CMPUT 365: Value Functions and Bellman Equations

Rupam Mahmood

Jan 24, 2022

Admin

Due dates C1M3:

- Practice quiz: Tues Jan 25
- Graded quiz: Sat Jan 29

Assignment 1:

- was released on last Sun Jan 23
- Two worksheet-like questions
- Due on this Sun Jan 30

Midterm:

- Based on worksheet questions, book reading, and lectures

Coursera video: Specifying policies

- Definition

Formally, a *policy* is a mapping from states to probabilities of selecting each possible action. If the agent is following policy π at time t, then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$. Like p, π is an ordinary function; the "|" in the middle of $\pi(a|s)$ merely reminds us that it defines a probability distribution over $a \in \mathcal{A}(s)$ for each $s \in \mathcal{S}$. Reinforcement learning methods specify how the agent's policy is changed as a result of its experience.

- Notations for deterministic and stochastic policies

- Invalid actions

All environment need to do is provide the next reward. The environment doesn't need to know $v_\pi(s)$ and $q_\pi(s,a)$, agent 自己知道. 如果已经有知道 $v_\pi(s)$ & $q_\pi(s,a)$, 之后的强化也是 computation 来得出.

Coursera video: Value Functions

- Definition of state value function

If the agent is following particular policy π from state s, what is the expected return the agent will get?

The *value function* of a state s under a policy π, denoted $v_\pi(s)$, is the expected return when starting in s and following π thereafter. For MDPs, we can define v_π formally by

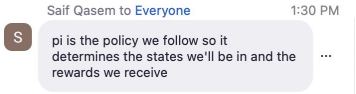
$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S}, \qquad (3.12)$$

you can't change G_t 确定的量

following a particular policy

you can change S_t

(v_π 是 depend on policy 的，所以要加上下标 π!!)



- Definition of action-value function

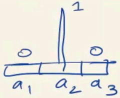
$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]. \qquad (3.13)$$

given. doesn't matter how you arrive them.

$\pi(a|s)$

$\pi(a_2|s) = 1$

$a_2 = \pi(s)$



reward depends on the previous state and action.

$$E[R_t \mid S_{t-1} = s] = E[R_{t+1} \mid S_t = s]$$

∴ 这里的 r must depend on π (policy) 因为这里 given 的只有 s. 没有 a. 如果不 depend on policy. 它就不知道 action (即 a) 是什么, 就求不出来 reward.

Coursera video: History of RL

- Rich S. Sutton
- Andy Barto
- A. Harry Klopf
- Philosophy of science: empiricism, logical positivism, Karl Popper
- Phenomenology: Hubert Dreyfus

Coursera video: Bellman equations

- Bellman equations

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], \quad \text{for all } s \in \mathcal{S},$$

the value function — i.e. s, an outcome — current action

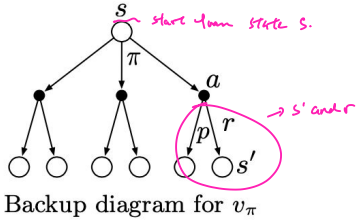
we can write the state value recursively.

$=$: equality (equal to) (logical equality)

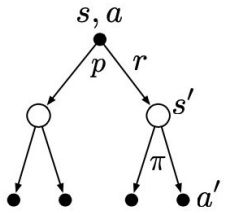
$\doteq$: by def sth is equal (you can't redefine sth that has been defined), $\therefore v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$.

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \quad \text{for all } s \in \mathcal{S}, \qquad (3.12)$$

is already define. — logically equal to.



Backup diagram for v_π



q_π backup diagram

- Let's derive it next time

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] && \text{(from (3.9))} \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] && (4.3) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], && (4.4)
\end{aligned}$$

Deriving Bellman Equations

V_f

Bellman equation for v_π

LOTE
LOTUS
MP
LE

$E[E[X|Y]] = E[X]$

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

↙ LOTE (加入新的变量. $E[X|Z] = E[E[X|Y,Z]|Z]$)

$$= E_\pi\Big[E_\pi[G_t \mid \underbrace{S_t = s}_{\text{旧变量}}, \underbrace{A_t}_{\text{新变量}}] \,\Big|\, S_t = s\Big]$$

↓ LOTUS ($\Rightarrow E[g(X)] = \sum_{x\in X} g(x)\cdot P(X=x)$)

$$= \sum_a E_\pi[G_t \mid S_t = s, A_t = a]\, P(\underbrace{A_t = a}_{\text{新变量}} \mid \underbrace{S_t = s}_{\text{旧变量, given}})$$

($E[g(X)]$ 变为 $\sum_{x\in X} g(x)\,P(X=x)$)

$$= \sum_a \pi(a|s)\, q_\pi(s,a)$$

$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

↙ LOTE

$$= E_\pi\Big[E_\pi[G_t \mid \underbrace{S_t = s, A_t = a}_{\text{旧变量}}, \underbrace{R_{t+1}, S_{t+1}}_{\text{新变量}}] \,\Big|\, S_t = s, A_t = a\Big]$$

↓ LOTUS

$g(X)$, 没有写出新加入的 R.V. 指为什么.

$$= \sum_{s',r} E_\pi[G_t \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s']$$

变为 $\sum_{x\in X} g(x)$, 注此处 x 由新加入的 R.V. 指为的来历!!

写出了新加入的 R.V. (即 R_{t+1} 和 S_{t+1}) 指为什么.

$$p(s', r \mid \underbrace{s, a}_{\text{旧变量, given}})$$

$$= \sum_{s',r} p(s',r|s,a)\, E_\pi[\underbrace{R_{t+1} + \gamma G_{t+1}}_{G_t = R_{t+1} + \gamma G_{t+1}} \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s']$$

↓LE + MP

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \mathop{E}_{\pi}\left[G_{t+1} \mid S_{t+1}=s'\right]\right]$$

By M.P. (G_{t+1} 仅由当前状态 S_{t+1} 决定.)

