1. (20)

For the average reward setting, the differential state-value function is defined as:

$$v_\pi(s) \doteq E_\pi\left[\sum_{k=t+1}^{\infty}(R_k - r(\pi)) \mid S_t = s\right],$$

where the average reward is

$$r(\pi) = \lim_{h\to\infty} E_\pi\left[\frac{1}{h}\sum_{t=0}^{h-1} R_{t+1}\right].$$

Derive the following Bellman equation for v_π:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r|s, a)\left[r - r(\pi) + v_\pi(s')\right].$$

Show steps.

$$V_{\pi_\theta}(s) \doteq \mathbb{E}_{\pi_\theta}\left[\sum_{k=t+1}^{\infty}(R_k - r(\pi_\theta)) \mid S_t = s\right]$$

$$= \mathbb{E}_{\pi_\theta}\left[R_{t+1} - r(\pi_\theta) + \sum_{k=t+2}^{\infty}(R_{t+2} - r(\pi_\theta)) \mid S_t = s\right]$$

$$= \mathbb{E}_{\pi_\theta}\left[R_{t+1} - r(\pi_\theta) + \mathbb{E}\left[\sum_{k=t+2}^{\infty}(R_{t+2} - r(\pi_\theta)) \mid S_{t+1}, \underbrace{S_t = s}_{\text{无效}}\right] S_t\right] \Rightarrow \text{L.T.E } \mathbb{E}[X|Z] = \mathbb{E}[\mathbb{E}[X|Y,Z]|Z]$$

$$= \mathbb{E}_{\pi_\theta}\left[R_{t+1} - r(\pi_\theta) + \mathbb{E}\left[\sum_{k=t+2}^{\infty}(R_{t+2} - r(\pi_\theta)) \mid S_{t+1}\right] S_t\right] \Rightarrow \text{Markov property. 上一步这里当前状态, 与过去无关.}$$

$$= \mathbb{E}_{\pi_\theta}\left[R_{t+1} - r(\pi_\theta) + V_{\pi_\theta}(S_{t+1}) \mid S_t = s\right] \Rightarrow V_{\pi_\theta}(s) \doteq \mathbb{E}_{\pi_\theta}\left[\sum_{k=t+1}^{\infty}(R_k - r(\pi_\theta)) \mid S_t = s\right]$$

$$= \sum_{A, S', r} P_{\pi_\theta}(A_t = a, S_{t+1} = s', R_{t+1} = r \mid S_t = s)\left[r - r(\pi_\theta) + V_{\pi_\theta}(s')\right] \Rightarrow \text{L.O.T.E } \mathbb{E}_Z[g(x)] = \sum_{x \in X} g(x) P(X = x)$$

$$= \sum_{a} \pi_\theta(a|s) \cdot \sum_{s', r} P(s', r \mid s, a) \cdot \left[r - r(\pi_\theta) + V_{\pi_\theta}(s')\right]$$

Ps: 用 $q_{\pi_\theta}(s, a)$ 那个 bellman equation 或用 $V_{\pi_\theta}(s)$ 都

2. (20)
Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π.

Tabular TD(0) for estimating v_π

1. Input: the policy π to be evaluated
2. Algorithm parameter: step size $\alpha \in (0, 1]$
3. Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$
4. Loop for each episode:
5. Initialize S
6. Loop for each step of episode:
7. $A \leftarrow$ action given by π for S => estimate q_π, ∴在 loop 前选 action. (跟 sarsa 不一样)
8. Take action A, observe R, S'
9. $V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$
10. $S \leftarrow S'$
11. until S is terminal

基本上都是 Sarsa (on-policy TD control) 类, 只是不用 ← action given by π for S 改为 Q.

