Reinforcement learning

Leonardo Toffalini

2026-02-21

Outline

1. Informal introduction to RL

1.1 Characteristics of RL

What makes RL different?

- Not supervised (learning from labeled data)
- Not unsupervised (learning patterns in unlabeled data)
- Only *reward* signal
- Feedback may be delayed
- Heavily time dependent
- The agent has affect on the data

1.2 Examples of RL

- Robotics
- Video games
- Self-driving
- Finance
- Natural language processing (recently)
- Recommendation systems
- Many more…

2. Setting the scene

2.1 Markov decision processes

Definition 2.1.1 A Markov decision process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}_a, \mathcal{R}_a)$, where

- $\mathcal{S}$ is the set of all states
- $\mathcal{A}$ is the set of all possible actions
- $\mathcal{P}$ is the transition probability function

$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, \quad A_t = a]$$

- $\mathcal{R}$ is the reward function

$$\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, \quad A_t = a]$$

2.1 Markov decision processes

Definition 2.1.2 A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \dots, S_t].$$

Remark

In a Markov decision process, the successor state is *solely* influenced by the current state.

That is, an MDP has no memory of previous states.

Is this a limitation?

2.2 Illustrative step in an MDP

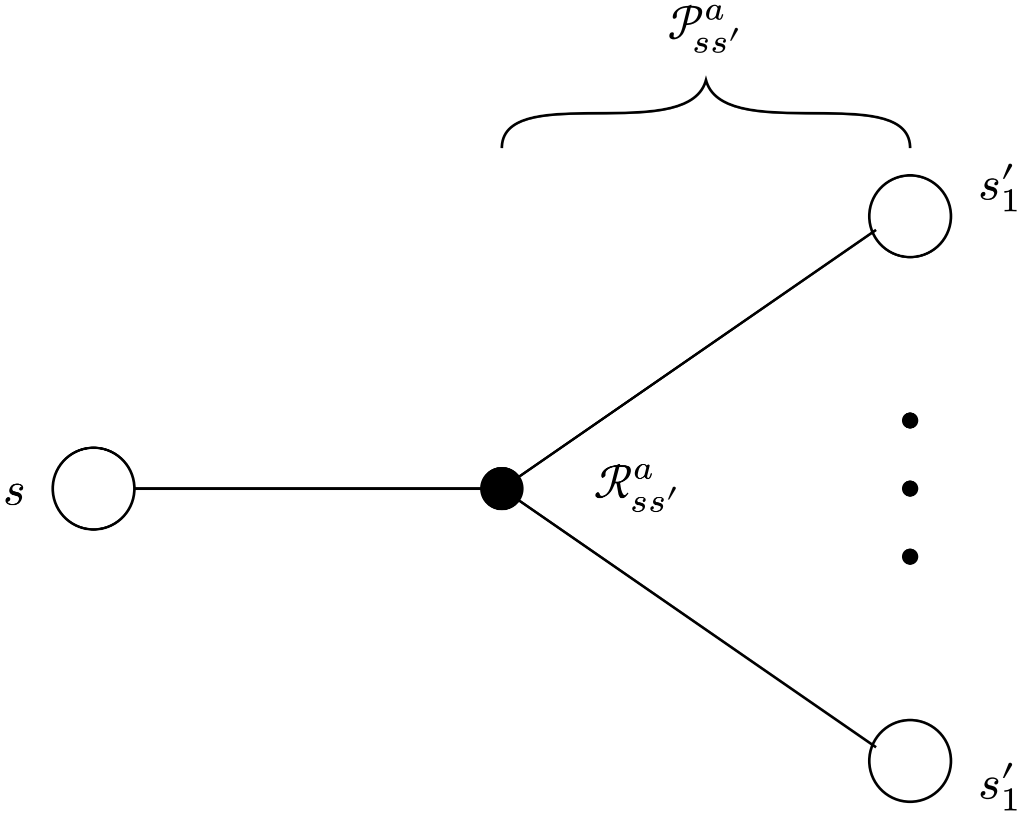


Figure 1: hollow circles = states, full circles = actions

2.3 Reward hypothesis

Definition 2.3.1 The reward hypothesis claims that all goals can be described by the maximization of expected cumulative reward.

