Suggested Solutions to Reinforcement Learning

Renjie Lu

March 13, 2018

1 Part I

The basic idea: given a police π, we obtain an estimate of $V_\pi(s)$; given V_π, we obtain an estimate of $Q(s,a)$, then update the current police π via $Q(s,a)$. If we know the whole structure of the Markov Decision Process (MDP), namely, $\Pr(S_{t+1}, R_{t+1}|S_t, A_t)$ is known, it is easy to obtain V_π and Q. However, it is impossible in reality. Hence, this problem lets us to develop the MC and TD algorithms to estimate V_π and Q under an empirical environment.

Note that initial action-value estimate may encourage the exploration in the stationary problem (but the nonstationary case), see Section 2.6.

1.1 Chapter 2

1. Exercise 2.1

   Denote by A^* the greedy action,

$$\begin{aligned}&\Pr(A^*)\\=&\Pr(A^*|\text{random})\Pr(\text{random})+\Pr(A^*|\text{nonrandom})\Pr(\text{nonrandom})\\=&1/4+1/2=3/4.\end{aligned}$$

2. Exercise 2.2

   It is easy to show by writing down the $Q_t(A_i), i=1,\dots,4$ for each time step.

3. Exercise 2.3

   Skip.

4. Exercise 2.4

   Skip.

5. Exercise 2.5

   See the supplement.

6. Exercise 2.6

   The highly optimistic initial values encourage the exploration due to the expectation of each bandit is zero. So, low optimal action percentage is trivial. The spike can be explained by the influence of the initial values decreases.

7. Exercise 2.7

   It is easy to show, and thus omitted.

8. Exercise 2.8

   (a) no information case. Police π: we select action 1 with probability p. Then

$$\begin{aligned}\mathbb{E}_\pi\left(R_t\right) &= \mathbb{E}_\pi\left(R_t|\text{A}\right)\Pr\left(\text{A}\right) + \mathbb{E}_\pi\left(R_t|\text{B}\right)\Pr\left(\text{B}\right)\\ &= 0.5\left(0.2-0.1p\right) + 0.5\left(0.8+0.1p\right)\\ &= 0.5.\end{aligned}$$

   This shows that if we do not know the information in advance, the best expectation is equal to 0.5. This implies the return is independent of the police.

   (b) Information case. Police π: we always select action 2 if we in case A, otherwise we select action 1. Then

$$\begin{aligned}\mathbb{E}_\pi\left(R_t\right) &= \mathbb{E}_\pi\left(R_t|\text{A}\right)\Pr\left(\text{A}\right) + \mathbb{E}_\pi\left(R_t|\text{B}\right)\Pr\left(\text{B}\right)\\ &= 0.1 + 0.45 = 0.55.\end{aligned}$$

9. Exercise 2.9

   See the supplement. (problem may exists, incomplete.)

1.2 Chapter 3

Note that

$$
\begin{aligned}
\mathbb{E}_\pi\left(R_{t+1}|S_t=s\right) &= \sum_r r\Pr_\pi\left(R_{t+1}=r|S_t=s\right)\\
&= \sum_r\sum_{s'} r\Pr_\pi\left(S_{t+1}=s', R_{t+1}=r|S_t=s\right)\\
&= \sum_r\sum_{s'}\sum_a r\Pr_\pi\left(S_{t+1}=s', R_{t+1}=r, A_t=a|S_t=s\right)\\
&= \sum_r\sum_{s'}\sum_a r\Pr\left(S_{t+1}=s', R_{t+1}=r|S_t=s, A_t=a\right)\pi\left(A_t=a|S_t=s\right)\\
&= \sum_a \pi\left(A_t=a|S_t=s\right)\sum_{r,s'} r\Pr\left(S_{t+1}=s', R_{t+1}=r|S_t=s, A_t=a\right)\\
&= \sum_a \pi\left(a|s\right)\sum_{r,s'} r\Pr\left(s', r|s, a\right)
\end{aligned}
$$

and by $\{S_t=s\}\subset\sigma\left(S_{t+1}\right)$ since for each t, state space is equal,