原理总结: (由下面资料而来):

① Tabular TD (0): estimate v_π, 所以假设 π 不变, 为 prediction.
update 的 V 跟 v_π, ∴更新 $V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$

② Sarsa: on-policy TD control, estimate $Q \to q_*$, 收敛到 optimal action value q_*, ∴为 update policy.
update 的 Q 跟 q_*, ∴更新 $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$

③ Q-learning: off-policy TD control. estimate $\pi \to \pi_*$, 收敛到 optimal policy π_*. ∴为 update policy.
behavior policy 为选择 max value 的 action
target policy (要更新的) 为 ε-greedy.
更新: $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$

④ Expected Sarsa: 加入 prediction, control, on-policy, off-policy.
更新: $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S, A)]$

⑤ Model based Q-planning: 在 Q learning 的基础上加一步 model learning, 即:
Send S, A to a sample model, and obtain
a sample next reward R, a sample next state S'

⑥ Tabular Dyna Q: 即 Q learning + Q planning.
- 前面用 direct RL 得到从 environmental interaction (experiences) 中学) 更新 $Q(S, A)$. 更新方式与 Q learning 一样.
- 后面用之前得到 (模拟) 的 S, A 来更新 model. 更新方式也与 Q-learning 一样.

⑦ Tabular Dyna Q+: 即 Q learning + Q planning. 大体都与 Dyna Q 一样.
但需将 model 更新中的 R 提为 $R + \kappa\sqrt{\tau}$.
并记得在前面 direct RL 处 initialize $\tau(s, a) = 0$, 更新 $\tau(s, a) += 1$.

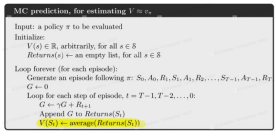
2. (20)
Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π.

Tabular TD(0) for estimating v_π

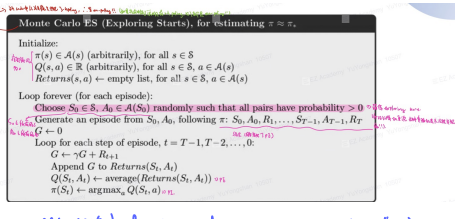
Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$
Loop for each episode:
  Initialize S
  Loop for each step of episode:
    $A \leftarrow$ action given by π for S
    Take action A, observe R, S'
    $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
    $S \leftarrow S'$
  until S is terminal

具体解法：① Monte-Carlo: 一直走到 episode 的末尾才更新，∴更新的是 G_t 以后的所有得到的实际 reward 全部加起来

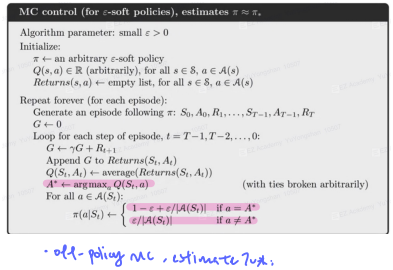
• MC prediction, estimate V_π



• MC exploring starts, estimate π_* (control)

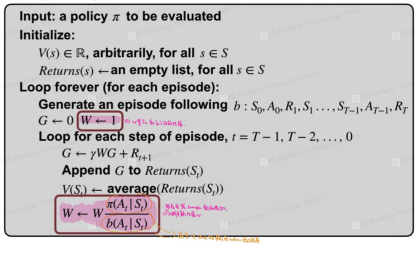


• MC control for ε-greedy, estimate π_* (predict)



• off-policy MC, estimate π_*

Off policy MC



Difference between prediction and control in pseudocode

tabular TD(0) for q_π: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$

^ Which is different than the following

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Loop for each step of episode:
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma Q(S', A') - Q(S, A)\right]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

approximating qpi

What would you modify in the above to get the pseudo code for TD(0)?

$$\mathbb{E}_{A\sim b}\left[\rho(A|s)(R+\gamma V(S')) - V(s) \mid S=s\right]$$

(s: given，∴不用被包括在求和里.)

$$= \sum_{a,s',r} P_b(A=a, S'=s', R=r \mid S=s)\,\rho(a|s)\,(r+\gamma V(s'))$$ ⇒ LOTUS $\mathbb{E}[g(x)] = \sum_{x\in X} g(x)\cdot P(X=x)$

在b(a|s)中用given的s得到了一个a，∴就可以在这里把a放进去得到 s', r.

