# Markov Decision Process (2)

Ali Gooya

Reference for slides:

[David Silver (Deep Mind)]

# Learning outcomes

- Extend the value function for MDPs
- Establish Bellman equations for MDPs
- Define optimal value functions and policies
- Introduce Bellman's optimality equations

# Value Function for MDP

## Definition

The *state-value function* $v_\pi(s)$ of an MDP is the expected return starting from state $s$, and then following policy $\pi$

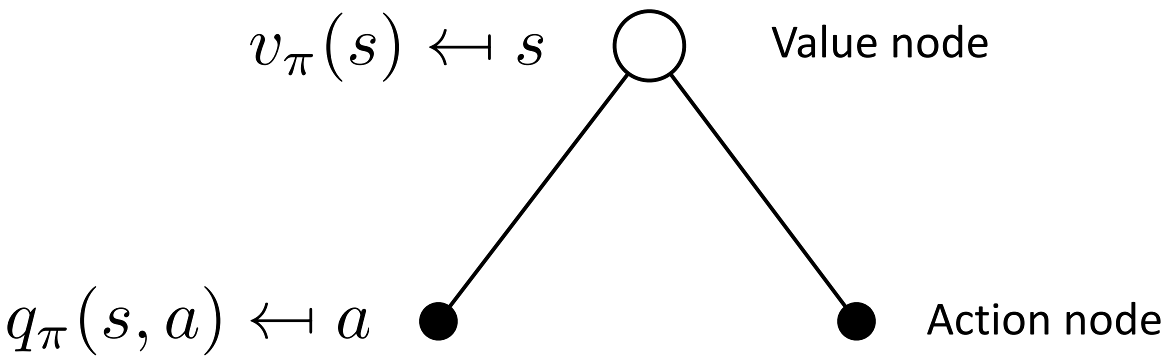
$$v_\pi(s) = \mathbb{E}_\pi \left[ G_t \mid S_t = s \right]$$

## Definition

The *action-value function* $q_\pi(s, a)$ is the expected return starting from state $s$, taking action $a$, and then following policy $\pi$

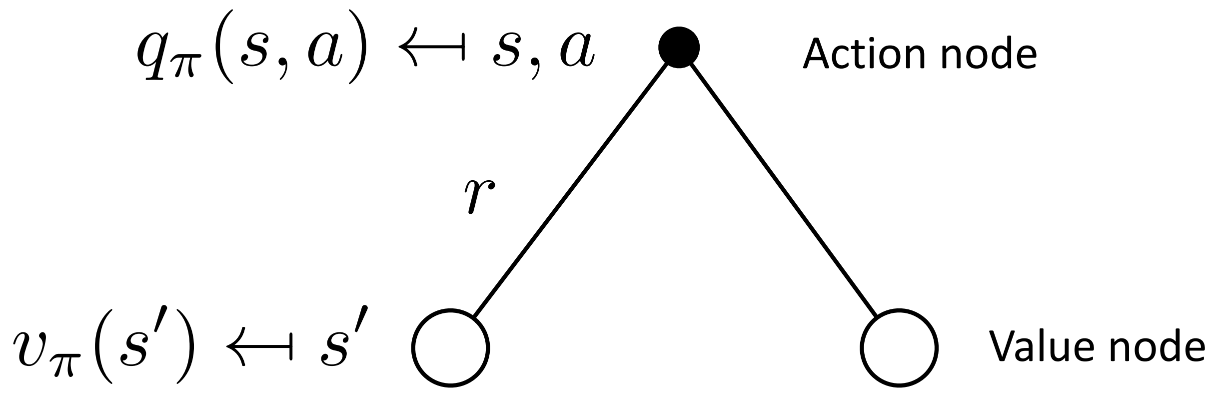
$$q_\pi(s, a) = \mathbb{E}_\pi \left[ G_t \mid S_t = s, A_t = a \right]$$