→ 用以下的 constant 提出来.

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

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

Show that

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

where

$$r_\pi(s) \doteq E_\pi\left[R_t \mid S_{t-1} = s \right]$$

$$= \sum_r r \, P_\pi\left(R_t = r \mid S_{t-1} = s \right)$$

$$\downarrow \text{LOTP}$$

$$= \sum_r r \sum_a \pi(a|s) P\left(R_t = r \mid S_{t-1} = s, A_{t-1} = a \right).$$

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

depend on policy (which action is not given.)

$$r_\pi(s) \doteq E_\pi[R_t \mid S_{t-1} = s]$$

expected reward of state s.

↳ 期望值定義 ≐

$$= \sum_r r \, P_\pi(R_t = r \mid S_{t-1} = s)$$

$$= \sum_r r \sum_a \pi(a|s) P(R_t = r \mid S_{t-1} = s, A_{t-1} = a)$$

Properties of expectations

- ✓ Linearity: $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$
- ✓ Linearity: $\mathbf{E}[aX] = a\mathbf{E}[X]$
- ✓ Non-multiplicativity: $\mathbf{E}[XY] \neq \mathbf{E}[X]\,\mathbf{E}[Y]$
- ✓ Law of the unconscious statistician: $\mathbf{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\,P(X=x)$

sum of two different R.V.

R.V.

constant

R.V.

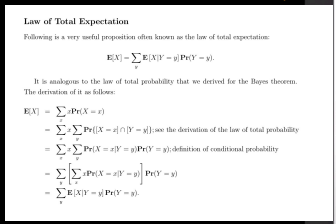
A R.V. which itself is a function of another R.V.

$$= \sum_{y \in Y} y\,P(Y=y)$$

$$= \sum g(x)\,P(g(X)=g(x))$$

$$= \sum_{x \in X} g(x)\,P(X=x)$$

bellman-eq.mp4



Law of Total Expectation

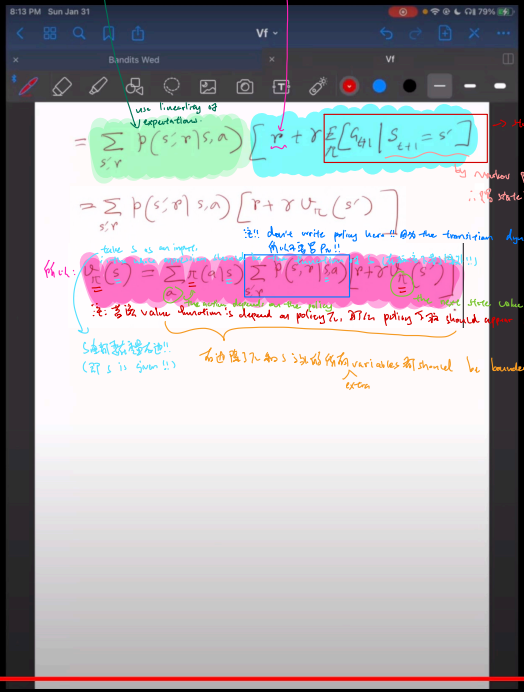
Following is a very useful proposition often known as the law of total expectation:

$$\mathbf{E}[X] = \sum_y \mathbf{E}[X|Y=y]\Pr(Y=y).$$

It is analogous to the law of total probability that we derived for the Bayes theorem. The derivation of it se follows:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_x x\Pr(X=x) \\
&= \sum_x x \sum_y \Pr[(X=x|\cap)(Y=y)]; \text{see the derivation of the law of total probability} \\
&= \sum_x x \sum_y \Pr(X=x|Y=y)\Pr(Y=y); \text{definition of conditional probability} \\
&= \sum_y \left[\sum_x x\Pr(X=x|Y=y)\right]\Pr(Y=y) \\
&= \sum_y \mathbf{E}[X|Y=y]\Pr(Y=y).
\end{aligned}$$

bellman-eq.mp4



14:26 / 19:37

Expectations & conditional expectations

- An expected value of a random variable is a weighted average of possible outcomes, where the weights are the probabilities of those outcomes

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \, P(X=x)$$

- An expected value of a random variable conditional on another event is a weighted average of possible outcomes, where the weights are the conditional probabilities of those outcomes given the event

$$\mathbf{E}[X \mid Y=y] = \sum_{x \in \mathcal{X}} x \, P(X=x \mid Y=y)$$

- Expectation conditional on a random variable $\mathbf{E}[X \mid Y]$ itself is a random variable, which is a function of another random variable Y

$$v_{k+1}(s) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s]$$

this is not value function!!

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma v_k(s')\Big],$$

Coursera video: Optimal policies and Optimal Value Functions

- Def'n:

$$\pi \geq \pi' \text{ if and only if } v_\pi(s) \geq v_{\pi'}(s) \text{ for all } s \in \mathcal{S}$$

$$v_*(s) \doteq \max_\pi v_\pi(s)$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a)$$

- Bellman optimality equations:

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) = \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_*(s')\right]$$

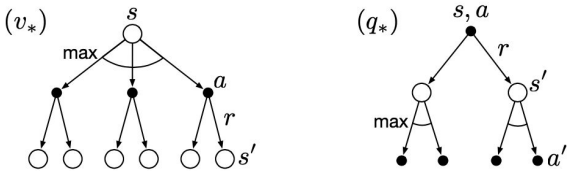


Figure 3.4: Backup diagrams for v_* and q_*

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right]$$
$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right].$$

Additional videos

- Notations for MDPs

- Deriving the Bellman equation with all steps shown