$$= \sum_{a,s',r} b(a|s)\,p(s',r|s,a)\,\frac{\pi(a|s)}{b(a|s)}\,(r+\gamma V(s'))$$

$$= \sum_{a,s',r} p(s',r|s,a)\,\pi(a|s)\,(r+\gamma V(s'))$$

$$= \mathbb{E}_{A\sim\pi}\left[R+\gamma V(S') - V(S) \mid S=s\right]$$

3. (20)

Show that the following off-policy TD(0) update for v_π is correct for transition $S, A\sim b, S', R$:

$$V(S) \leftarrow V(S) + \alpha\left[\rho(A|S)\left(R+\gamma V(S')\right) - V(S)\right]$$

by showing the following:

$$E_{A\sim b}\left[\rho(A|S)\left(R+\gamma V(S')\right) - V(S) \mid S=s\right]$$

$$= E_{A\sim\pi}\left[R+\gamma V(S') - V(S) \mid S=s\right].$$

Here, $\rho(A|S) = \dfrac{\pi(A|S)}{b(A|S)}$, π and b are policy distributions with assumption $\pi(A|S) > 0 \Rightarrow b(A|S) > 0$.

4. (20)

Give the specification of the off-policy Expected Sarsa control method.

Ans. For the transition S, A, R, S', where the action $A \sim b$ is drawn from policy distribution b, update the action value the following way:

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \sum_{a'} \pi(a'|S')Q(S',a') - Q(S,A)\right],$$

where the target policy $\pi \neq b$ (off-policy) and uses greedification (control) such as ϵ-greedy; $\alpha > 0$.

(20)

5. An agent is in a 3-state MDP, $S = \{1, 2, 3\}$, where each state has two actions $A = \{1, 2\}$. Assume the agent observed the following trajectory:

$S_0 = 1, A_0 = 1, R_1 = 1, S_1 = 2, A_1 = 2, R_2 = -1,$
$S_2 = 3, A_0 = 1, R_3 = 2, S_3 = 1, A_3 = 1, R_4 = 2, S_4 = 1.$

The agent uses Tabular Dyna-Q.
Which of the following are possible (or not possible) simulated transition $\{S, A, R, S'\}$ given the above observed trajectory with a deterministic model and random search control.

i. $\{S = 1, A = 1, R = 2, S' = 1\}$
ii. $\{S = 3, A = 1, R = 2, S' = 1\}$
iii. $\{S = 1, A = 1, R = 1, S' = 2\}$
iv. $\{S = 2, A = 2, R = -1, S' = 3\}$
v. $\{S = 3, A = 2, R_3 = 2, S' = 1\}$.

Just mention possible or not possible for each.

Worksheet 9: Planning, Learning & Acting

CMPUT 397
March 19, 2021

2. An agent is in a 4-state MDP, $\mathcal{S} = \{1, 2, 3, 4\}$, where each state has two actions $\mathcal{A} = \{1, 2\}$. Assume the agent saw the following trajectory,

$$
\begin{aligned}
&S_0 = 1, A_0 = 2, R_1 = -1, \\
&S_1 = 1, A_1 = 1, R_2 = 1, \\
&S_2 = 2, A_2 = 2, R_3 = -1, \\
&S_3 = 2, A_3 = 1, R_4 = 1, \\
&S_4 = 3, A_4 = 1, R_5 = 100, \\
&S_5 = 4
\end{aligned}
$$

and uses Tabular Dyna-Q with 5 planning steps for each interaction with the environment.

(a) Once the agent sees S_5, how many Q-learning updates has it done with **real experience**? How many updates has it done with **simulated experience**?