$$
\begin{aligned}
\mathbb{E}_\pi\left(R_{t+2}|S_t=s\right) &= \mathbb{E}_\pi\left\{\mathbb{E}_\pi\left(R_{t+2}|S_{t+1}\right)|S_t=s\right\}\\
&= \sum_{s'} f\left(s'\right)\Pr_\pi\left(S_{t+1}=s'|S_t=s\right), \quad \text{let } \mathbb{E}_\pi\left(R_{t+2}|S_{t+1}\right)=f\left(S_{t+1}\right)\\
&= \sum_{s'} f\left(s'\right)\sum_r\sum_a \Pr_\pi\left(S_{t+1}=s', R_{t+1}=r, A_t=a|S_t=s\right)\\
&= \sum_{s'}\sum_r\sum_a f\left(s'\right)\Pr\left(S_{t+1}=s', R_{t+1}=r|S_t=s, A_t=a\right)\pi\left(A_t=a|S_t=s\right)\\
&= \sum_a \pi\left(A_t=a|S_t=s\right)\sum_{r,s'} f\left(s'\right)\Pr\left(S_{t+1}=s', R_{t+1}=r|S_t=s, A_t=a\right)\\
&= \sum_a \pi\left(a|s\right)\sum_{r,s'} f\left(s'\right)\Pr\left(s', r|s, a\right)\\
&= \sum_a \pi\left(a|s\right)\sum_{r,s'} \mathbb{E}_\pi\left(R_{t+2}|S_{t+1}=s'\right)\Pr\left(s', r|s, a\right).
\end{aligned}
$$

Also, it is reasonable that (can be proved)

$$\Pr_\pi\left(S_{t+1}=s',R_{t+1}=r|S_t=s,A_t=s\right)=\Pr\left(s',r|s,a\right).$$

In R.L., return may not mean reward.

1. Exercise 3.1

   Skip.

2. Exercise 3.2

   MDP framework may not work on the goal-learning task without interaction.

3. Exercises 3.3

   Skip.

4. Exercise 3.4

   Suppose our stochastic police is $\pi\left(a|s\right)$, all states are finite and $S_t=s$ is given,

$$\begin{aligned}\mathbb{E}_\pi\left(R_{t+1}|S_t=s\right)&=\sum_r r\Pr_\pi\left(r|s\right)\\&=\sum_r r\sum_{s',s,a}\Pr\left(r,s'|s,a\right)\pi\left(a|s\right).\end{aligned}$$

5. Exercise 3.5 (unclear) ???

s	a	$s^{'}$	r	$\Pr(s', r \| s, a)$
high	search	high	r_{reseach}	α
high	search	low	r_{reseach}	$1-\alpha$
low	search	high	-3	$1-\beta$
low	search	low	r_{reseach}	β
high	wait	high	r_{wait}	1
high	wait	low	r_{wait}	0
low	wait	high	r_{wait}	0
low	wait	low	r_{wait}	1
low	recharge	high	0	1
low	recharge	low	0	0

6. Exercise 3.6 (unclear)

7. Exercise 3.7

   Suppose T is the terminal time. Then,

$$\begin{aligned} G_t &= \sum_{i=1}^{T-t} R_{t+i}\gamma^{i-1} \\ &= -\gamma^{T-t-1}, \end{aligned}$$

   where γ is the discount rate. It can be seen that G_t at each time is not a constant.

8. Exercise 3.8

   From the expression of G_t in Exercise 3.7, if there is no discount rate, we have G_t is always equal to -1. Such return design does not encourage to short the escape time. Therefore, we need to add the discount rate.

9. Exercise 3.9

Suppose $G_T = 0$, where T is the terminal time.

$$\begin{aligned} G_4 &= R_5 + 0.5 G_5 = 2, \\ G_3 &= R_4 + 0.5 G_4 = 3 + 1 = 4, \\ &\dots \\ G_0 &= \dots . \end{aligned}$$

10. Exercise 3.10 and 3.11

    Skip.

11. Exercise 3.12

$$\begin{aligned} 0.7 = v_\pi(s) &= 0.25 \left(\sum_{s', r} \Pr(s', r|s, a)\left(r + 0.9 v_\pi(s')\right) \right) \\ &= 0.25 \left(0.9 \sum_{s', r} \Pr(s', r|s, a)\, v_\pi(s') \right) \\ &= 0.25 \times 0.9 \times 3 = 0.675 \approx 0.7. \end{aligned}$$