Remark

Do you agree?

Can you find a counterexample?

2.4 Policy

Definition 2.4.1 A policy π is a distribution over the action space conditioned on the state space, that is

$$\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s].$$

Remark

The next action depends *only* on the current state and nothing else.

The policy may be deterministic if $\pi(a_i \mid s) = 1$ for some i and $\pi(a_j \mid s) = 0$ for any other $j \neq i$.

2.5 Value function

Definition 2.5.1 The return G_t is the total discounted reward from time step t, that is

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$

Remark

We can control the importance of *immediate* versus *future* rewards with $\gamma \in [0, 1]$.

2.5 Value function

Definition 2.5.2 The state-value function of a state s is the expected discounted return starting from state s, that is

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

Definition 2.5.3 The action-value function of a state-action pair (s, a) is the expected discounted return starting from state s and taking action a, that is

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

3. Bellman equations

3.1 Bellman expectation equations

Proposition 3.1.1

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

Proof.

$$\begin{aligned} v_\pi(s) &:= \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1}] \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] \end{aligned}$$

$\square$

3.1 Bellman expectation equations

Proposition 3.1.2

$$q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

Proof.

$$\begin{aligned} q_\pi(s,a) &:= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1}, A_{t+1}] \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \end{aligned}$$

$\square$

3.1 Bellman expectation equations

Proposition 3.1.3

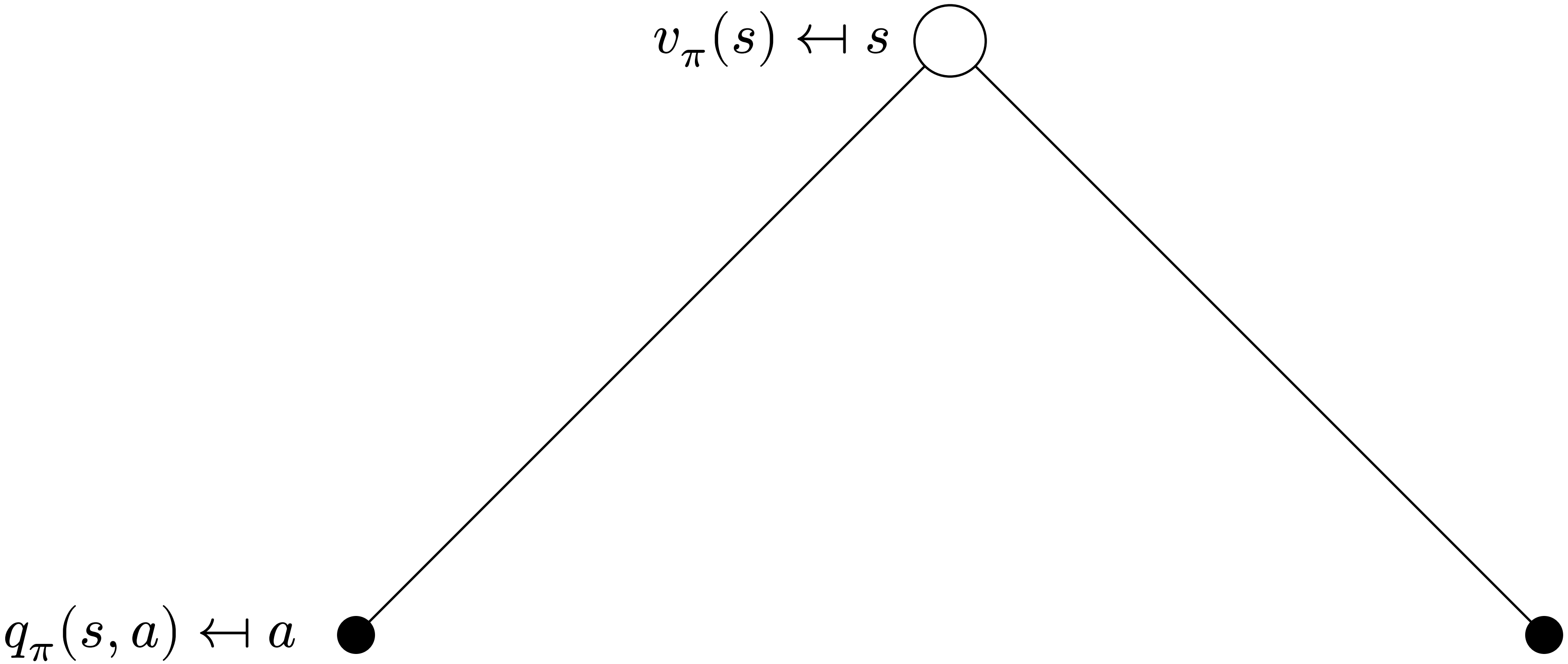
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$$