(b) Which of the following are possible (or not possible) simulated transitions $\{S, A, R, S'\}$ given the above observed trajectory with a deterministic model and random search control?

i. $\{S = 1, A = 1, R = 1, S' = 2\}$

ii. $\{S = 2, A = 1, R = -1, S' = 3\}$

iii. $\{S = 2, A = 2, R = -1, S' = 2\}$

iv. $\{S = 1, A = 2, R = -1, S' = 1\}$

v. $\{S = 3, A = 1, R = 100, S' = 5\}$

Tabular Dyna-Q

Initialize $Q(s, a)$ and $Model(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Loop forever:
- (a) $S \leftarrow$ current (nonterminal) state
- (b) $A \leftarrow \varepsilon$-greedy(S, Q)
- (c) Take action A; observe resultant reward, R, and state, S'
- (d) $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
- (e) $Model(S, A) \leftarrow R, S'$ (assuming deterministic environment)
- (f) Loop repeat n times:
  - $S \leftarrow$ random previously observed state
  - $A \leftarrow$ random action previously taken in S
  - $R, S' \leftarrow Model(S, A)$
  - $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$

a).

b). i). iii). iv).

For each update we do based on the real experience, we do n many updates using the simulated experience.

The question shows that there are 5 time steps of state-action pairs and their subsequent next state and reward that we observed from the real experience, and the question also says that we will do 5 planning steps, so n=5

So 5 updates based on real experience, and for each of these updates we do 5 planning steps

-> 5x5 = 25

对于我们基于真实经验所做的每一次更新，我们使用模拟经验做n次更新。

问题显示，有5个时间步骤的状态-动作对及其随后的下一个状态和奖励是我们从真实经验中观察到的，问题还说我们将做5个计划步骤，所以n=5

所以有5个基于真实经验的更新，对于每个更新，我们做5个规划步骤

-> 5x5 = 25

Worksheet 12: Control with Approximation

CMPUT 397
April 7, 2022

4. (*Exercise 10.6 S&B*) Suppose there is an MDP that under any policy produces the deterministic sequence of rewards $+1, 0, +1, 0, +1, 0, \ldots$ going on forever. Technically, this is not allowed because it violates ergodicity; there is no stationary limiting distribution μ_π and the limit

$$\lim_{t\to\infty} \mathbb{E}[R_t | S_0, A_{0:(t-1)} \sim \pi]$$

does not exist. Nevertheless, the average reward,

$$\lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t | S_0, A_{0:(t-1)} \sim \pi]$$

is well defined; What is it? Now consider two states in this MDP. From A, the reward sequence is exactly as described above, starting with a $+1$, whereas, from B, the reward sequence starts with a 0 and then continues with $+1, 0, +1, 0, \ldots$ The differential return

$$G_t = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \ldots$$

A: 1 0 1 0 1 0 ...
B: 0 1 0 1 0 1 ...

is not well defined for this case as the limit does not exist. To repair this, one could alternately define the value of a state as

$$v_\pi(s) = \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}_\pi[R_{t+1} | S_0 = s] - r(\pi)\right)$$

Under this definition, what are the values of states A and B?

Markov reward process: $C_1 = \frac{1}{1} = 1$

$C_2 = \frac{1+0}{2} = \frac{1}{2}$

$C_3 = \frac{1+0+1}{3} = \frac{2}{3}$ → 勝ち=odd.

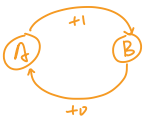
$C_4 = \frac{1+0+1+0}{4} = \frac{1}{2}$ ∴ even number: all half

$C_5 = \frac{1+0+1+0+1}{5} = \frac{3}{5}$ → 勝ち=odd. odd number: reducing. (converging to $\frac{1}{2}$, ∴ never going below 0.5 !!)

$C_6 = \frac{1+0+1+0+1+0}{6} = \frac{1}{2}$

$C_7 = \frac{4}{7}$ → 勝ち=odd

∴ $\lim_{h\to\infty} C_h = \frac{1}{2}$ is the average reward.



→ BからBの average reward は $\frac{1+0}{2} = \frac{1}{2}$.

Here, we have a period of 2: ∴ $V_\pi(A) = \frac{1}{2}((R_1 - r(\pi)) + (R_2 - r(\pi)))$

Starting from A: A→B 1回目 (1, 0, 1, 0, 1, ...) B→B average reward

$= \frac{1}{2}((1 - r(\pi)) + (0 - r(\pi)))$

$= \frac{1}{2}(1 - \frac{1}{2}) + (0 - \frac{1}{2}))$

$= \frac{1}{2}(1 - \frac{1}{2})$

$= \frac{1}{4}$.