12. Exercise 3.13.

$$\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi\left(R_{t+1}|S_t=s,A_t=a\right)+\gamma\mathbb{E}_\pi\left(\sum_{k=0}^\infty \gamma^k R_{t+k+2}|S_t=s,A_t=a\right)\\
&= \sum_{r,s'} r\Pr\left(s',r|r,a\right)+\gamma\mathbb{E}_\pi\left(q_\pi\left(S_{t+1},A_{t+1}\right)|s,a\right)\\
&= \sum_{r,s'} r\Pr\left(s',r|r,a\right)+\gamma\sum_{a',s'} q_\pi\left(s',a'\right)\Pr_\pi\left(S_{t+1}=s',A_{t+1}=a'|s,a\right)\\
&= \sum_{r,s'} r\Pr\left(s',r|r,a\right)+\gamma\sum_{a',s'} q_\pi\left(s',a'\right)\Pr_\pi^{(1)}\left(A_{t+1}=a'|s,a,S_{t+1}=s'\right)\Pr_\pi^{(2)}\left(S_{t+1}=s'|s,a\right)\\
&= \sum_{r,s'} r\Pr\left(s',r|r,a\right)+\gamma\sum_{a',s'} q_\pi\left(s',a'\right)\pi\left(a'|s'\right)\sum_r \Pr\left(s',r|s,a\right)\\
&= \sum_{r,s'}\left[r+\sum_{a'} q_\pi\left(s',a'\right)\pi\left(a'|s'\right)\right]\Pr\left(s',r|r,a\right).
\end{aligned}$$

13. Exercise 3.14

$$v_c(s)=\mathbb{E}\left[\sum_{k=0}^\infty \gamma^k\left(R_{t+k+1}+c\right)|S_t=s\right]=v(s)+\frac{c}{1-\gamma}.$$

14. Exercise 3.15

For episodic task, adding a positive constant will encourage to extend the duration of task.

$$\mathbb{E}\left[\sum_{k=0}^T \left(R_{t+k+1}+c\right)|S_t=s\right]=v(s)+Tc,$$

where T is the terminal time. We cam maximize the previous equation by increasing T.

15. Exercise 3.16

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi q_\pi(s,A)\\
&= \sum_a q_\pi(s,a)\,\pi(a|s).
\end{aligned}$$

16. Exercise 3.17

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi\left(R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a\right) \\ &= \mathbb{E}\left\{\left[R_{t+1}+\gamma\mathbb{E}_\pi\left(G_{t+1}|S_{t+1}\right)\right]|S_t=s,A_t=a\right\} \\ &= \mathbb{E}\left[R_{t+1}+\gamma v_\pi\left(S_{t+1}\right)|S_t=s,A_t=a\right] \\ &= \sum_{r,s^{'}}\left[r+\gamma v_\pi\left(s^{'}\right)\right]\Pr\left(s^{'},r|s,a\right). \end{aligned}$$

17. Exercise 3.18-3.19

Skip.

18. Exercise 3.20

Note that $v_*(s)$ has been given in the book, we just use it. Here, we take $S_t=h$ and $A_t=s$ for example,

$$\begin{aligned} q_*(h,s) &= \mathbb{E}\left[R_{t+1}+\gamma v_*\left(S_{t+1}\right)|S_t=h,A_t=s\right] \\ &= r_s+\alpha v_*(h)+(1-\alpha)v_*(l). \end{aligned}$$

19. Exercise 3.21 (π_* may be reverse)

d

20. Exercise 3.22

If $\gamma=0$: π: at the top state, we turn to left since

$$G_t=R_{t+1}$$

.

If $\gamma = 0.9$: π: at the top state, we turn to right since

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots$$

and $2 \times 0.9 \times (1 + 0.9^2 + \dots) > 1 + 0.9^2 + 0.9^4 + \dots$.

If $\gamma = 0.5$: π: randomly choose left or right in an equal way.

21. Exercise 3.23-3.24

    From Exercise 3.16-3.17,

$$\begin{aligned} v_* (s) &= \max_a q_* (s, a), \\ q_* (s, a) &= \sum_{r, s^{'}} \left[r + \gamma v_* \left(s^{'} \right) \right] \Pr \left(s^{'}, r | s, a \right). \end{aligned}$$

22. Exercise 3.25-3.26

$$\begin{aligned} \pi_* (s) &= \operatorname*{argmax}_a q_* (s, a) \\ &= \operatorname*{argmax}_a \sum_{r, s^{'}} \left[r + \gamma v_* \left(s^{'} \right) \right] \Pr \left(s^{'}, r | s, a \right). \end{aligned}$$

