Attempts to exercise in Reinforcement Learning book Chapter 7

Mengliao Wang

August 20, 2017

Exercise 7.1:

First of all, according to definition of G_t and $G_{t:t+n}$ we have the following equations:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + ... + \gamma^{T-t-1} R_T \\ G_{t+n} &= R_{t+n+1} + \gamma R_{t+n+2} + ... + \gamma^{T-t-n-1} R_T \\ G_{t:t+n} &= R_{t+1} + \gamma R_{t+2} + ... + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n}) \end{aligned}$$

From these equation we can have

$$G_t = G_{t:t+n} - \gamma^n V_{t+n-1}(S_{t+n}) + \gamma^n G_{t+n} \tag{1}$$

Then by applying equation 1, the difference between G_t and $V_{t+n-1}(S_t)$ can be written as:

$$\begin{aligned} G_t - V_{t+n-1}(S_t) &= G_{t:t+n} - \gamma^n V_{t+n-1}(S_{t+n}) + \gamma^n G_{t+n} - V_{t+n-1}(S_t) && (2) \\ &= [G_{t:t+n} - V_{t+n-1}(S_t)] + \gamma^n [G_{t+n} - V_{t+n-1}(S_{t+n})] && (3) \\ &= \delta_t + \gamma^n [\delta_{t+n} + \gamma^n [G_{t+2n} - V_{t+2n-1}(S_{t+2n})]] && (4) \\ &= \sum_{k=0}^{t+kn<T} \gamma^{kn} \delta_{t+kn} && (5) \end{aligned}$$

Exercise 7.3:

A larger random walk task will make the simulated sequence significantlly longer, which allows us to run TD approach on a bigger n. If we change the number of states to be smaller, it will be beneficial for smaller n values, since less simulated sequences will have more steps than n. However, I do not think changing the left-side outcome from 0 to -1 would make a differece in the best value of n here.

Exercise 7.4:

Here is the pseudocode for per-reward off-policy state value algorithm:

Initialize:
  an arbitrary behaviour policy b such that $b(a|s) > 0$
  $V(s)$ arbitrarily
  π to be a fixed given policy
  step size $\alpha \in (0, 1]$, small $\epsilon > 0$, a positive integer n
All store and access opeartions (for S_t, A_t , and R_t) can take their index mode n
repeat(For each episode):
  Initialize and store $S_0 \neq$ terminal
  Select and store an action $A_0 \sim b(\cdot|S_0)$

$T \leftarrow \infty$
for $t = 0, 1, 2, ...$ **do**
if $t < T$ **then**
Take action A_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**
$T \leftarrow t + 1$
else
Select and store an action $A_{t+1} \sim b(\cdot|S_{t+1})$
$\tau \leftarrow t - n + 1$
if $\tau \geq 0$ **then**
$G \leftarrow 0$
$\rho \leftarrow 1$
for $k = \tau + 1, \tau + 2, ..., \min(\tau + n, T)$ **do**
$\phi \leftarrow \rho \cdot [1 - \frac{\pi(A_k|S_k)}{b(A_k|S_k)}]$
$\rho \leftarrow \rho \cdot \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$
$G \leftarrow G + \gamma^{k-\tau-1}\rho R_k + \gamma^{k-\tau-1}\phi V(S_k)$
if $\tau + n < T$ **then**
$G \leftarrow G + \gamma^n V(S_{\tau+n})$
$V(S_\tau) \leftarrow V(S_\tau) + \alpha[G - V(S_\tau)]$
if $\tau = T - 1$ **then**
Break For Loop
until $True$

Exercise 7.5:

Similiar to above, here is the pseudocode for per-reward off-policy action value algorithm:
Initialize:
an arbitrary behaviour policy b such that $b(a|s) > 0$
$Q(s, a)$ arbitrarily
π to be a fixed given policy, or ϵ-greedy with respect to $Q(s, a)$
step size $\alpha \in (0, 1]$, small $\epsilon > 0$, a positive integer n
All store and access opeartions (for S_t, A_t , and R_t) can take their index mode n
repeat(For each episode):
Initialize and store $S_0 \neq$ terminal
Select and store an action $A_0 \sim b(\cdot|S_0)$
Store $Q(S_0, A_0)$ as Q_0
$T \leftarrow \infty$
for $t = 0, 1, 2, ...$ **do**
if $t < T$ **then**
Take action A_t
Observe and store the next reward as R_{t+1} and the next state as S_{t+1}
Store $\sum_a \pi(a|S_t)Q(S_t, a)$ as $\bar{Q}_{t+1}$
if S_{t+1} is terminal **then**
$T \leftarrow t + 1$
else
Select and store an action $A_{t+1} \sim b(\cdot|S_{t+1})$
$\tau \leftarrow t - n + 1$
if $\tau \geq 0$ **then**
$G \leftarrow 0$

