CMPUT365 Winter 2022 Final2: Sample Final

导师： Baihong Qi

Final Exam

1. (20)

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

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

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.

LOTE: Law of total expectation

1. $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \mid Y))$;

运算过程 [编辑]

$$
\begin{aligned}
\mathrm{E}(\mathrm{E}(X|Y)) &= \sum_y \mathrm{E}(X|Y=y) \cdot \mathrm{P}(Y=y) \\
&= \sum_y \left(\sum_x x \cdot \mathrm{P}(X=x|Y=y) \right) \cdot \mathrm{P}(Y=y) \\
&= \sum_y \sum_x x \cdot \mathrm{P}(X=x|Y=y) \cdot \mathrm{P}(Y=y) \\
&= \sum_y \sum_x x \cdot \mathrm{P}(Y=y|X=x) \cdot \mathrm{P}(X=x) \\
&= \sum_x \sum_y x \cdot \mathrm{P}(Y=y|X=x) \cdot \mathrm{P}(X=x) \\
&= \sum_x x \cdot \mathrm{P}(X=x) \cdot \left(\sum_y \mathrm{P}(Y=y|X=x) \right) \\
&= \sum_x x \cdot \mathrm{P}(X=x) \\
&= \mathrm{E}(X).
\end{aligned}
$$

MP: Markov Property

- Conditional Probability depends only upon the present states

LOTUS

Law of the unconscious statistician: $\mathbf{E}[\, g(X)\,] = \sum_{x \in \mathscr{X}} g(x)\, \mathrm{P}(X=x)$

Easy Edu

Ans.

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

$$= E_\pi\left[R_{t+1} - r(\pi) + \sum_{k=t+2}^{\infty}\left(R_k - r(\pi)\right) \middle| S_t = s\right]$$

$$= E_\pi\left[R_{t+1} - r(\pi) + E_\pi\left[\sum_{k=t+2}^{\infty}\left(R_k - r(\pi)\right) \middle| S_{t+1}, S_t = s\right] \middle| S_t = s\right]$$

[LOTE]

$$= E_\pi\left[R_{t+1} - r(\pi) + E_\pi\left[\sum_{k=t+2}^{\infty}\left(R_k - r(\pi)\right) \middle| S_{t+1}\right] \middle| S_t = s\right]$$

[MP]

$$= E_\pi\left[R_{t+1} - r(\pi) + v_\pi(S_{t+1}) \middle| S_t = s\right]$$

[v_π definition]

$$= \sum_{a,s',r} P_\pi\left(A_t = a, S_{t+1} = s', R_{t+1} = r \middle| S_t = s\right)\left[r - r(\pi) + v_\pi(s')\right]$$

[LOTUS]

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

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
8. Take action A, observe R, S'
9. $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$
10. $S \leftarrow S'$
11. until S is terminal

Sarsa

3. Sarsa算法(on-policy)

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)
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
        $S \leftarrow S'$; $A \leftarrow A'$;
    until S is terminal

Ans.

* In step 3, replace "$V(s)$" with "$Q(s,a)$ for all $a \in A$"

  and "$V(\text{terminal}) = 0$" with "$Q(\text{terminal}, \cdot) = 0$"

* Switch step 6 & 7
* Add before step 9 "$A' \leftarrow$ action given by π for S'"

* Replace step 9 with

  "$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S', A') - Q(S,A)]$"

* Add to step 10: "$A \leftarrow A'$"

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:

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

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

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

$\pi(A|S) > 0 \Rightarrow b(A|S) > 0.$

Easy Edu

Let's first work on the target

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

$$= \sum_{a,s',r} P_b(A=a, S'=s', R=r \mid S=s)\,\rho(a|s)(r+\gamma V(s'))$$

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

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

$$= \mathbb{E}_{A\sim\pi}\left[R+\gamma V(S')\,\middle|\, S=s\right] \quad (1).$$

On the other hand,

$$\mathbb{E}_{A\sim b}\left[V(S)\,\middle|\, S=s\right] = V(s).$$

Therefore, $\mathbb{E}_{A\sim b}\left[V(S)\,\middle|\, S=s\right] = \mathbb{E}_{A\sim\pi}\left[V(S)\,\middle|\, S=s\right]$. (2)

Therefore, the identity follows from (1)-(2).

4. (20)

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

Sarsa

3. Sarsa算法(on-policy)

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

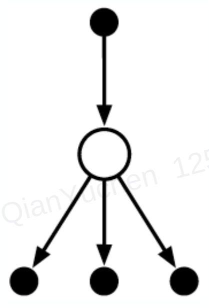
Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

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)
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
        $S \leftarrow S'$; $A \leftarrow A'$;
    until S is terminal

Sarsa

4. Expected Sarsa公式 (off-policy)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \mathbb{E}[Q(S_{t+1}, A_{t+1}) \mid S_{t+1}] - Q(S_t, A_t) \Big]$$
$$\leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t) \Big]$$



- St+1时look back
- current state 是 St+1

5. Sarsa与Expected Sarsa的关系

- Sarsa是on-policy, Expected Sarsa多数情况下是off-policy

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \Big]$$

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:

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

where the target policy $\underset{\text{off-policy}}{\underline{\pi \neq b}}$ and uses $\underset{\text{control}}{\underline{\text{greedification}}}$ 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.

Planning & Learning

3. Dyna Q

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) $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$ ← **direct RL**
(e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment) ← **model learning**
(f) Loop repeat n times:
$S \leftarrow$ random previously observed state
$A \leftarrow$ random action previously taken in S
$R, S' \leftarrow Model(S,A)$
$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$

← **planning**