1.3 Chapter 4

1. Exercise 4.1

   From Exercise 3.17, we have

$$q_\pi (s, a) = \sum_{r, s^{'}} \left[r + \gamma v_\pi \left(s^{'} \right) \right] \Pr \left(s^{'}, r | s, a \right).$$

Hence,

$$\begin{aligned} q_\pi\left(11,\text{down}\right) &= \sum_{r,s^{'}}\left[r+v_\pi\left(s^{'}\right)\right]\Pr\left(s^{'},r|11,\text{down}\right) \\ &= \sum_{r,s^{'}}\left[r+v_\pi\left(s^{'}\right)\right]\Pr\left(s^{'},r|11,\text{down}\right) \\ &= \left(-1+0\right)\times 1=-1. \end{aligned}$$

2. Exercise 4.2

   (a) $v_\pi\left(15\right)=0$ since we cannot go to this state if all transitions are unchanged.

   (b) Due to

$$\begin{aligned} v_\pi\left(15\right) &= \sum_a \pi\left(a|15\right)\sum_{s^{'},r}\Pr\left(s^{'},r|15,a\right)\left[r+v_\pi\left(s^{'}\right)\right] \\ &= 0.25\left[-4+\sum_{i=12}^{15}v_\pi\left(i\right)\right] \\ &= -20. \end{aligned}$$

3. Exercise 4.3

   From Exercise 3.13, we have

$$q_\pi\left(s,a\right)=\sum_{r,s'}\left[r+\sum_{a'}q_\pi\left(s',a'\right)\pi\left(a'|s'\right)\right]\Pr\left(s',r|r,a\right).$$

   Then,

$$q_{k+1}\left(s,a\right)=\sum_{r,s'}\left[r+\sum_{a'}q_k\left(s',a'\right)\pi\left(a'|s'\right)\right]\Pr\left(s',r|s,a\right).$$

4. Exercise 4.4

   Skip.

5. Exercise 4.5

   See supplement.

6. Exercise 4.6

   Suppose we only consider deterministic police $\pi(s)$. From Exercise 4.3, given a police $\pi(s)$, we have an update equation for $q_k(s,a)$,

$$q_{k+1}(s,\pi(s)) = \sum_{r,s'} \left[r + q_k(s',\pi(s'))\right] \Pr(s',r|s,\pi(s)),$$

   for all $s \in$ state space. After we obtain the updated action value function $q_{k+1}(s,\pi(s))$, a new police is improved by

$$\begin{aligned}\pi^{'}(s) &= \arg\max_a q_{k+1}(s,a)\\ &= \arg\max_{\pi(s)} q_{k+1}(s,\pi(s)).\end{aligned}$$

7. Exercise 4.7

   For my understanding,

$$\pi(s) = \begin{cases}\arg\max\limits_a q(s,a), & 1-\epsilon,\\ \text{random police}, & \epsilon.\end{cases}$$

   and changing the stopping rule.

8. Exercise 4.8

   No idea :(

9. Exercise 4.9

   See supplement.

10. Exercise 4.10

    From Exercise 4.6, it is easy to see the update rule is

$$q_{k+1}(s,\pi(s)) = \max_a \left\{\sum_{r,s'} \left[r + \sum_{a'} q_\pi(s',a')\,\pi(a'|s')\right] \Pr(s',r|r,a)\right\}.$$

1.4 Chapter 5

1. Exercise 5.1

   Skip.

2. Exercise 5.2

   Skip.

3. Exercise 5.3

   For $Q(s,a)$, we start from S_t, A_t, it hence follows that

$$\begin{aligned}\tilde{\rho}_t^T &= \prod_{k=t}^{T-1} \frac{\Pr_\pi\left(S_{k+1}, A_{k+1} | S_k, A_k\right)}{\Pr_u\left(S_{k+1}, A_{k+1} | S_k, A_k\right)} \\ &= \prod_{k=t}^{T-1} \frac{\pi\left(A_{k+1} | S_{k+1}\right) \Pr\left(S_{k+1} | S_k, A_k\right)}{u\left(A_{k+1} | S_{k+1}\right) \Pr\left(S_{k+1} | S_k, A_k\right)} \\ &= \prod_{k=t}^{T-1} \frac{\pi\left(A_{k+1} | S_{k+1}\right)}{u\left(A_{k+1} | S_{k+1}\right)}.\end{aligned}$$

4. Exercise 5.4

   The weighted important sampling is quite biased if the process continues few steps. This implies that the game stops early.