Proof.

$$\begin{aligned} v_\pi(s) &:= \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[\mathbb{E}_\pi[G_t \mid S_t, A_t] \mid S_t = s] \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a) \end{aligned}$$

□

3.2 Backup diagram



3.3 Bellman expectation equations

Proposition 3.3.1

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$

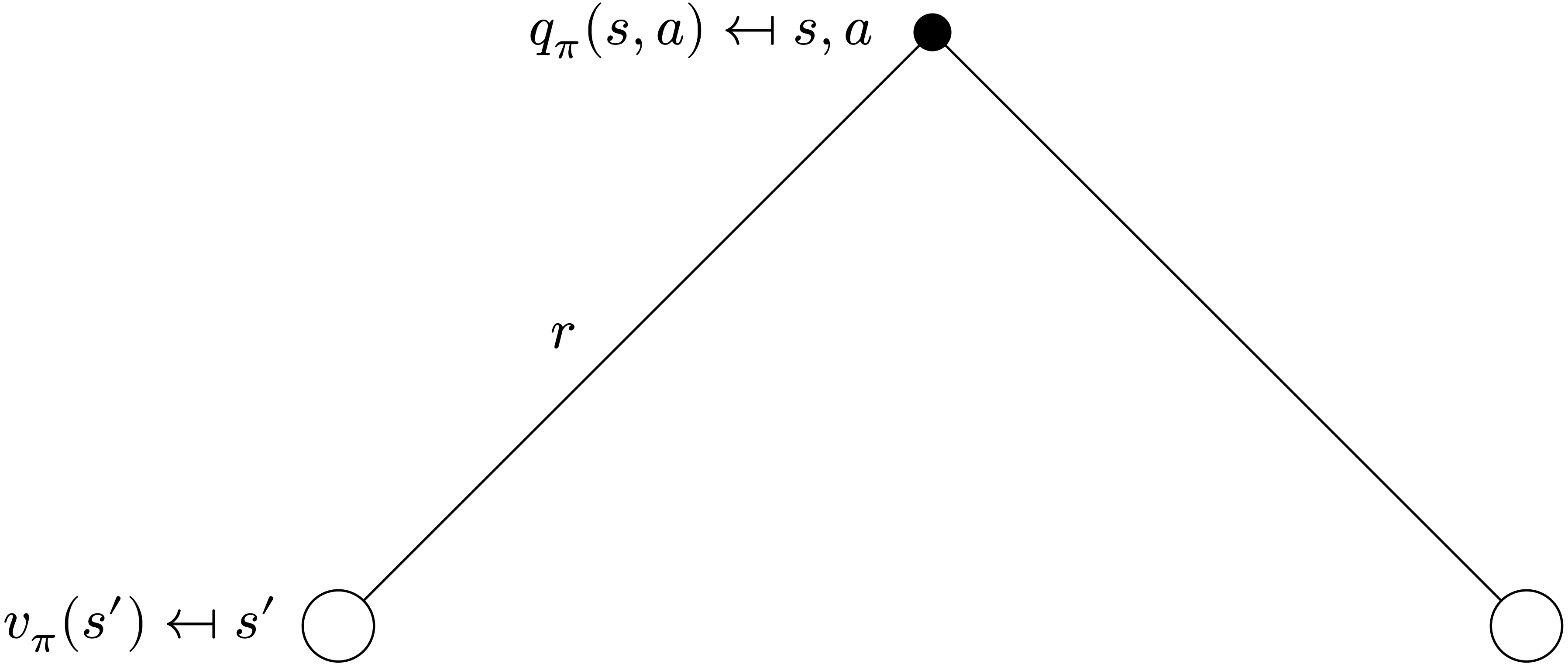
3.3 Bellman expectation equations

Proof.

