Value Functions & Bellman Equations

Rupam Mahmood

January 29, 2020

The Bellman equation for v_π

overbrace ($\frown$) means changes

$$v_\pi(s) \doteq E_\pi[G_t \mid S_t = s] = E_\pi\left[\overbrace{E_\pi}\left[G_t \mid S_t = s, \overbrace{A_t}\right]\right]$$

law of total expectations

$$= \overbrace{\sum_a \pi(a \mid s)}\, E_\pi\left[G_t \mid S_t = s, A_t \overbrace{= a}\right]$$

law of the unconscious statistician

$$= \sum_a \pi(a \mid s) E_\pi\left[\overbrace{E_\pi}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, \overbrace{R_{t+1}, S_{t+1}}\right]\right]$$

law of total expectations

$$= \sum_a \pi(a \mid s) \overbrace{\sum_{s',r} p(s', r \mid s, a)}\, E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, R_{t+1} \overbrace{= r}, S_{t+1} \overbrace{= s'}\right]$$

law of the unconscious statistician

$$= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r \overbrace{+} \gamma E_\pi\left[G_{t+1} \mid \overbrace{S_{t+1} = s'}\right]\right]$$

Markov property & linearity of expectation

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

definition of v_π

Worksheet question

(*Exercise 3.12 in* 2^{nd} *ed.*) Recall that the value $v_\pi(s)$ for state s when following policy π is the expected total reward (or discounted reward) the agent would receive when starting from state s and executing policy π. How can we write $v_\pi(s)$ in terms of the action values $q_\pi(s, a)$?

Optimal policies & values

Optimal state-value function: $$v_*(s) \doteq E_{\pi_*}[G_t \mid S_t = s] = \max_{\pi} v_\pi(s), \forall s$$

Optimal action-value function: $$q_*(s,a) \doteq E_{\pi_*}[G_t \mid S_t = s, A_t = a] = \max_{\pi} q_\pi(s,a), \forall s, a$$

$$v_*(s) = \sum_a \pi_*(a \mid s) q_*(s,a) \qquad = \max_a q_*(s,a)$$

An optimal policy: $\pi_*(a \mid s) = 1$ **if** $a = \bar{\arg\max_b}\, q_*(s,b)$, $\quad 0$ **otherwise**

where $\bar{\arg\max}$ **is** $\arg\max$ **with ties broken in a fixed way**

Bellman optimality equations

value under π: $$v_\pi(s) \doteq E_\pi[G_t \mid S_t = s] = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]$$

optimal value: $$v_*(s) \doteq E_{\pi_*}[G_t \mid S_t = s] = \sum_a \pi_*(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$$

$$= \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$$

Writing action-value functions wrt state-value functions

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

$$= E_\pi\left[E_\pi\left[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s, A_t = s, R_{t+1}, S_{t+1}\right]\right]$$

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

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

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

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

The Bellman equation for q_π

Bellman equation with expected reward $r(s, a)$

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

$$\sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)r = \sum_a \pi(a|s) \sum_r r \sum_{s'} P(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a)$$

$$= \sum_a \pi(a|s) \sum_r rP(R_{t+1} = r|S_t = s, A_t = a)$$

due to the law of total probabilities

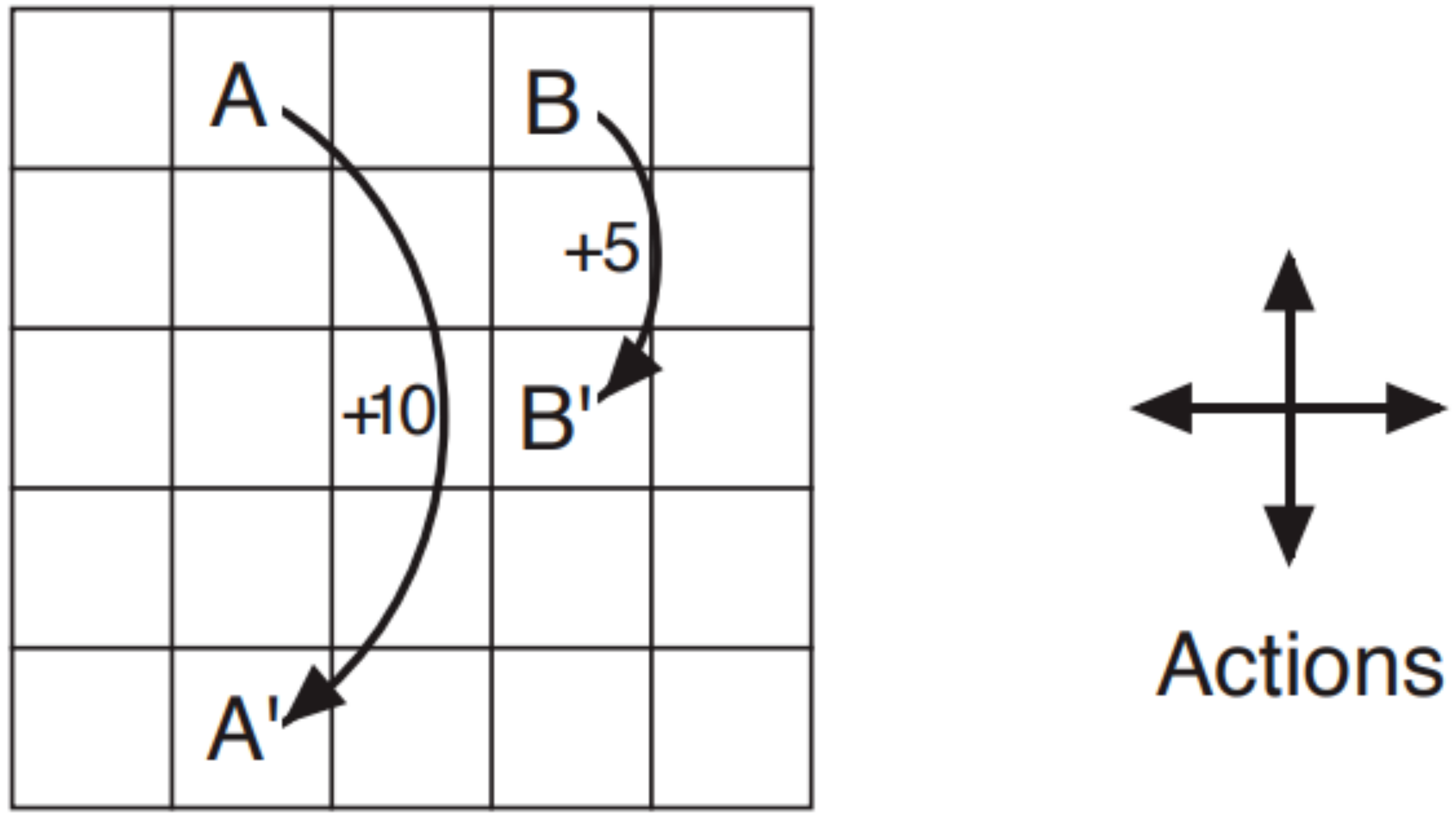
$$= \sum_a \pi(a|s)E[R_{t+1}|S_t = s, A_t = a]$$

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

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

Worksheet question

Consider the gridworld and value function in the figure below. Using your knowledge of the transition dynamics and the values (numbers in each grid cell), write down the policy corresponding to taking the greedy action with respect to the values in each state. Create a grid with the same dimension as the figure and draw an arrow in each square denoting the greedy action.



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0