# Recurrent forms of $v_\pi$ and $q_\pi$ (i)




$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t | s_t = s] \\
&= \sum_a \pi(a|s) q_\pi(s,a)
\end{aligned}
$$

# Recurrent forms of $v_\pi$ and $q_\pi$ (ii)




$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[G_t | s_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) | s_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} | s_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} | s_t = s, A_t = a] \\
&= \mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a v_\pi(s')
\end{aligned}
$$

# Recurrent form of $v_\pi$

- Substituing $q_\pi(s,a)$ in $v_\pi(s)$, we obtain:

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s)\Big(\mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a v_\pi(s')\Big) \qquad (\star) \\
&= \sum_a \pi(a|s)\mathcal{R}_s^a + \gamma \sum_{s'}\sum_a \pi(a|s) P_{ss'}^a v_\pi(s') \\
&= \mathcal{R}_s^\pi + \gamma \sum_{s'} \mathcal{P}_{ss'}^\pi v_\pi(s')
\end{aligned}
$$

- Which reduces the MDP to MRP, giving the Bellman equation:

$$
\begin{aligned}
\mathsf{v}_\pi &= \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathsf{v}_\pi \\
\mathsf{v}_\pi &= [\mathsf{I} - \gamma \mathcal{P}^\pi]^{-1} \mathcal{R}^\pi
\end{aligned}
$$

# Recurrent form for $q_\pi$

- Substituing $v_\pi(s)$ in $q_\pi(s,a)$ we obtain:

$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a \left( \sum_{a'} \pi(a'|s') q_\pi(s',a') \right)$$

- Which can be solved likewise Bellman's equation for value functions.

- However, we want to solve for **optimal** action/value functions, which is not addressed above.

# Optimal value function

## Definition

The *optimal state-value function* $v_*(s)$ is the maximum value function over all policies