$\rho \leftarrow 1$
for $k = \tau+1, \tau+2, ..., \min(\tau+n-1, T)$ **do**
$\phi \leftarrow \rho \cdot [1 - \frac{\pi(A_{k-1}|S_{k-1})}{b(A_{k-1}|S_{k-1})}]$
$G \leftarrow G + \gamma^{k-\tau-1}\rho R_k + \gamma^{k-\tau}\phi\bar{Q}_k$
$\rho \leftarrow \rho \cdot \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$
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 $\tau = T - 1$ **then**
Break For Loop
until $True$

Exercise 7.6:

According to definition 7.10, we can have the following:

$$\begin{aligned}
G_t - V(S_t) &= \rho_t(R_{t+1} + \gamma G_{t+1}) + (1-\rho_t)V(S_t) - V(S_t) \\
&= \rho_t[R_{t+1} + \gamma G_{t+1} - V(S_t)] \\
&= \rho_t[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] + \gamma\rho_t[G_{t+1} - V(S_{t+1})] \\
&= \rho_t\delta_t + \gamma\rho_t\rho_{t+1}\delta_{t+1} + \gamma^2\rho_t\rho_{t+1}[G_{t+2} - V(S_{t+2})] \\
&= \rho_{t:t}\delta_t + \gamma\rho_{t:t+1}\delta_{t+1} + \gamma^2\rho_{t:t+2}\delta_{t+2} + ... + \gamma^{T-t-1}\rho_{t:T-1}\delta_{T-1} + \gamma^{T-t}\rho_{t:T}[G_T - V(S_T)] \\
&= \sum_{k=t}^{T-1}\gamma^{k-t}\rho_{t:k}\delta_k
\end{aligned}$$

Here we define $\rho_{t:k} = \prod_{l=t}^{k}\rho_l$

Exercise 7.7:

Suppose the definition 7.11 in the book is written as $G_{t:h} = R_{t+1} + \gamma(\rho_{t+1}G_{t+1:h} + (1-\rho_{t+1})\bar{Q}_{t+1})$, then we can have:

$$\begin{aligned}
G_t - Q_t &= R_{t+1} + \gamma[\rho_{t+1}G_{t+1} + (1-\rho_{t+1})\bar{Q}_{t+1}] - Q_t \\
&= (1-\rho_{t+1})(R_{t+1} + \gamma\bar{Q}_{t+1} - Q_t) + \rho_{t+1}R_{t+1} - \rho_{t+1}Q_t + \gamma\rho_{t+1}G_{t+1} \\
&= (1-\rho_{t+1})\delta'_t + \rho_{t+1}(R_{t+1} + \gamma Q_{t+1} - Q_t) + \gamma\rho_{t+1}G_{t+1} - \gamma\rho_{t+1}Q_{t+1} \\
&= (1-\rho_{t+1})\delta'_t + \rho_{t+1}\delta_t + \gamma\rho_{t+1}(G_{t+1} - Q_{t+1}) \\
&= (1-\rho_{t+1})\delta'_t + \rho_{t+1}\delta_t + \gamma\rho_{t+1}(1-\rho_{t+2})\delta'_{t+1} + \gamma\rho_{t+1}\rho_{t+2}\delta_{t+1} + \gamma^2\rho_t\rho_{t+1}(G_{t+2} - Q_{t+2}) \\
&= (1-\rho_{t+1})\delta'_t + \rho_{t+1}\delta_t + \gamma\rho_{t+1}(1-\rho_{t+2})\delta'_{t+1} + \gamma\rho_{t+1}\rho_{t+2}\delta_{t+1} + ... \\
&\quad + \gamma^{T-t-1}\rho_{t+1:T-1}(1-\rho_T)\delta'_{T-1} + \gamma^{T-t-1}\rho_{t+1:T}\delta_{T-1} + \gamma^{T-t}\rho_{t+1:T}(G_T - Q_T) \\
&= \sum_{k=t}^{T-1}\gamma^{k-t}\rho_{t+1:k}(1-\rho_{k+1})\delta'_k + \sum_{k=t}^{T-1}\gamma^{k-t}\rho_{t+1:k+1}\delta_k
\end{aligned}$$

Here for convenience we write $Q(S_t, A_t)$ as Q_t. Similar to above we define $\rho_{t:k} = \prod_{l=t}^{k}\rho_l$, where $\rho_{t:t-1} = 1$. Also we have $\delta'_t = R_{t+1} + \gamma\sum_a \pi(a|S_{t+1})Q(S_{t+1}, a) - Q(S_t, A_t)$, and $\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$