5. Exercise 5.5

   Yes, since in this case, we have only one state. The every-visit MC is equal to the first-visit MC.

6. Exercise 5.6-5.7

   Skip.

7. Exercise 5.8 (Programming)

   The whole framework is similar to Exercies 8.4. It thus skips.

8. Exercise 5.9

   We need only consider the sampling ratio for $Q(s,a)$ derived in Exercise 5.3.

1.5 Chapter 6

1. Exercise 6.1

   $$\begin{aligned} G_t - V_t\left(S_t\right) &= R_{t+1} + \gamma G_{t+1} - V_t\left(S_t\right) + \gamma\left[V_{t+1}\left(S_{t+1}\right) - V_{t+1}\left(S_{t+1}\right)\right] + \gamma\left[V_t\left(S_{t+1}\right) - V_t\left(S_{t+1}\right)\right] \\ &= \delta_t + \gamma\left[G_{t+1} - V_{t+1}\left(S_{t+1}\right)\right] + \gamma\alpha\delta_t^{'} = \dots, \end{aligned}$$

   where $\delta_t^{'} = R_{t+1} + \gamma V_t\left(S_{t+2}\right) - V_t\left(S_{t+1}\right)$.

2. Exercise 6.2

   Skip.

3. Exercise 6.3

   In the first episode, we stop at state zero.

4. Exercise 6.4

   See the relationship between MC error and TD error. It implies that both use a very small α will have similar results.

5. Exercise 6.5

   The α refers to the size of step. A large α may cause that the state value V_t cannot go to its optimal value due the large step size. In this case, the initial value may not be important since it is equally treat each state.

6. Exercise 6.6

   By "absorbing probability", we have

   $$\mathrm{Pr}_i = \sum_{j=1}^{k} p_{ij}\mathrm{Pr}_j,$$

   where p_{ij} is the transition probability and Pr_i is the probability that we start from state i and end at an absorbing state.

7. Exercise 6.7

   Since

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi\left(R_{t+1}+\gamma G_{t+1}|S_t=s\right)\\
&= \mathbb{E}_\pi\left(R_{t+1}|S_t=s\right)+\mathbb{E}_\pi\left(\gamma G_{t+1}|S_t=s\right)\\
&= \mathbb{E}_b\left(\rho_{t:t}R_{t+1}|S_t=s\right)+\mathbb{E}_\pi\left[\gamma\mathbb{E}_\pi\left(G_{t+1}|S_{t+1}\right)|S_t=s\right]\\
&= \mathbb{E}_b\left(\rho_{t:t}R_{t+1}|S_t=s\right)+\mathbb{E}_\pi\left[\gamma v_\pi\left(S_{t+1}\right)|S_t=s\right]\\
&= \mathbb{E}_b\left(\rho_{t:t}R_{t+1}|S_t=s\right)+\gamma\mathbb{E}_b\left[\rho_{t:t}v_\pi\left(S_{t+1}\right)|S_t=s\right]\\
&= \mathbb{E}_b\left\{\rho_{t:t}\left[R_{t+1}+\gamma v_\pi\left(S_{t+1}\right)\right]|S_t=s\right\},
\end{aligned}$$

   So,

$$V\left(S_t\right)=V\left(S_t\right)+\alpha\left\{\rho_{t:t}\left[R_{t+1}+\gamma V\left(S_{t+1}\right)\right]-V\left(S_t\right)\right\}, a\sim b\left(A_t|S_t\right).$$

8. Exercise 6.8

   Skip.

9. Exercise 6.9-6.10

10. Exercise 6.11

    Since in the update formula, A_{t+1} is chosen from $\arg\max_a Q\left(S_{t+1},a\right)$ independent of the current policy.

11. Exercise 6.12

$$Q_1\left(S_t,A_t\right)=Q_1\left(S_t,A_t\right)+\alpha\left[R_{t+1}+\gamma\sum_a\pi\left(a|S_{t+1}\right)Q_2\left(S_{t+1},A^*\right)-Q\left(S_t,A_t\right)\right],$$

    where $A^*=\arg\max_a Q_1\left(s,a\right)$ and $\pi\left(a|s\right)=\begin{cases}1-\varepsilon+\varepsilon/\left|\mathfrak{A}\left(s\right)\right| & ,a=A^*\\ \varepsilon/\left|\mathfrak{A}\left(s\right)\right| & ,a\neq A^*\end{cases}$, see Algorithm “On-policy first-visit MC control (for $\varepsilon-$soft policies)”

12. Exercise 6.13

    No idea :(.