$$
\begin{aligned}
q_\pi(s,a) &:= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\
&= \mathbb{E}_\pi\left[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\
&= \mathbb{E}_\pi\left[R_t + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s\right] \mid S_t = s, A_t = a\right] \\
&= \mathbb{E}_\pi\left[R_t + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right] \\
&= \mathcal{R}_s^a + \gamma \mathbb{E}_\pi\left[v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right] \\
&= \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')
\end{aligned}
$$

$\square$

3.4 Backup diagram



3.5 Bellman expectation equations

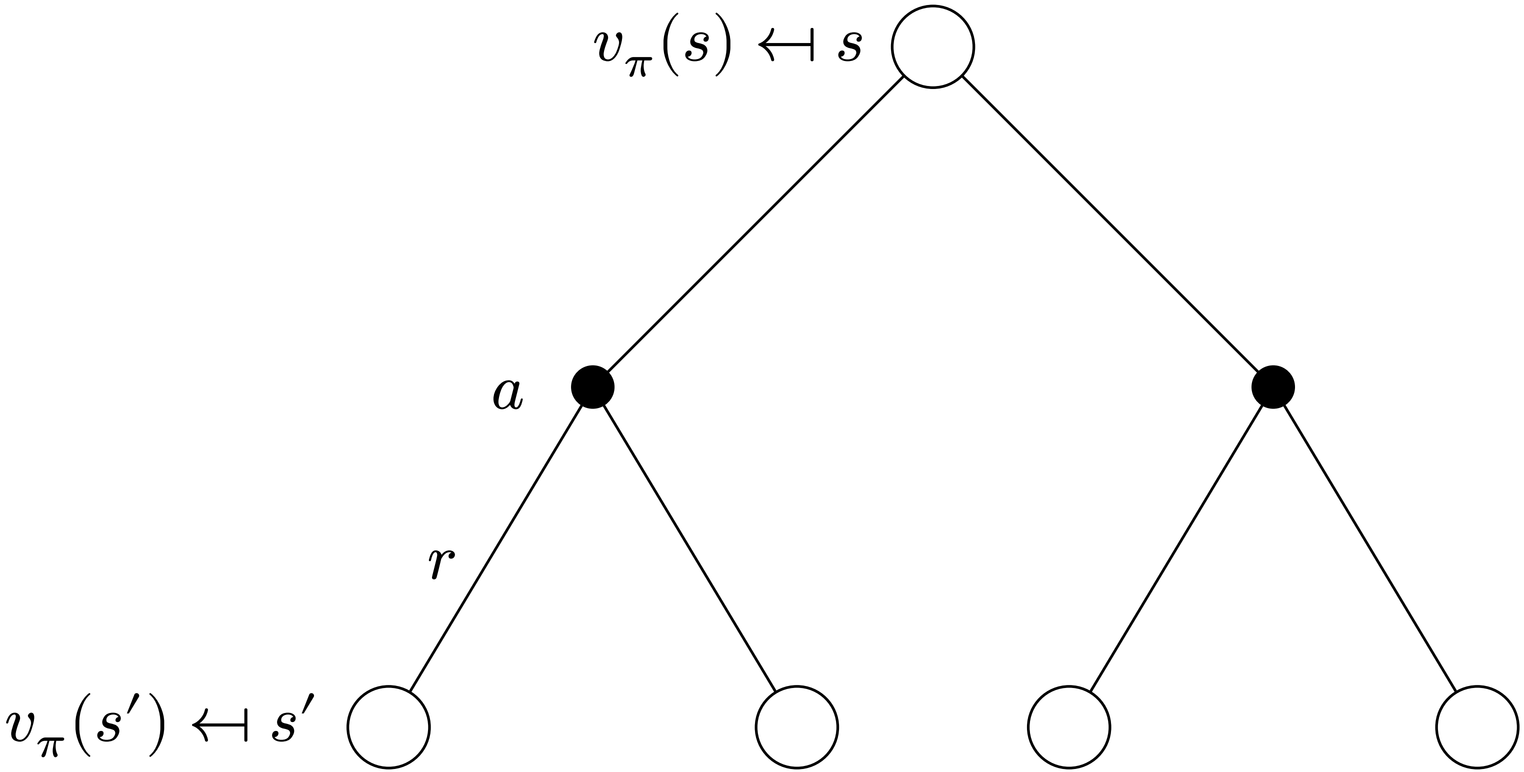
Proposition 3.5.1

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a v_\pi(s') \right)$$