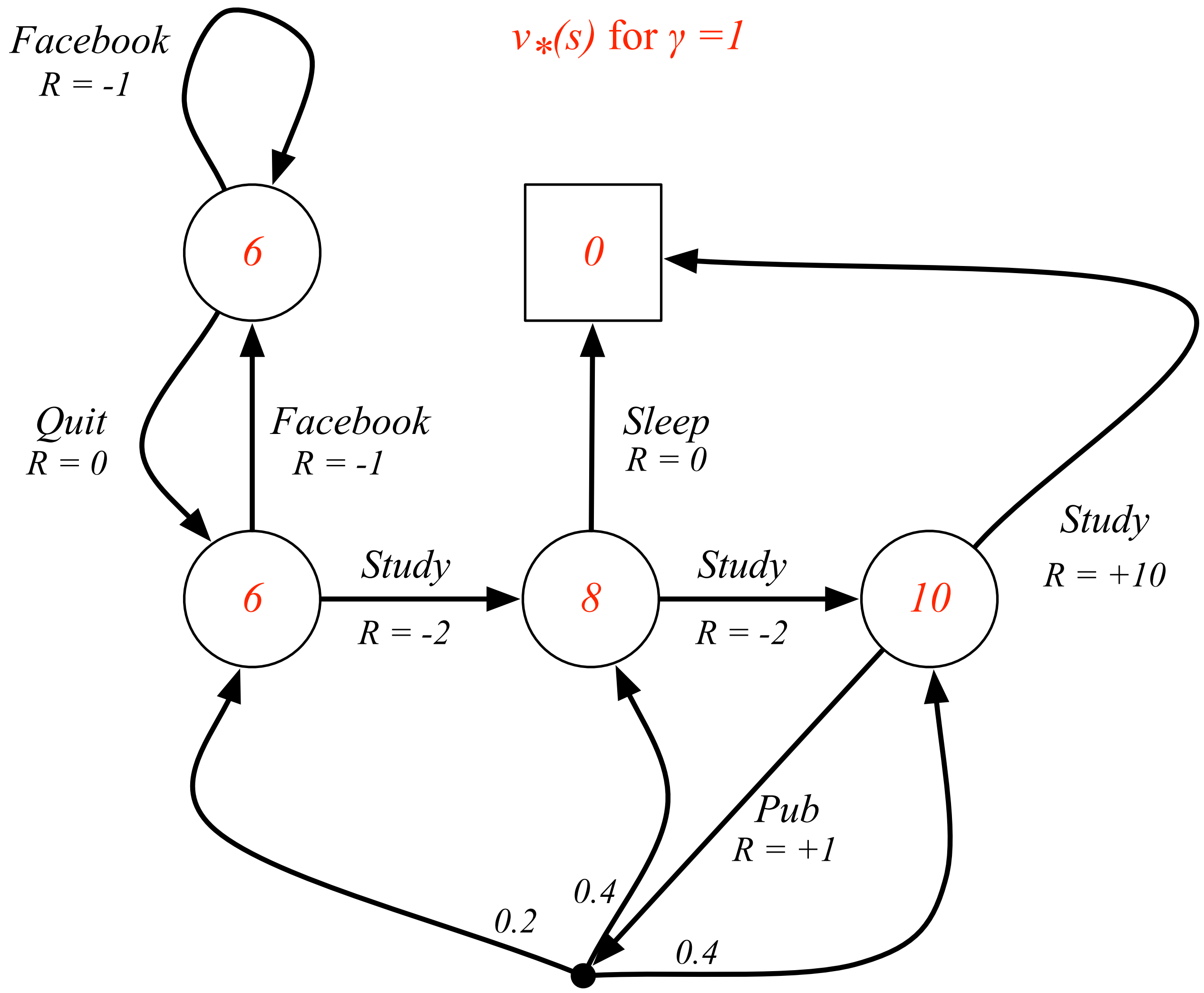
$$v_*(s) = \max_\pi v_\pi(s)$$

The *optimal action-value function* $q_*(s, a)$ is the maximum action-value function over all policies

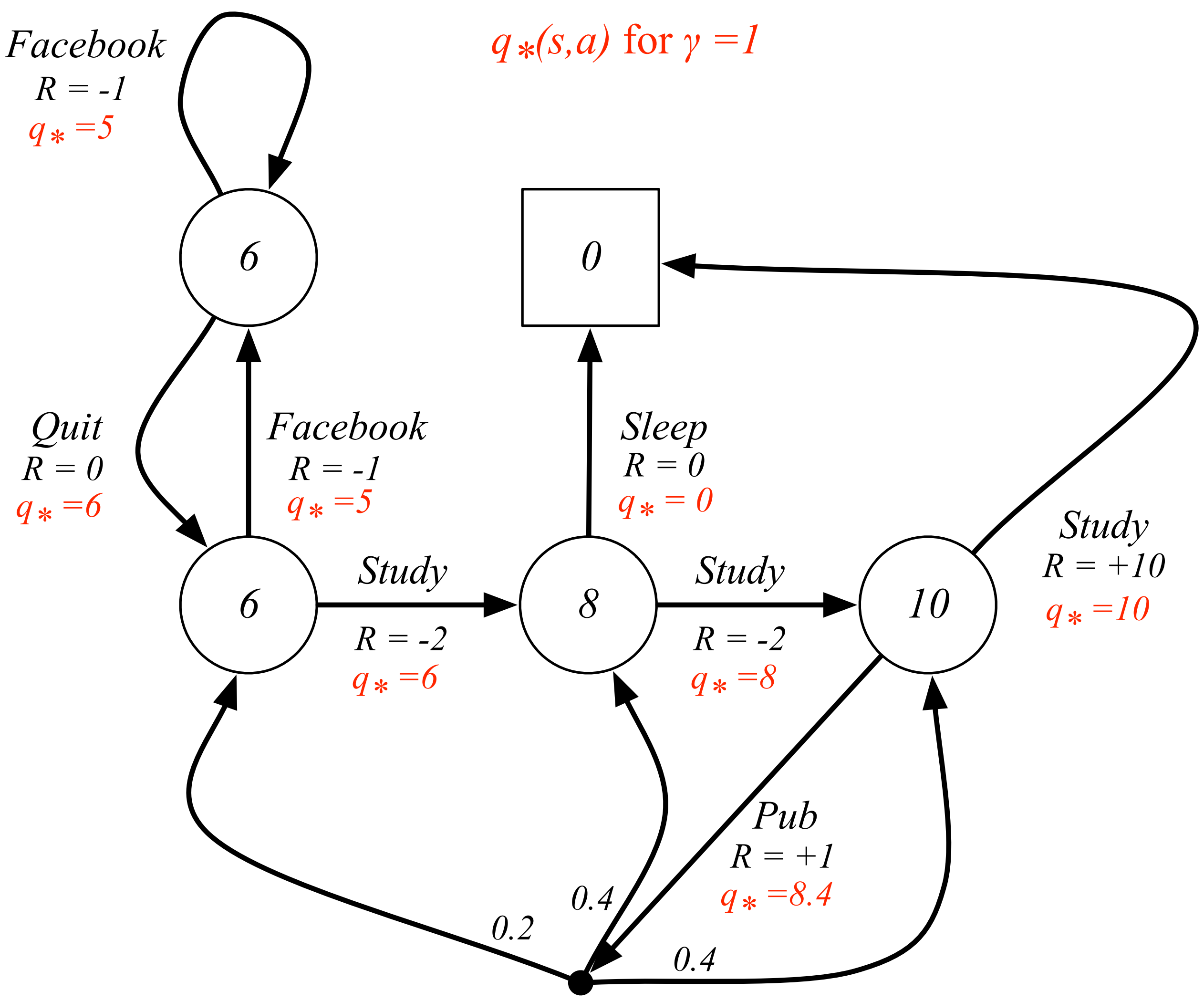
$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