    Actually, to my knowledge to "afterstate value function", I think by using it the updating involves the information of customers, it thus reduce the randomness and increase the convergence rate.

1.6 Chapter 1.7

1. Exercise 7.1

$$\begin{aligned} G_{t:t+n} - V\left(S_t\right) &= R_{t+1} + \gamma G_{t+1:t+n} - V\left(S_t\right) + \gamma\left[V\left(S_{t+1}\right) - V\left(S_{t+1}\right)\right] \\ &= \delta_t + \gamma\left[G_{t+1:t+h} - V_{t+1}\left(S_{t+1}\right)\right] = \cdots = \sum_{k=0}^{n-1} \gamma^k \delta_{t+k}. \end{aligned}$$

2. Exercise 7.2 (Programming)

   We use the gamble defined in Exercise 4.9. See supplement.

3. Exercise 7.3

   Intuitively, setting the reward at the most left state as -1 will encourage to postpone the terminal time of one episode. This implies that we can test the large n.

4. Exercise 7.4

   Skip (Change the formula of G).

5. Exercise 7.5

   Skip.

6. Exercise 7.6

$$G_{t:t+n} - V\left(S_t\right) = \rho_t\left\{\delta_t + \gamma\left[G_{t+1:t+n} - V\left(S_t\right)\right]\right\} = \ldots$$

7. Exercise 7.7

   Skip.

8. Exercise 7.8 (Programming)

   We use state transition problem, see supplement.

1.7 Chapter 8

1. Exercise 8.1

   In this case, I think there exists a suitable searching step for multiple bootstrap method such that this method does well as the planning method. Note that (A) multiple bootstrap method have a searching routine but planning method randomly select occurred states to update the action value function and (B) for both models, especially for the latter, it is easy to obtain the complete information for the environment.

2. Exercise 8.2

   The reward of Q^+ is $r + k\sqrt{\tau} > r$.

3. Exercise 8.3

   Both methods find the optimal solution quickly, and the reward of Q^+ is slightly larger than Q's since τ is quite small (even zero).

4. Exercise 8.4

   I use maze for demonstration. See supplement.

2 Part II

Note that "the proceeding state is changed" in the backward view, is presented to be the TD error δ_t times the eligibility trace vector e_t. For example, when $\lambda = 0$, e_t is only a function of S_t and $\boldsymbol{\theta}_t$. When $\lambda < 1$, clearly e_t is a function of S_t, $\boldsymbol{\theta}_t$ and e_{t-1}(is a function of S_{t-1}, $\boldsymbol{\theta}_{t-1}$ and e_{t-2})

2.1 Chapter 9

1. Exercise 9.1

   Each distinct feature will occur $n+1$ times, and there are k distinct features.

2. Exercise 9.2

   It is not hard to see that $n=2$ and $c_{1,1}=c_{1,2}=0$, $c_{2,1}=1, c_{2,2}=0$, ..., $c_{9,1}=c_{9,2}=2$.

3. Exercise 9.3

   There are $(n+1)^k$ combinations.

4. Exercise 9.4

   Skip.

2.2 Chapter 10

1. Exercise 10.1

   An episode is too long.

2. Exercise 10.2

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t+\alpha\left[R_{t+1}+\gamma\sum_a\pi\left(a|S_{t+1}\right)q\left(S_{t+1},a,\boldsymbol{w}_t\right)-q\left(S_t,A_t,\boldsymbol{w}_t\right)\right]\nabla q\left(S_t,A_t,\boldsymbol{w}_t\right),$$

3. Exercise 10.3

   High standard error of steps per episode is caused by the high standard error of parameter $\boldsymbol{w}$. This implies that $\hat{q}\left(S,a,\boldsymbol{w}\right)$ is not a good estimator for $G_{t:t+n}$ as n becomes large.