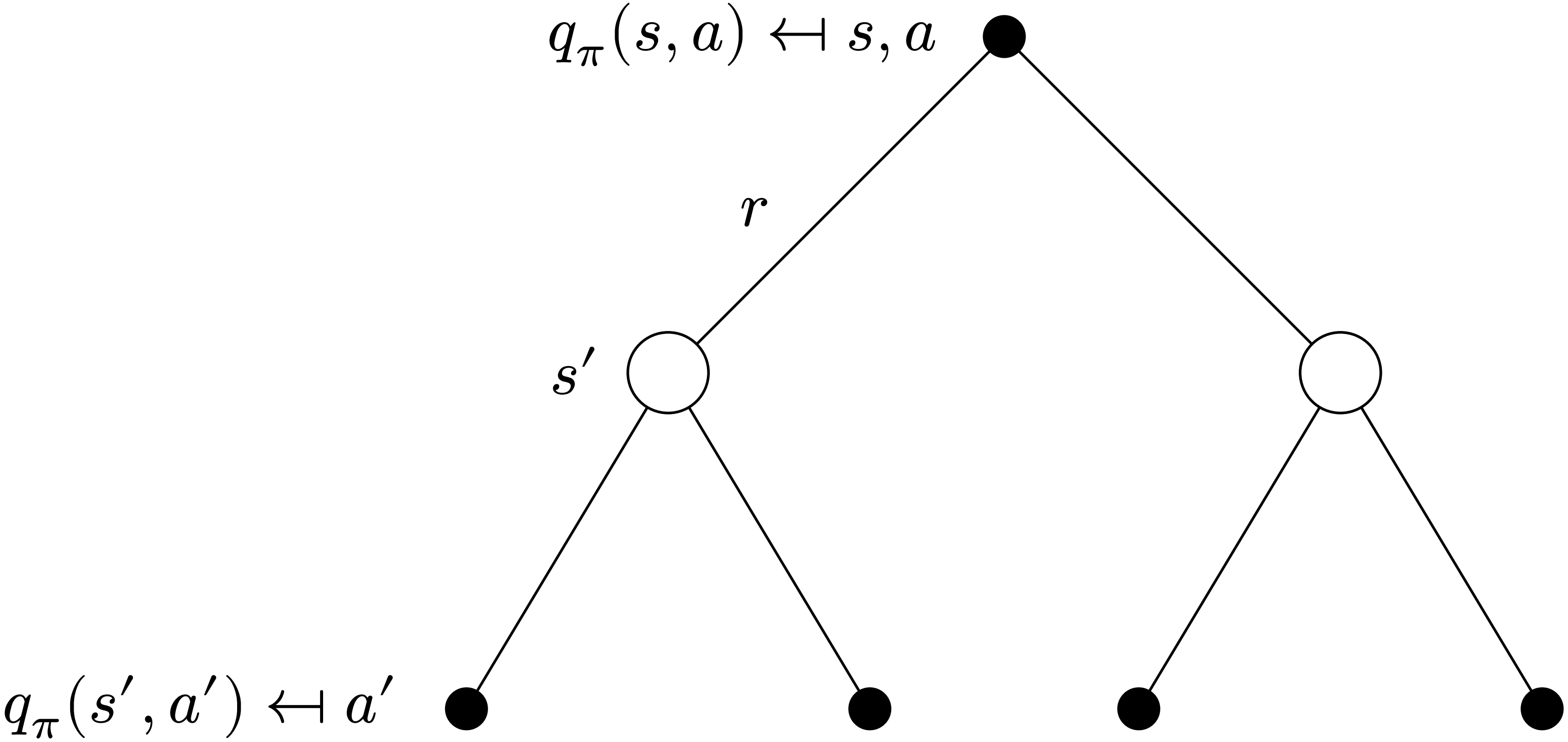
Proposition 3.5.2

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s', a')$$

3.6 Backup diagram



3.6 Backup diagram



3.7 Bellman optimality equations

Definition 3.7.1 The optimal state-value function is defined as

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

Definition 3.7.2 The optimal action-value function is defined as

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

3.7 Bellman optimality equations

Definition 3.7.3 A policy π^* is optimal if for all other policies π

$$v_{\pi^*}(s) \geq v_\pi(s) \quad \forall s \in \mathcal{S}.$$

Remark

That is, π^* is optimal if the value function induced by following π^* is optimal.

