Exercise Set 2

6.0 VU AKNUM Reinforcement Learning

May 21, 2020

Exercise (4.1) By definition the value functions are undiscounted. Since $p(T, -1|11, \texttt{down}) = 1$, the deterministic successor state $s' = T$. By definition, $v(T) = 0$. Thus, the value $q_\pi(11, \texttt{down}) = -1 + 0 = -1$.

Exercise (4.2) The value is calculated using $v_\pi(15) = \sum_a \pi(a|15) - 1 + v_\pi(s')$. Using the state value table in Figure 4.1, this leads to

$v_\pi(15) = .25 * (-1 - 22) + .25 * (-1 - 20) + .25 * (-1 - 14) + .25 * (-1 + v_\pi(15)) \implies$
$v_\pi(15) = -14.75 + .25 * (-1 + v_\pi(15)) = -15 + .25v_\pi(15) \implies$
$.75v_\pi(15) = -15 \implies$
$v_\pi(15) = -20$

If a new state $13'$ is introduced with $p(15, -1|13', \texttt{down}) = 1$, the value function would not change since $v_\pi(15) = v_\pi(13)$ and $p(15, -1|13', \texttt{down}) = p(13, -1|13, \texttt{down}) = 1$.

Exercise (4.3) For 4.3 and 4.4:
$q_\pi(s, a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$
$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$
$= \mathbb{E}_\pi\Big[R_{t+1} + \gamma \sum_{a',s'} q_\pi(s', a') \Bigm| S_t = s, A_t = a\Big]$
$= \sum_{s',r} p(s', r|s, a)[r + \gamma \sum_{a'} \pi(a'|s')q_\pi(s', a')]$

For 4.5:
$q_{k+1}(s, a) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = s]$
$= \sum_{s',r} p(s', r|s, a)[r + \sum_{a'} \pi(a'|s')q_k(s', a')]$

Exercise (4.5) Solution can be seen in algorithm 1.

Exercise (4.6) For step 3, one would determine if the policy is stable only with greedy, not exploritative actions. Also, since the policy is stochastic, $old{-}action$ would be chosen diferently and the update to $\pi(s|a)$ would take the ϵ-soft characteristic into account.
For step 2, the value updates would have to deal with a stochastic policy, not with a deterministic one. Also, the $\delta < \theta$ comparison should respect the exploritative aspect.
For step 1, ϵ needs to be defined as parameter and π needs to be a stochastic ϵ-soft policy.

Algorithm 1 Policy iteration using action values

Input: $\theta > 0$

Initialize $Q(s,a) \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$.
$\delta \leftarrow 0$
while $\delta < \theta$ **do**
  for $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$ **do**
    $q \leftarrow Q(s,a)$
    $Q(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a' \in \mathcal{A}(s')} \pi(a'|s')Q(s',a')]$
    $\delta \leftarrow \max(\delta, |q - Q(s,a)|)$
  end for
end while
$policy - stable \leftarrow true$
for $s \in \mathcal{S}$ **do**
  $old - action \leftarrow \pi(s)$
  $\pi(s) \leftarrow \arg\max_a Q(s,a)$
  if $old - action$ and $\pi(s)$ are not equi-probable **then**
    $policy - stable \leftarrow false$
  end if
end for
if $policy - stable$ **then**
  return V, π
else
  Go to policy evaluation
end if
return false

Exercise (4.10) $q_{k+1}(s) \doteq \max_a \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_k(S_{t+1}, a') \mid S_t = s, A_t = a]$
$= \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} q_k(s', a')]$

Exercise (5.3) The backup diagram for the MC estimation of $q_\pi(s)$ is similar to the MC estimation of $v_\pi(s)$ depicted on page 95. The major difference is that the backup diagram starts with an action node instead of a state node.

Exercise (5.9) The sample average update rule is as follows:

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n) \tag{1}$$

This can easily be adapted by replacing the rewards with returns:

$$V_{n+1} = V_n + \frac{1}{n}(G_n - V_n) \tag{2}$$

The first-visit Monte Carlo prediction algorithm can be seen in Algorithm 2.