2.3 Chapter 11

1. Exercise 11.1

$$w_{t+n} = w_{t+n-1} - \alpha \rho_{t:t+n-1} \delta_t \nabla v \left(S_t, w_{t+n-1}\right),$$

where

(a) for episode case,

$$\begin{aligned}\delta_t &= G_{t:t+n} - v\left(S_t, w_{t+n-1}\right) \\ &= R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n v\left(S_{t+n}, w_{t+n-1}\right) - v\left(S_t, w_{t+n-1}\right).\end{aligned}$$

(b) for continuous case,

$$\begin{aligned}\delta_t &= G_{t:t+n} - v\left(S_t, w_{t+n-1}\right) \\ &= R_{t+1} - \bar{R}_t + R_{t+2} - \bar{R}_{t+1} + \ldots \gamma R_{t+n} - \bar{R}_{t+n-1} + v\left(S_{t+n}, w_{t+n-1}\right) - v\left(S_t, w_{t+n-1}\right).\end{aligned}$$

2. Exercise 11.2

   Combine (7.16) and (7.17)

3. Exercise 11.3

$$\begin{aligned}&\mathbb{E}\left[G_t - \hat{v}\left(S_t, w\right)\right]^2 \\ =&\mathbb{E}\left[G_t - v_\pi\left(S_t\right)\right]^2 + \bar{VE}\left(w\right) + 2\mathbb{E}\left[G_t - v_\pi\left(S_t\right)\right]\left[v_\pi\left(S_t\right) - \hat{v}\left(S_t, w\right)\right] \\ =&\mathbb{E}\left[G_t - v_\pi\left(S_t\right)\right]^2 + \bar{VE}\left(w\right) + 2\mathbb{E}\left\{\left[v_\pi\left(S_t\right) - \hat{v}\left(S_t, w\right)\right]\mathbb{E}\left[G_t - v_\pi\left(S_t\right)|S_t\right]\right\} \\ =&\mathbb{E}\left[G_t - v_\pi\left(S_t\right)\right]^2 + \bar{VE}\left(w\right).\end{aligned}$$

2.4 Chapter 12

1. Exercise 12.1

(a) .

$$
\begin{aligned}
G_{t:t+n} &= R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}\left(S_{t+n}, w_{t+n-1}\right) \\
&= R_{t+1} + \gamma \left[R_{t+2} + \cdots + \gamma^{n-2} R_{t+n} + \gamma^{n-1} \hat{v}\left(S_{t+n}, w_{t+n-1}\right)\right] \\
&= R_{t+1} + \gamma G_{t+1:t+n}.
\end{aligned}
$$

(b) The generalization

$$
\begin{aligned}
G_t^\lambda &= (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t \\
&= (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1}\left(R_{t+1} + \gamma G_{t+1:t+n}\right) + \lambda^{T-t-1}\left(R_{t+1} + \gamma G_{t+1}\right) \\
&= R_{t+1} + (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} \gamma G_{t+1:t+n} + + \lambda^{T-t-1} \gamma G_{t+1} \\
&= R_{t+1} + (1-\lambda) \gamma G_{t+1:t+1} + \gamma \left[(1-\lambda) \sum_{n=2}^{T-t-1} \lambda^{n-1} G_{t+1:t+n} + \lambda^{T-t-1} G_{t+1}\right] \\
&= R_{t+1} + \gamma \left[(1-\lambda) \hat{v}\left(S_{t+1}, w_t\right) + \lambda G_{t+1}^\lambda\right].
\end{aligned}
$$

2. Exercise 12.2

Given the final time T, find a λ such that

$$
(1-\lambda) \sum_{n=1}^{T/2} \lambda^n = \frac{1}{2}.
$$

3. Exercise 12.3

Suppose $w_t = w$ for all t. By Exercise 12.1,

$$
\begin{aligned}
G_t^\lambda - \hat{v}\left(S_t, w\right) &= R_{t+1} + (1-\lambda) \gamma \hat{v}\left(S_{t+1}, w\right) + \gamma \lambda G_{t+1}^\lambda - \hat{v}\left(S_t, w\right) \\
&= R_{t+1} + \gamma \hat{v}\left(S_{t+1}, w\right) - \hat{v}\left(S_t, w\right) + \gamma \lambda \left[G_{t+1}^\lambda - \hat{v}\left(S_{t+1}, w\right)\right] \\
&= \delta_t + \gamma \lambda \left[G_{t+1}^\lambda - \hat{v}\left(S_{t+1}, w\right)\right] = \ldots.
\end{aligned}
$$