3.7 Bellman optimality equations

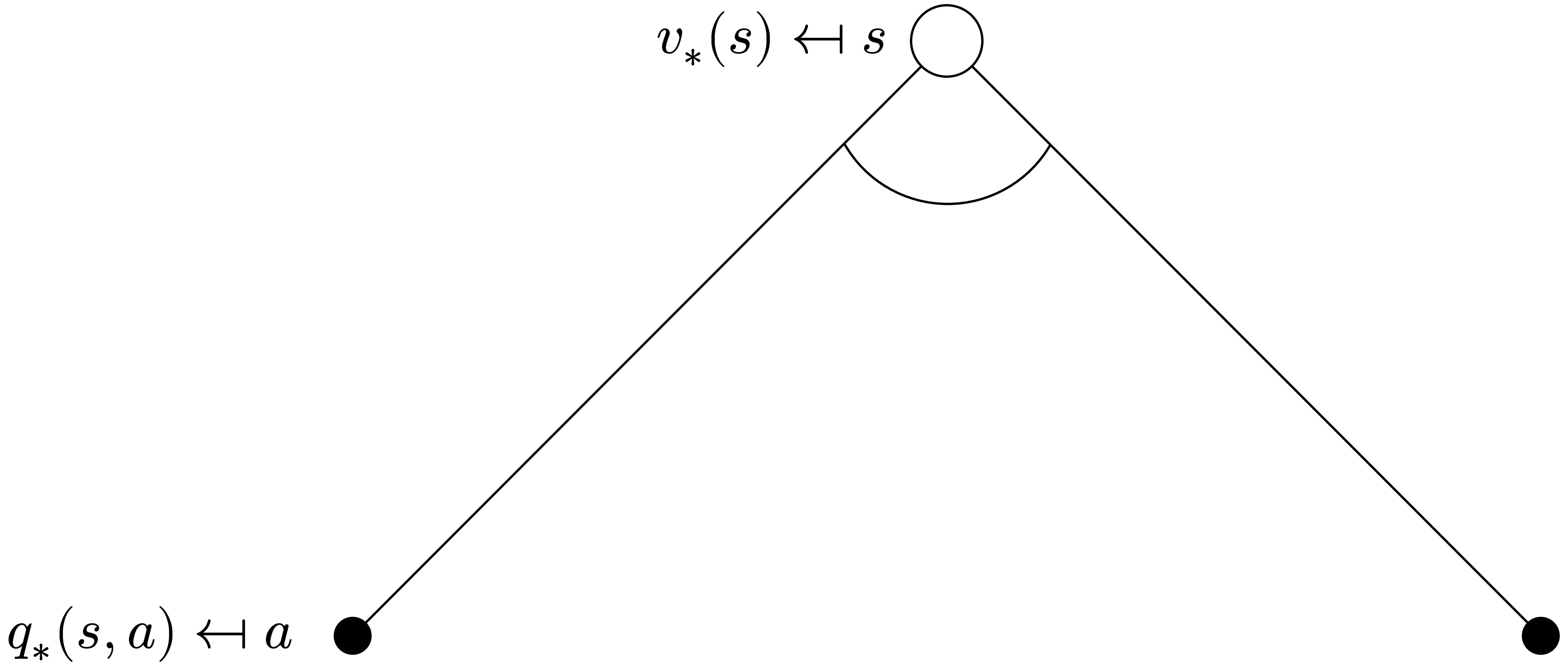
Proposition 3.7.4

$$v_*(s) = \max_a q_*(s, a)$$

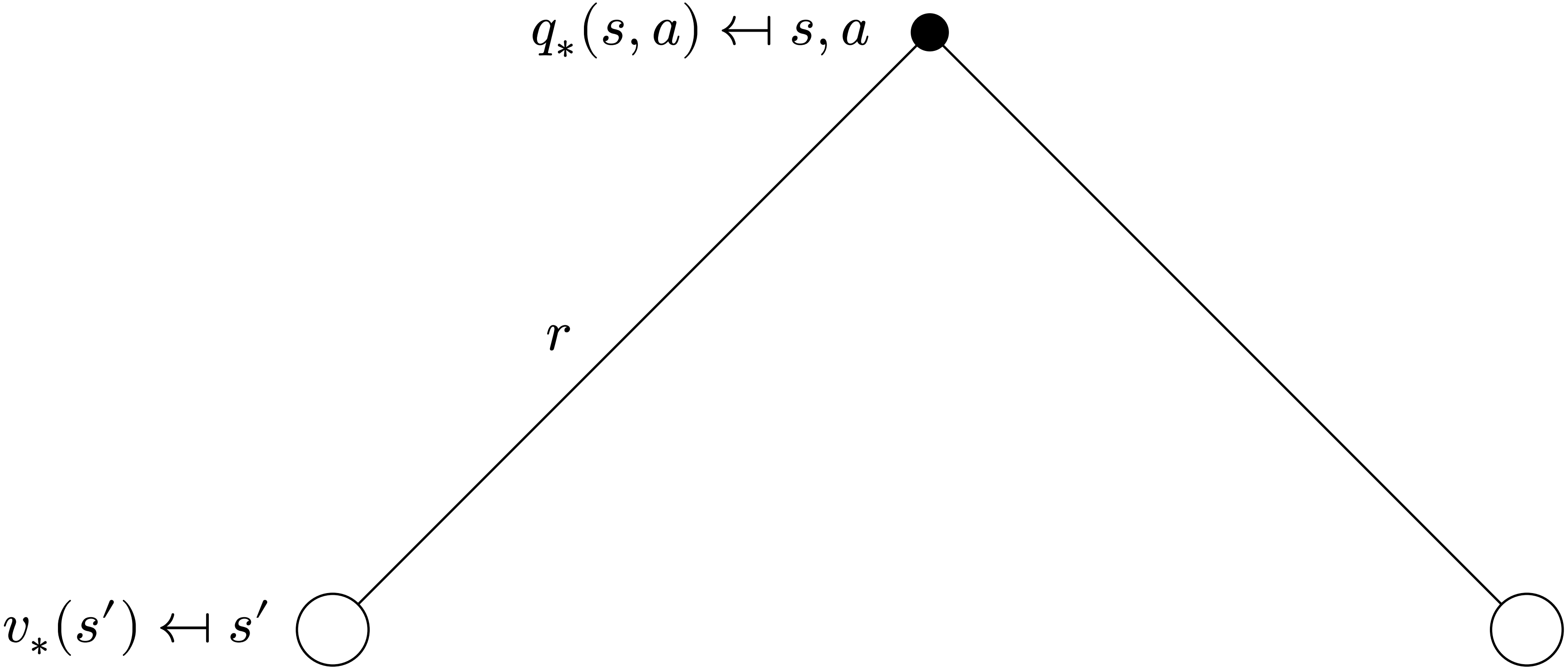