Starting from B: $V_\pi(B) = \frac{1}{2}((R_1 - r(\pi)) + (R_2 - r(\pi)))$ B→A 1回目, ∴ $R_1 = 0$, $R_2 = 1$.

$= \frac{1}{2}((0 - \frac{1}{2}) + (1 - \frac{1}{2}))$ → B→B average reward.

$= -\frac{1}{4}$

Confirming accuracy through Bellman equation:

$V_\pi(A) = \frac{1}{4}$

$V_\pi(B) = -\frac{1}{4}$

$\therefore V_\pi(A) = 1 - r(\pi) + V_\pi(B)$

$= 1 - \frac{1}{2} - \frac{1}{4}$

$= \frac{1}{2} - \frac{1}{4}$

$= \frac{1}{4}$ ⇒ correct !!

$V_\pi(B) = 0 - r(\pi) + V_\pi(A)$

$= -\frac{1}{2} + \frac{1}{4}$

$= -\frac{1}{4}$ ⇒ correct !!

using Abel sum: $V_\pi(B) = \lim_{\gamma\to 1} ((R_1 - r(\pi)) + \gamma(R_2 - r(\pi)) + \gamma^2(R_3 - r(\pi)) + \ldots$

$= \lim_{\gamma\to 1} (1 - \frac{1}{2} + \gamma(0 - \frac{1}{2}) + \gamma^2(1 - \frac{1}{2}) + \ldots)$

$= \lim_{\gamma\to 1} (\frac{1}{2} - \frac{\gamma}{2} + \frac{\gamma^2}{2} \ldots)$

$= \frac{1}{2} \lim_{\gamma\to 1} (1 - \gamma + \gamma^2 - \gamma^3 + \ldots)$

$= \frac{1}{2} \lim_{\gamma\to 1} \frac{1}{1+\gamma}$

$= \frac{1}{2} \cdot \frac{1}{2}$

$= \frac{1}{4}$

Worksheet 12: Control with Approximation

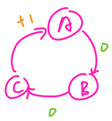
CMPUT 397
April 7, 2022

5. (*Exercise 10.7 S&B*) Consider a Markov reward process consisting of a ring of three states A, B, and C, with state transitions going deterministically around the ring. A reward of +1 is received upon arrival in A and otherwise the reward is 0. What are the differential values of the three states using

$$v_\pi = \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}_\pi[R_{t+1} | S_0 = s] - r(\pi) \right)$$

⇒ 这里指 you have to find the answer equal this way!! 不是说你要用这方法求 answer!!

Markov reward process:



$\therefore$ Average reward $r(\pi) = \frac{1+0+0}{3} = \frac{1}{3}$

Solve by Bellman equation:

$v_\pi(C) = (1 - \frac{1}{3}) + v_\pi(A)$ (这一步得到的 reward; Average reward; 这一步到达的 state.)

$v_\pi(B) = (0 - \frac{1}{3}) + v_\pi(C)$

$v_\pi(A) = (0 - \frac{1}{3}) + v_\pi(B)$

$v_\pi(A) = \frac{1}{3}(3(R_1 - r(\pi)) + 2(R_2 - r(\pi)) + R_3 - r(\pi))$

$= \frac{1}{3}(3(0 - \frac{1}{3}) + 2(0 - \frac{1}{3}) + 1(1 - \frac{1}{3}))$

从 A 开始走，reward 依次为：0 → 0 → 1

$= \frac{1}{3}(-1 - \frac{2}{3} + \frac{2}{3})$

$= -\frac{1}{3}$

$\therefore$ can find $v_\pi(B)$ and $v_\pi(C)$ by Bellman equation. 但因为用了 Bellman equation 来 find answer, ∴ 必须用另一种方式验证.

$v_\pi(C) = 1 - \frac{1}{3} + v_\pi(A) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$.