Algorithm 2 First-visit MC policy evaluation using sample averages

---

Input: π
  Initialize $V(s) \forall s \in \mathcal{S}$ arbitrarily
  $N(s) \leftarrow 0 \forall s \in \mathcal{S}$
  while true **do**
    Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    for all steps in episode desc **do**
      $G \leftarrow \gamma G + R_{t+1}$
      if $S_t \notin \{S_0, \ldots, S_{t-1}\}$ **then**
        $N(S_t) \leftarrow N(S_t) + 1$
        $V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}(G - V(S_t))$
      end if
    end for
  end while

Exercise (5.10) The value estimate for weighted importance sampling is:

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \tag{3}$$

Let $C_n = \sum_{k=1}^{n} W_k$, then the update rule is derived by:
$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}$
$= \frac{1}{C_n} \left(\sum_{k=1}^{n} W_k G_k\right)$
$= \frac{1}{C_n} \left(W_n G_n + \sum_{k=1}^{n-1} W_k G_k\right)$
$= \frac{1}{C_n} \left(W_n G_n + (C_{n-1}) \frac{1}{C_{n-1}} \sum_{k=1}^{n-1} W_k G_k\right)$

$= \frac{1}{C_n}(W_n G_n + C_{n-1} V_n)$
$= \frac{1}{C_n}(W_n G_n + (C_n - W_n) V_n)$
$= \frac{1}{C_n}(W_n G_n + C_n V_n - W_n V_n)$
$= V_n + \frac{1}{C_n}(W_n G_n - W_n V_n)$
$= V_n + \frac{W_n}{C_n}(G_n - V_n)$

Exercise (5.13) The importance weighted reward is given by:

$$\rho_{t:T-1} R_{t+1} = \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_{T-1}|S_{T-1})}{b(A_{T-1}|S_{T-1})} R_{t+1} \tag{4}$$

The expectation of this is:
$\mathbb{E}[\rho_{t:T-1} R_{t+1}] = \mathbb{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_{T-1}|S_{T-1})}{b(A_{T-1}|S_{T-1})} R_{t+1}\right]$
$= \mathbb{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right] \mathbb{E}\left[\frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})}\right] \cdots \mathbb{E}\left[\frac{\pi(A_{T-1}|S_{T-1})}{b(A_{T-1}|S_{T-1})}\right] \mathbb{E}[R_{t+1}]$
Using Equation 5.13, we know that $\mathbb{E}\left[\frac{\pi(A_k|S_k)}{b(A_k|S_k)}\right] = 1 \forall k > t$. Thus, the expectation simplifies to:
$\mathbb{E}[\rho_{t:T-1} R_{t+1}] = \mathbb{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right] 1 \ldots 1 \mathbb{E}[R_{t+1}] = \mathbb{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right] \mathbb{E}[R_{t+1}] = \mathbb{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_{t+1}\right]$

Exercise (5.14) The weighted importance sampling estimator is given by:

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \tilde{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \tilde{G}_{t:T(t)}\right)}{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1}\right)} \tag{5}$$

In order to adapt this for action values, the importance ratios have to be shifted. The ratio is defined by:

$$\rho_{t:T-1} \doteq \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)} \tag{6}$$

Since action A_t is already defined for $q(s, a)$, the estimator has to be adapted to:

$$Q(s,a) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t+1:h-1} \tilde{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t+1:T(t)-1} \tilde{G}_{t:T(t)}\right)}{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t+1:h-1} + \gamma^{T(t)-t-1} \rho_{t+1:T(t)-1}\right)} \tag{7}$$

The resulting algorithm can be seen in Algorithm 3. It is highly likely that the algorithm is not optimal and could be improved towards linear time.

Algorithm 3 Off-policy MC control using truncated weighted-importance sampling

Input: π

for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$ **do**
  $Q(s,a) \in \mathcal{R}$ arbitrarily
  $P(s,a) \leftarrow 0$
  $R(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(s,a)$
end for
while true **do**
  $b \leftarrow$ any soft policy
  Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G_T \leftarrow 0$
  $\rho \leftarrow 1$
  $W(T) \leftarrow 1$
  for all steps in episode desc **do**
    $G_T \leftarrow \gamma G_T + R_{t+1}$
    $R(s,a) \leftarrow R(s,a) + (1-\gamma)\sum_{h=t+1}^{T-1} W(h) \sum_{k=1}^{h} R_k + \gamma^{T-t-1}\rho G_T$
    $P(s,a) \leftarrow P(s,a) + (1-\gamma)\sum_{h=t+1}^{T-1} W(h) + \gamma^{T-t-1}\rho$
    $Q(S_t, A_t) \leftarrow \frac{R(s,a)}{P(s,a)}$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$
    if $A_t \neq \pi(S_t)$ **then** exit loop
    end if
    $W(t-1) \leftarrow W(t)\gamma^{t-1}\frac{1}{b(A_t|S_t)}$
    $\rho \leftarrow \rho\frac{1}{b(A_t|S_t)}$
  end for
end while