- The optimal value function specifies the best possible performance in the MDP.
- An MDP is "solved" when we know the optimal value fn.

# Optimal value function for student MDP



# Optimal action-value function for the student MDP



# Optimal Policy $\pi_*$

Define a partial ordering over policies

$$\pi \geq \pi' \text{ if } v_\pi(s) \geq v_{\pi'}(s), \forall s$$

**Theorem**

*For any Markov Decision Process*

- *There exists an optimal policy $\pi_*$ that is better than or equal to all other policies, $\pi_* \geq \pi, \forall \pi$*
- *All optimal policies achieve the optimal value function, $v_{\pi_*}(s) = v_*(s)$*
- *All optimal policies achieve the optimal action-value function, $q_{\pi_*}(s, a) = q_*(s, a)$*

- Hence, to solve an MDP's $\pi_*$, we look for the optimal policy $v_*$

# Finding $v_*$ to discover optimal policy

- We start by

$$
\begin{aligned}
v_*(s) &= \max_\pi v_\pi(s) \\
&= \max_\pi \sum_a \pi(a|s) q_\pi(s,a) \\
&= \underbrace{1}_{\pi_*(a|s)} \times \max_a \underbrace{\{\max_\pi q_\pi(s,a)\}}_{q_*(s,a)} \\
&= \max_a q_*(s,a)
\end{aligned}
$$

Decompose policy in the current state ($\pi_*$) and successors ($q_*$)

