Value Functions & Bellman Equations

Rupam Mahmood

January 27, 2020

The goal of a bandit agent

Maximize expected reward R

$$\pi(a) = P(A = a)$$

$$v_\pi = E_\pi[R] = E_\pi[E[R \mid A]] = E_\pi[q_*(A)]$$

Choose policy π **that maximizes** v_π

The goal of an agent

Contextual Bandits

Maximize expected reward *R* **for all state** *S*

$$\pi(a \mid s) = P(A = a \mid S = s)$$

$$v_\pi(s) = E_\pi[R \mid S = s] = E_\pi[E[R \mid S = s, A]] = E_\pi[q_*(s, A)]$$

Choose policy π **that maximizes**
v_π **for all state** *S*

MDPs

Maximize expected sum of discounted future rewards *R* **from all states** *S*

Maximize expected return *G* **from all states** *S*

return: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$$= R_{t+1} + \gamma G_{t+1}$$

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

Choose policy π **that maximizes**
v_π **for all state** *S*

State-value functions w.r.t. action-value functions

Contextual Bandits

$$v_\pi(s) = E_\pi[R \mid S = s]$$

$$= E_\pi[E[R \mid S = s, A]] = E_\pi[q_*(s, A)]$$

Law of the unconscious statistician: $E[\, g(X) \,] = \sum P(X{=}x)\, g(x)$

$$= \sum_a P(A_t = a \mid S_t = s) q_*(s, a)$$

$$= \sum_a \pi(a \mid s) q_*(s, a)$$

MDPs

state-value function:

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

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

$$= \sum_a P(A_t = a \mid S_t = s) q_\pi(s, a)$$

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

action-value function:

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

The Bellman equation for v_π

return: $G_t = R_{t+1} + \gamma G_{t+1}$

state-value function:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]; \text{ for all } s$$

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

Optimal policies & values

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

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

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

An optimal policy: $\pi_*(a \,|\, s) = 1$ **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

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

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

Writing action-value functions wrt state-value functions

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

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

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

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

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

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

The Bellman equation for q_π

Bellman equation with expected reward

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

$$\begin{aligned}
\sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r \quad &= \sum_a \pi(a \mid s) \sum_r r \sum_{s'} P(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\
&= \sum_a \pi(a \mid s) \sum_r r P(R_{t+1} = r \mid S_t = s, A_t = a) \\
&= \sum_a \pi(a \mid s) E[R_{t+1} \mid S_t = s, A_t = a] \\
&= \sum_a \pi(a \mid s) r(s, a)
\end{aligned}$$

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