Exercise 3.7. What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state-action pair (s, a). As a hint, the backup diagram corresponding to this equation is given in Figure 1 (right). Show the sequence of equations analogous to (3.12), but for action values.

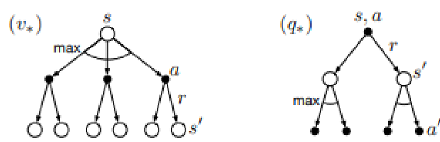

Figure 3.4: Backup diagrams for v_* and q_*

Figure 1: Backup diagrams for v_* and q_*.

Nguyen Tuan Ngoc `credit by: Tianlin Liu`

Solution. One direct corollary of the Conditional Expectation Theorem is that, informally, we have

$$\mathbb{E}[X \mid \text{info}] = \sum_i \mathbb{E}[X \mid \text{info}, F_i]\mathbb{P}(F_i \mid \text{info}).$$

http://www.stat.yale.edu/ pollard/Courses/600.spring08/Handouts/elem.conditioning.pdf

Before finding out what is $q_\pi(s', a')$, first we take a closer look at how $v_\pi(s)$ is deducted in the text:

$$\begin{aligned}
v_\pi(s) &:= \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right] \\
&= \underbrace{\mathbb{E}_\pi\left[R_{t+1} \mid S_t = s\right]}_{\text{Part 1}} + \underbrace{\mathbb{E}_\pi\left[\gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]}_{\text{Part 2}}
\end{aligned}$$

$$\begin{aligned}
\text{Part 1} :&= \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s \right] \\
&= \sum_a \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s, A_t = a \right] \underbrace{P(A_t = a \mid S_t = s)}_{\pi(a|s)} \\
&= \sum_a \pi(a \mid s) \sum_r r P(R_{t+1} = r \mid S_t = s, A_t = a) \\
&= \sum_a \pi(a \mid s) \sum_{s',r} r p(s', r \mid s, a).
\end{aligned}$$

$$\begin{aligned}
\text{Part 2} :&= \mathbb{E}_\pi \left[\gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right] \\
&= \sum_a \sum_{s'} \underbrace{\mathbb{E}_\pi \left[B \mid S_t = s, A_t = a, S_{t+1} = s' \right]}_{v_\pi(s')} \pi(a \mid s) P(S_{t+1} = s' \mid S_t = s, A_t = a) \\
&= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) v_\pi(s').
\end{aligned}$$

Thus,

$$\begin{aligned}
v_\pi(s) &= \text{Part 1} + \text{Part 2} \\
&= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \forall s \in \mathcal{S}.
\end{aligned}$$

Now, using the exactly the same idea

$$\begin{aligned}
q_\pi(s, a) &= \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right] \\
&= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \\
&= \mathbb{E} \left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a \right] \\
&= \underbrace{\mathbb{E} \left[R_{t+1} \mid S_t = s, A_t = a \right]}_{\text{Part 1}} + \underbrace{\mathbb{E} \left[\gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a \right]}_{\text{Part 2}}
\end{aligned}$$

$$
\begin{aligned}
\text{Part 1} &:= \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right] \quad \textit{(Expectation of immediate reward)} \\
&= \sum_r r P(R_{t+1} = r \mid S_t = s, A_t = a) \quad \textit{(Definition of expectation)} \\
&= \sum_r r \sum_{s'} \underbrace{P(R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a)}_{\text{Joint probability of reward and next state}} \quad \textit{(Marginalizing over next state } s'\textit{)} \\
&= \sum_r r \sum_{s'} \underbrace{P(R_{t+1} = r \mid S_t = s, A_t = a, S_{t+1} = s')}_{\text{Reward model}} \underbrace{P(S_{t+1} = s' \mid S_t = s, A_t = a)}_{\text{State transition model}} \quad \textit{(Cpf)} \\
&= \sum_{s'} \underbrace{P(S_{t+1} = s' \mid S_t = s, A_t = a)}_{\text{State transition probability}} \sum_r r \underbrace{P(R_{t+1} = r \mid S_t = s, A_t = a, S_{t+1} = s')}_{\text{Reward probability given next state}} \quad \textit{(Rs)} \\
&= \sum_{s',r} r \underbrace{p(s', r \mid s, a)}_{\text{Joint probability of next state and reward}} \quad \textit{(Using joint probability def for env dynamics).}
\end{aligned}
$$

Notation:

- **Cpf:** Conditional probability factorization
- **Rs:** Rearrange summations

$$
\begin{aligned}
\text{Part 2} &:= \mathbb{E}\left[\underbrace{\gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2}}_{:=B} \mid S_t = s, A_t = a\right] \\
&= \sum_{s'} \mathbb{E}\left[B \mid S_t = s, A_t = a, S_{t+1} = s'\right] P(S_{t+1} = s' \mid S_t = s, A_t = a) \\
&= \sum_{s'} \sum_{a'} \underbrace{\mathbb{E}\left[B \mid S_t = s, A_t = a, S_{t+1} = s', A_{t+1} = a'\right]}_{\gamma q_\pi(s', a')} p(s \mid s, a) \underbrace{p(a' \mid s')}_{\pi(a' \mid s')} \\
&= \sum_{s'} p(s' \mid s, a) \sum_{a'} \gamma q_\pi(s', a') \pi(a' \mid s') \\
&= \sum_{s',r} p(s', r \mid s, a) \sum_{a'} \gamma q_\pi(s', a') \pi(a' \mid s').
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s, a) &= \text{Part 1} + \text{Part 2} \\
&= \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \sum_{a'} q_\pi(s', a') \pi(a' \mid s')\right].
\end{aligned}
$$