- With the following deterministic policy at state s

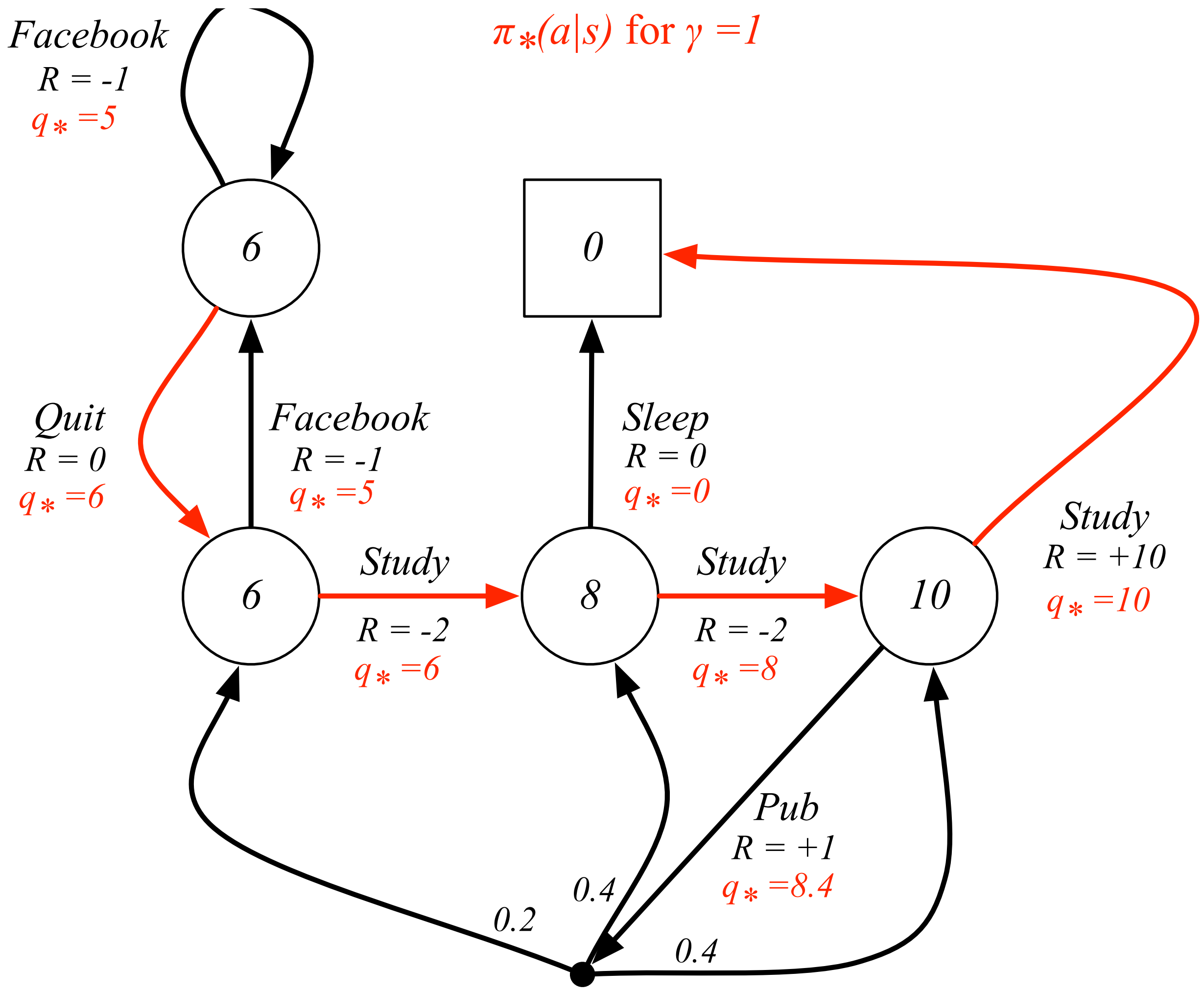
$$
\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \operatorname*{argmax}_{a \in \mathcal{A}} q_*(s,a) \\ 0 & \textit{otherwise} \end{cases}
$$

# Example of $\pi_*(a|s)$ in student MDP

- If we know $q_*(s,a)$, we can always find the optimal policy according to

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \underset{a \in \mathcal{A}}{\text{argmax}}\; q_*(s,a) \\ 0 & \textit{otherwise} \end{cases}$$

