ELEN6885 Reinforcement Learning HW4

Tong Wu, tw2906

Problem 1

For the start state s, the left hand side of the equation can be written as:

$$|\mathbb{E}_\pi[G_t^{(n)}|S_t = s] - V_\pi(s)|$$

$$= |E_\pi[R_{t+1} + R_{t+2} + \ldots + \gamma^n \sum_{k=0}^{\infty} \gamma^k R_{t+n+k+1}|S_t = s] - v_\pi(s)|$$

$$= \gamma^n | \sum_{s'} P_\pi(S_{t+n} = s'|S_t = s)(V_t(s') - V_\pi(s')|$$

Then, maximising all states then the left hand size of equation could be added max_s, then the inequality equation can be written as:

$$max_s | \sum_{s'} E_\pi[G_t^{(n)}|S_t = s] - V_\pi(s)| \leq \gamma^n max_s |V_t(s) - V_\pi(s)|$$

Problem 2

Part 1

On-line updating methods updates each value function during the episode, after the increment step is calculated.

For off-line updating methods, the update of each value function is accumulated during the episode and will update the value function in the end of each episode.

Part 2

For learning rate of $\alpha = 0.1$

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

$$V_1(A) = V_0(A) + \alpha \times (1 + 2 + 1 - V_0(A)) = 0 + 0.1 \times (4 - 0) = 0.4$$

$$V_2(A) = V_1(A) + \alpha \times (1 - V_1(A)) = 0.4 + 0.1 \times (1 - 0.4) = 0.46$$

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

$$V_1(A) = \alpha \times (1 + 2 + 1 - V_0(A)) = 0.1 \times 4 = 0.4$$

$$V_2(A) = \alpha \times (1 - V_0(A)) = 0.1 \times 1 = 0.1$$

$$V(A) = V_0(A) + V_1(A) + V_2(A) = 0 + 04 + 0.1 = 0.5$$

Part 3

For learning rate of $\alpha = 0.1$

On-line TD(0) method:

$$\lambda = 0$$

$V_1(A) = V_0(A) + \alpha(1 - V_0(A)) = 0 + 0.1 \times (1 - 0) = 0.1$

$V_2(A) = V_1(A) + \alpha(1 - V_0(A)) = 0.1 + 0.1 \times (1 - 0.1) = 0.19$

Off-line TD(0) method:

$\lambda = 0$

$V_1(A) = \alpha(1 - V_0(A)) = 0.1 \times (1 - 0) = 0.1$

$V_2(A) = \alpha(1 - V_0(A)) = 0.1 \times (1 - 0) = 0.1$

$V(A) = V_1(A) + V_2(A) = 0.1 + 0.1 = 0.2$

Part 4

$\alpha = 0.1,\ \lambda = 0.5$

On-line forward-view TD(λ) method:

$G_0^\lambda = (1-\lambda)G_0^{(1)} + (1-\lambda)\lambda G_0^{(2)} + \lambda^2 G_0^{(3)} = (1-0.5) \times 1 + (1-0.5) \times 3 + 0.5^2 \times 4 = 2.25$

$G_2^\lambda = 1$

$V_1(A) = V_0(A) + \alpha(G_0^\lambda - V_0(A)) = 0 + 0.1(2.25 - 0) = 0.225$

$V_2(A) = V_1(A) + \alpha(G_2^\lambda - V_1(A)) = 0.225 + 0.1(1 - 0.225) = 0.3025$

Off-line forward-view TD(λ) method:

$V_1(A) = \alpha(G_0^\lambda - V_0(A)) = 0.1 \times (2.25 - 0) = 0.225$

$V_2(A) = \alpha(G_2^\lambda - V_0(A)) = 0.1 \times (1 - 0) = 0.1$

$V(A) = V_1(A) + V_2(A) = 0.1 + 0.225 = 0.325$

Part 5

$\alpha = 0.1,\ \lambda = 0.5$

On-line backward-view TD(λ) method:

$V_1(A) = V_0(A) + \alpha(1 + V_0(B) - V_0(A))E_0(A) = 0.1$

$V_1(B) = V_0(B) + \alpha(1 + V_0(B) - V_0(A))E_0(B) = 0$

$V_2(A) = V_1(A) + \alpha(2 + V_1(B) - V_1(A))E_1(A) = 0.1 + 0.1 \times 2.1 \times 0.5 = 0.205$

$V_2(B) = V_1(B) + \alpha(2 + V_1(A) - V_1(B))E_1(B) == 0.21$

$V_3(A) = V_2(A) + \alpha(1 + V(T) - V_2(A))E_2(A) = 0.304375$

Off-line backward-view TD(λ) method:

