamma\lambda)^{k-t} \mathbf{1}(S_k = s) \\
&= \sum_{t=0}^{T-1} \alpha \mathbf{1}(S_k = s) \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k
\end{aligned}$$

Consider an individual update of the λ-return algorithm

$$\begin{aligned}
G_t - V_t(S_t) = &- V_t(S_t) + (1-\lambda)\lambda^0 \left[R_{t+1} + \gamma V_t(S_{t+1})\right] \\
&+ (1-\lambda)\lambda^1 \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_t(S_{t+2})\right] \\
&+ (1-\lambda)\lambda^2 \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V_t(S_{t+3})\right] \\
&\cdots \\
= &- V_t(s_t) \\
&+ (\gamma\lambda)^0 \left[R_{t+1} + \gamma V_t(S_{t+1}) - \gamma\lambda V_t(S_{t+1})\right] \\
&+ (\gamma\lambda)^1 \left[R_{t+2} + \gamma V_t(S_{t+2}) - \gamma\lambda V_t(S_{t+2})\right] \\
&+ (\gamma\lambda)^2 \left[R_{t+3} + \gamma V_t(S_{t+3}) - \gamma\lambda V_t(S_{t+3})\right] \\
&\cdots \\
= &\quad (\gamma\lambda)^0 \left[R_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)\right] \\
&+ (\gamma\lambda)^1 \left[R_{t+2} + \gamma V_t(s_{t+2}) - V_t(s_{t+1})\right] \\
&+ (\gamma\lambda)^2 \left[R_{t+3} + \gamma V_t(s_{t+3}) - V_t(s_{t+2})\right] \\
&\cdots \\
\approx &\sum_{k=t}^{\infty} (\gamma\lambda)^{k-t} \delta_k \\
\approx &\sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k
\end{aligned}$$

Thus, the right-hand side of the equation

$$\sum_{t=0}^{T-1} \alpha \left(G_t - V(S_t)\right) \mathbf{1}(S_t = s) = \sum_{t=0}^{T-1} \alpha \mathbf{1}(S_k = s) \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k$$

Therefore,

$$\sum_{t=0}^{T-1} \alpha \delta_t E_t(s) = \sum_{t=0}^{T-1} \alpha \left(G_t - V(S_t)\right) \mathbf{1}(S_t = s)$$

2.

Ans:
No. On-line methods will update value function once reaching state s. Backward and forward views use different information. They can't match exactly in every update.

P4

1.

Ans:

$$\hat{Q}(s,a,w) = 1$$
$$q_t^n = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^n q\left(S_{t+n}, \Lambda_{t+n}\right)$$

Thus

$q_1^1 = -1+1 = 0,\ q_1^2 = -1-1+1 = -1,\ q_1^3 = -1-1-1+1 = -2,\ q_1^4 = -1-1-1-1+0 = -4$
$q_2^1 = -1+1 = 0,\ q_2^2 = -1-1+1 = -1,\ q_2^3 = -1-1-1+0 = -3$
$q_3^1 = -1+1 = 0,\ q_3^2 = -1-1+0 = -2$
$q_4^1 = -1+0 = -1$

$$q_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} q_t^n$$

Thus, the λ-return q_t^λ corresponding to this episode is

$$\begin{aligned}
q_1^\lambda &= (1-0.5)\times(1\times 0 + 0.5\times(-1) + 0.25\times(-2)) + 0.125\times(-4) = -1\\
q_2^\lambda &= (1-0.5)\times(1\times 0 + 0.5\times(-1)) + 0.25\times(-3) = -1\\
q_3^\lambda &= (1-0.5)\times(1\times 0) + 0.5\times(-2) = -1\\
q_4^\lambda &= -1
\end{aligned}$$

2.

Ans:
For forward-view $TD(\lambda)$, target is the action-value λ-return.

$$\Delta w = \alpha\left(q_t^\lambda - \hat{q}\left(S_t, A_t, w\right)\right)\nabla_w \hat{q}\left(S_t, A_t, w\right)$$

Thus, the sequence of updates to weight w_1 is

$$\begin{aligned}
\Delta w_1^1 &= 0.5\times(-1-1)\times 1 = -1\\
\Delta w_1^2 &= 0.5\times(-1-1)\times 1 = -1\\
\Delta w_1^3 &= 0.5\times(-1-1)\times 1 = -1\\
\Delta w_1^4 &= 0.5\times(-1-1)\times 0 = 0
\end{aligned}$$

The total update to weight w_1 is -3

3.

Ans:
The $TD(\lambda)$ accumulating eligibility trace $\boldsymbol{e}_t$ when using linear value function approximation is

$$\boldsymbol{e}_t = \gamma\lambda e_{t-1} + x(s,a)$$

The sequence of eligibility traces corresponding to *right* action are

$$1,\ \frac{3}{2},\ \frac{7}{4},\ \frac{7}{8}$$

4.

Ans:
For back-view $TD(\lambda)$ we have

$$\begin{aligned}
&\delta_t = R_{t+1} + \gamma \hat{q}\left(S_{t+1}, A_{t+1}, w\right) - \hat{q}\left(S_t, \Lambda_t, w\right) \\
&\Delta w = \alpha \delta_t E_t
\end{aligned}$$

Thus, the sequence of updates to weight w_1 is

$$\begin{aligned}
\Delta w_1^1 &= 0.5 \times (-1 + 1 - 1) \times 1 = -\frac{1}{2} \\
\Delta w_1^2 &= 0.5 \times (-1 + 1 - 1) \times \frac{3}{2} = -\frac{3}{4} \\
\Delta w_1^3 &= 0.5 \times (-1 + 1 - 1) \times \frac{7}{4} = -\frac{7}{8} \\
\Delta w_1^4 &= 0.5 \times (-1 + 0 - 1) \times \frac{7}{8} = -\frac{7}{8}
\end{aligned}$$

The total update to weight w_1 is -3

5.

Ans:
When using off-line updates and linear function approximation, forward-view and backward-view $TD(\lambda)$ are equivalent to each other.