$v_\pi(B) = 0 - \frac{1}{3} + v_\pi(C) = -\frac{1}{3} + \frac{1}{3} = 0$.

验证：因为从 A 开始，往下 2 个 reward 都为 0，∴ A 的 valuable 程度最低，∴ $v_\pi(A)$ 为负.

B　　2 个 reward 分别为 0, 1.　∴ B 比 A 高，∴ $v_\pi(B) = 0$.

C　　下一个 reward 就是 1.　∴ C 最高，∴ $v_\pi(C) = 1$. 最大.

∴ 对.

8. (15)

What is the difference between the following two updates? Which one is a stochastic-gradient and a semi-gradient update?

a. $\underline{w}_{t+1} \doteq \underline{w}_t - \alpha \left[\nabla_{\underline{w}_t} \frac{1}{2} \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{\psi}) - \hat{v}(S_t, \underline{w}_t) \right)^2 \right] \Big|_{\psi = \underline{w}_t}$

求梯度 ∇w_t 因为 loss 是关于 w_t 的！不含 w_t！

$= \underline{w}_t + \frac{1}{2} \alpha [2(R_{t+1} + \gamma \hat{v}(S_{t+1}, \psi) - \hat{v}(S_t, \underline{w}_t)) \nabla_{\underline{w}_t} \hat{v}(S_t, \underline{w}_t)] \Big|_{\psi = \underline{w}_t}$

b. $\underline{w}_{t+1} \doteq \underline{w}_t - \alpha \nabla_{\underline{w}_t} \frac{1}{2} \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{w}_t) - \hat{v}(S_t, \underline{w}_t) \right)^2$.

对 w_t 求梯度 ∴ 将包含上述式中所有含有 w_t 的项.

$= \underline{w}_t + \frac{1}{2} \alpha \cdot 2(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{w}_t) - \hat{v}(S_t, \underline{w}_t)) \cdot (\nabla_{\underline{w}_t} \hat{v}(S_t, \underline{w}_t) - \nabla_{\underline{w}_t} \hat{v}(S_{t+1}, \underline{w}_t))$

↓ 调换顺序，得以将 2 前的符号变为正.

TD(0) 中没有，删为 TD(0) 中含有 ψ，不能表以 ∇w_t.

Ans.

(a) becomes

$$\underline{w}_{t+1} \doteq \underline{w}_t - \alpha \left[\nabla_{\underline{w}_t} \frac{1}{2} \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{\psi}) - \hat{v}(S_t, \underline{w}_t) \right)^2 \right] \Bigg|_{\psi = \underline{w}_t}$$

对这一项进行求导

因为也对 w_t 进行求偏导，∴需保留其余上含有 w_t 的一项.

$$= \underline{w}_t + \frac{1}{2} \alpha \left[2 \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{\psi}) - \hat{v}(S_t, \underline{w}_t) \right) \nabla_{\underline{w}_t} \hat{v}(S_t, \underline{w}_t) \right] \Bigg|_{\psi = \underline{w}_t}$$

$$= \underline{w} + \alpha \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{w}_t) - \hat{v}(S_t, \underline{w}_t) \right) \nabla_{\underline{w}_t} \hat{v}(S_t, \underline{w}_t),$$

which is the semi-gradient TD(0) update.

Therefore, this one is semi-gradient.

(b) becomes

$$\underline{w}_{t+1} \doteq \underline{w}_t - \alpha \nabla_{\underline{w}_t} \frac{1}{2} \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{w}_t) - \hat{v}(S_t, \underline{w}_t) \right)^2$$

$$= \underline{w}_t + \frac{1}{2} \alpha \, 2 \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \underline{w}_t) - \hat{v}(S_t, \underline{w}_t) \right) \times \left(\nabla_{\underline{w}_t} \hat{v}(S_t, \underline{w}_t) - \nabla_{\underline{w}_t} \hat{v}(S_{t+1}, \underline{w}_t) \right),$$

which is a gradient update and different from TD(0).

The main difference is the last term that does not exist in TD(0).