$V_1(A) = \alpha(1 + V(B) - V_0(A))E_0(A) = 0.1$

$V_2(A) = \alpha(2 + V(A) - V_0(A))E_1(A) = 0.1$

$V_3(A) = \alpha(1 + V(T) - V_0(A))E_2(A) = 0.125$

$V(A) = V_0(A) + V_1(A) + V_2(A) + V_3(A) = 0 + 0.1 + 0.1 + 0.125 = 0.325$

Problem 3

Part 1

Section a

The update to state s, the value of TD(1) should be equal to the every visit. Where the $\lambda = 1$ so according to the update equation:

$\sum_{t=0}^{T-1} \alpha \delta_t E_t(s)$

the total update should be equivalent to the right hand side of the equation:

$\sum_{t=0}^{T-1} \alpha (G_t - V(S_t)) 1(S_t = s)$

Section b

$E_0(s) = 0$

$E_t(s) = \gamma \lambda E_{t-1}(s) + 1(S_t = s)$

$$\begin{aligned} E_t(s) &= \sum_{i=1}^{n} 1 \times \gamma^{t-t_i} \\ &= \sum_{k=0}^{t} \gamma^{t-k} \times 1(S_k = s) \end{aligned}$$

Section c

$$\begin{aligned} \sum_{t=0}^{T-1} \alpha \delta_t E_t(s) &= \sum_{t=0}^{T-1} \alpha \delta_t \sum_{k=0}^{t} \gamma^{t-k} \times 1(S_k = s) \\ &= \sum_{k=0}^{T-1} 1(S_k = s) \sum_{t=k}^{T-1} \alpha \gamma^{t-k} (R_{t+1} + \gamma V(S_{t+1} - V(S_t))) \\ &= \sum_{t=0}^{T-1} \alpha \gamma^{t-k} R_{t+1} + \sum_{t=0}^{T-1} \alpha \gamma^{t-k+1} V(S_{t+1}) - \sum_{t=0}^{T-1} \alpha \gamma^{t-k} V(S_t) \\ &= \alpha (G_k - V(S_k)) \end{aligned}$$

Where the $\alpha(G_k - V(S_k))$ could be insert into:

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

Part 2

It is not possible to construct a version of on-line TD(λ) method that matches the on-line λ-return algorithm exactly. Since the λ-return use the future information which is can not get from the current step.

Problem 4

Part 1

$q_1 = (1-0.5) \times (0 + (-1) \times 0.5 + (-2) \times 0.25) + (-4) \times 0.125 = 0.5 \times (-0.5 - 0.5) - 0.5 = -1$

$q_2 = (1-0.5) \times (0 + (-1) \times 0.5) + (-3) \times 0.25 = 0.5 \times (-0.5) - 0.75 = -1$

$q_3 = (1-0.5) \times 0 + (-2) \times 0.5 = -1$

$q_4 = -1$

Part 2

According to the equation:

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

$\Delta w_1^1 = \alpha(q_t^\lambda - \hat{q}(S_t, A_t, w))\nabla_w \hat{q}(S_t, A_t, w) = 0.5 \times (-1-1) \times 1 = -1$

$\Delta w_1^2 = \alpha(q_t^\lambda - \hat{q}(S_t, A_t, w))\nabla_w \hat{q}(S_t, A_t, w) = -1$

$\Delta w_1^3 = \alpha(q_t^\lambda - \hat{q}(S_t, A_t, w))\nabla_w \hat{q}(S_t, A_t, w) = -1$

$\Delta w_1^4 = \alpha(q_t^\lambda - \hat{q}(S_t, A_t, w))\nabla_w \hat{q}(S_t, A_t, w) = 0$

Part 3

Where the linear value function approximation in trace e_t is

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

The sequence of eligibility traces corresponding to right action should be:

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

Part 4

The update should be:

$\Delta w_1^1 = \alpha\delta_1 e_1 = 0.5 \times (-1) \times 1 = -0.5$

$\Delta w_1^2 = \alpha\delta_2 e_2 = (-0.5) \times \frac{3}{2} = -\frac{3}{4}$

$\Delta w_1^3 = \alpha\delta_3 e_3 = (-0.5) \times \frac{7}{4} = -\frac{7}{8}$

$\Delta w_1^4 = \alpha\delta_4 e_4 = (-1) \times \frac{7}{8} = -\frac{7}{8}$

Part 5

Forward-view and backward-view TD(λ) is equivalent to each other.