- So we try to find $q_*(s,a)$ using a recurrent form

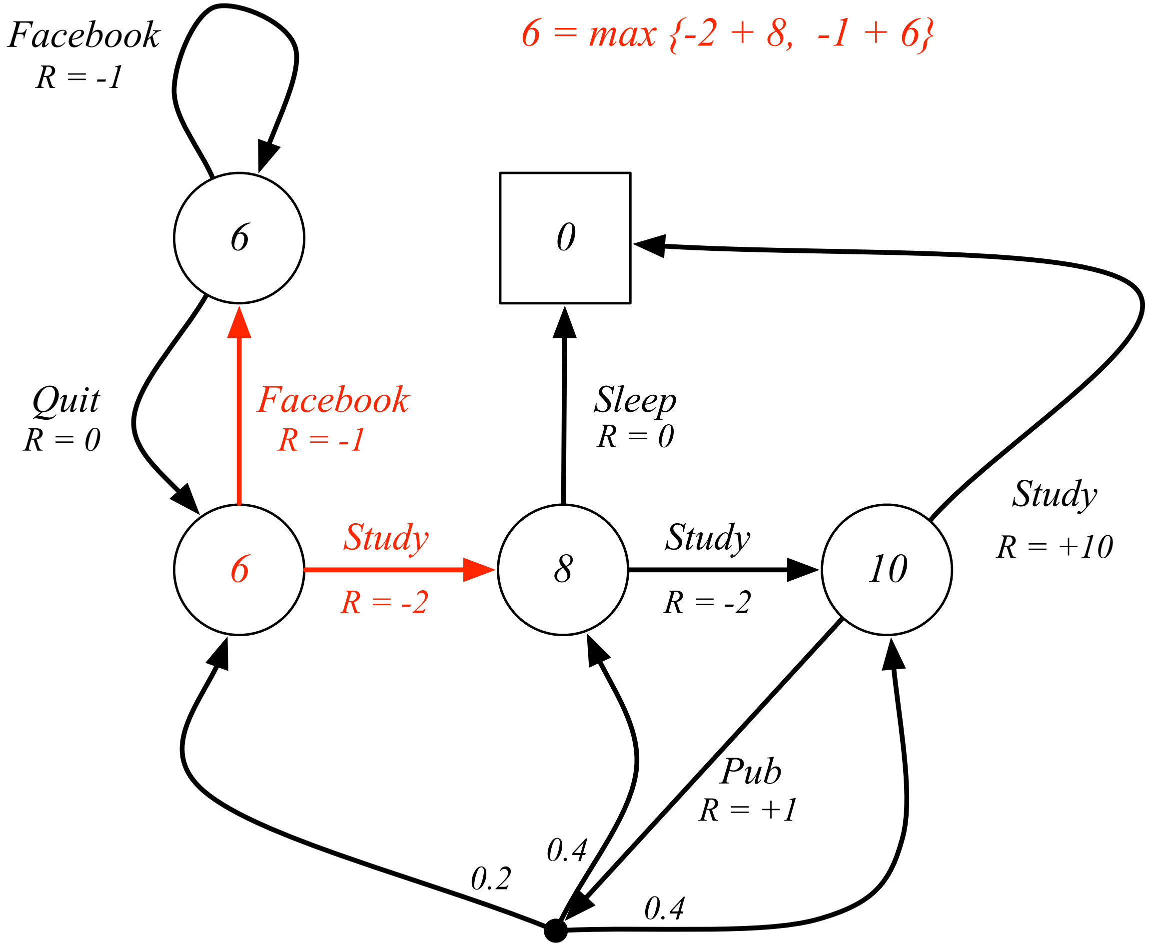


# Recurrent forms for $q_*(s, a)$ and $v_*(s)$

- Bellman optimality equations:

$$
\begin{aligned}
q_*(s,a) &= \max_{\pi}\{\mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a v_\pi(s')\} \\
&= \mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a v_*(s') \\
&= \mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} q_*(s', a')
\end{aligned}
$$

$$
v_*(s) = \max_{a}\{\mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a v_*(s')\}
$$



# Conclusions

- Bellman’s optimality equations are non-linear (due to Max operation)
- No closed form solution
- Iterative solution methods
  - Value iteration
  - Policy iteration
  - Q-learning
  - SARSA

- We will introduce these iterative methods in the rest of course.