4. Exercise 12.4

$$\begin{aligned} G_t^{\lambda} - V_t(S_t) &= R_{t+1} + (1-\lambda)\gamma V_t(S_{t+1}) + \gamma\lambda G_{t+1}^{\lambda} - V_t(S_t) + \gamma\left[V_t(S_{t+1}) - V_t(S_{t+1})\right] + V_{t-1}(S_t) - \\ &= R_{t+1} + \gamma V_t(S_{t+1}) - V_{t-1}(S_t) + \gamma\lambda\left[G_{t+1}^{\lambda} - V_t(S_{t+1})\right] - \Delta V_t(S_t) \\ &= u_t^{'} + \gamma\lambda\left[G_{t+1}^{\lambda} - V_t(S_{t+1})\right] - \Delta V_t(S_t) \\ &= u_t^{'} + \gamma\lambda u_{t+1}^{'} + \gamma^2\lambda^2\left[G_{t+2}^{\lambda} - V_t(S_{t+2})\right] - \gamma\lambda\Delta V_t(S_{t+1}) - \Delta V_t(S_t) = \ldots. \end{aligned}$$

We may use the 19 states random walk to be an experiment. The code is quite standard, and thus omits.

5. Exercise 12.5

Suppose that the value function is held constant, namely, $\triangle V_t(S) = V_t(S) - V_{t-1}(S) = 0$ for all t. Then, by the result of Exercise 12.4, we have the desired outcome.

6. Exercise 12.6

Just following (12.16). Skip.

7. Exercise 12.7

$$\begin{aligned} G_{t:h}^{\lambda_s} &= R_{t+1} + \gamma_{t+1}\left[(1-\lambda_{t+1})\,\hat{v}(S_{t+1}, w_t) + \lambda_{t+1} G_{t+1:h}^{\lambda_s}\right], \\ G_{t:h}^{\lambda_a} &= R_{t+1} + \gamma_{t+1}\left[(1-\lambda_{t+1})\,\hat{q}(S_{t+1}, A_{t+1}, w_t) + \lambda_{t+1} G_{t+1:h}^{\lambda_a}\right]. \end{aligned}$$

8. Exercise 12.8

The steps are similar to Exercise 12.3-4 and tedious, and thus omit.

9. Exercise 12.9 (truncated case)

Suppose all $G_T = G_{T+1} = \ldots,$

$$G_t^{\lambda_s} - \hat{v}\left(S_t, w\right) = \rho_t \sum_{k=t}^{T} \delta_k^s \prod_{i=t+1}^{k} \gamma_i \lambda_i \rho_i.$$

10. Exercise 12.10-14

    All are tedious, and thus we skip all of them.

11. Exercise 12.15 (Programming)

    Unsolved. This problem is not hard to be solved, but is worth trying.

2.5 Chapter 13

1. Exercise 13.1

   By assumptions on features, we have

   $$\frac{\exp\left(h\left(s,a,\theta\right)\right)}{\sum_b \exp\left(h\left(s,b,\theta\right)\right)} = \frac{\exp\left(\theta_1\right)}{\exp\left(\theta_1\right)+\exp\left(\theta_2\right)} = \frac{1}{1+\exp\left(\theta_2-\theta_1\right)}.$$

   Hence, as long as $\ln\left(\frac{1-p}{p}\right) = \theta_2 - \theta_1$, we have

   $$\frac{1}{1+\exp\left(\theta_2-\theta_1\right)} = p.$$

2. Exercise 13.2

   Skip.

3. Exercise 13.3

   (a) By $\exp\left(h\left(s,0,\boldsymbol{\theta}\right)\right)/\exp\left(h\left(s,1,\boldsymbol{\theta}\right)\right) = \exp\left(h\left(s,0,\boldsymbol{\theta}\right)-h\left(s,1,\boldsymbol{\theta}\right)\right) = \exp\left(-\boldsymbol{\theta}'\boldsymbol{\phi}\left(\boldsymbol{s}\right)\right)$, we have the desire result.

   (b) By computation, we have $\nabla\log\left(\pi\left(1|S_t,\boldsymbol{\theta}\right)\right) = \left[1-\pi\left(1|S_t,\boldsymbol{\theta}\right)\right]\boldsymbol{\phi}\left(\boldsymbol{s}\right)$.

   (c) See (b).

3 Part III

3.1 Chpater 17

1. Exercise 17.1

   Skip.