Proposition 3.7.5

$$q_*(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

3.8 Backup diagrams



3.8 Backup diagrams



3.9 Bellman optimality equations

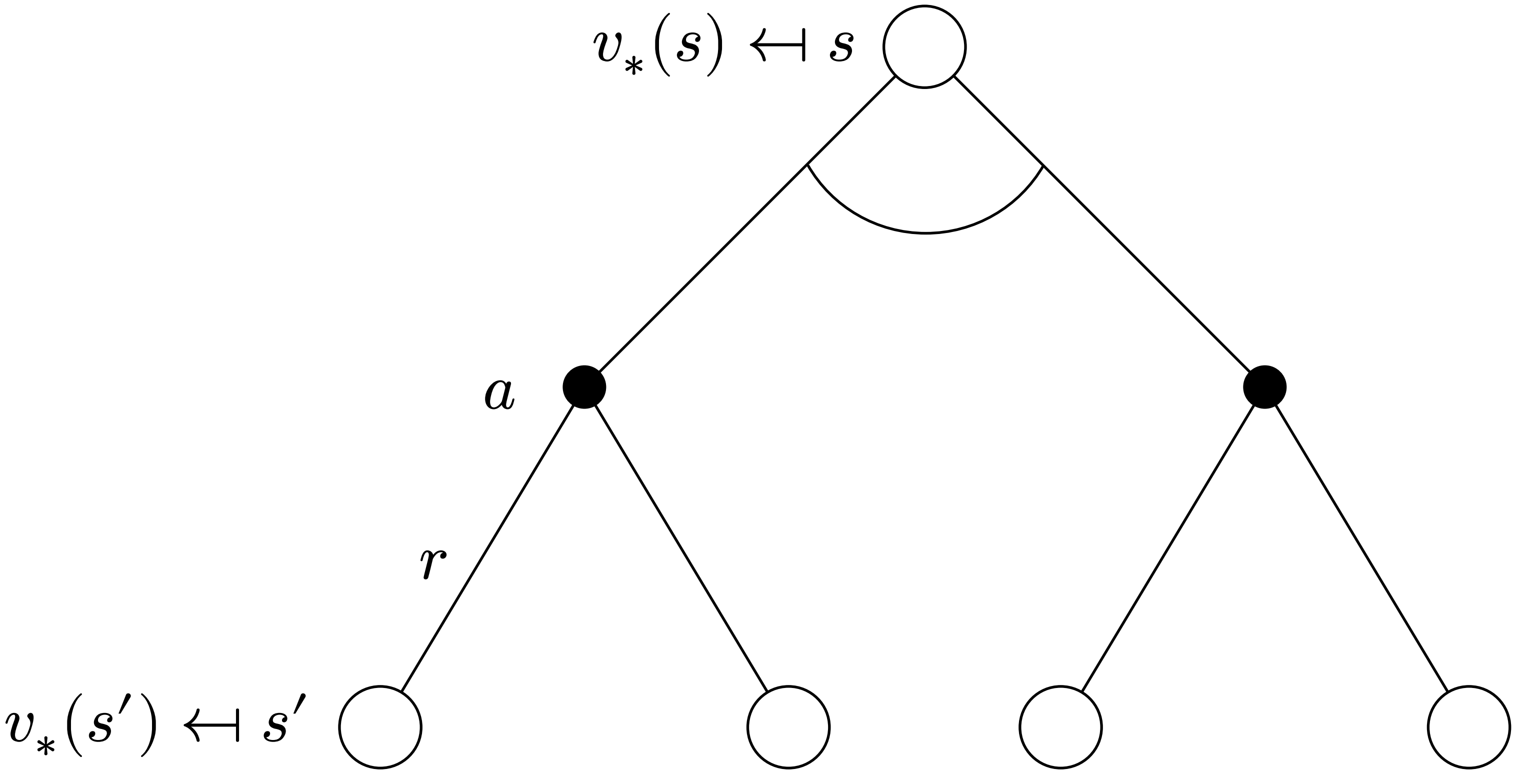
Proposition 3.9.1

$$v_*(s) = \max_a \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s') \right)$$

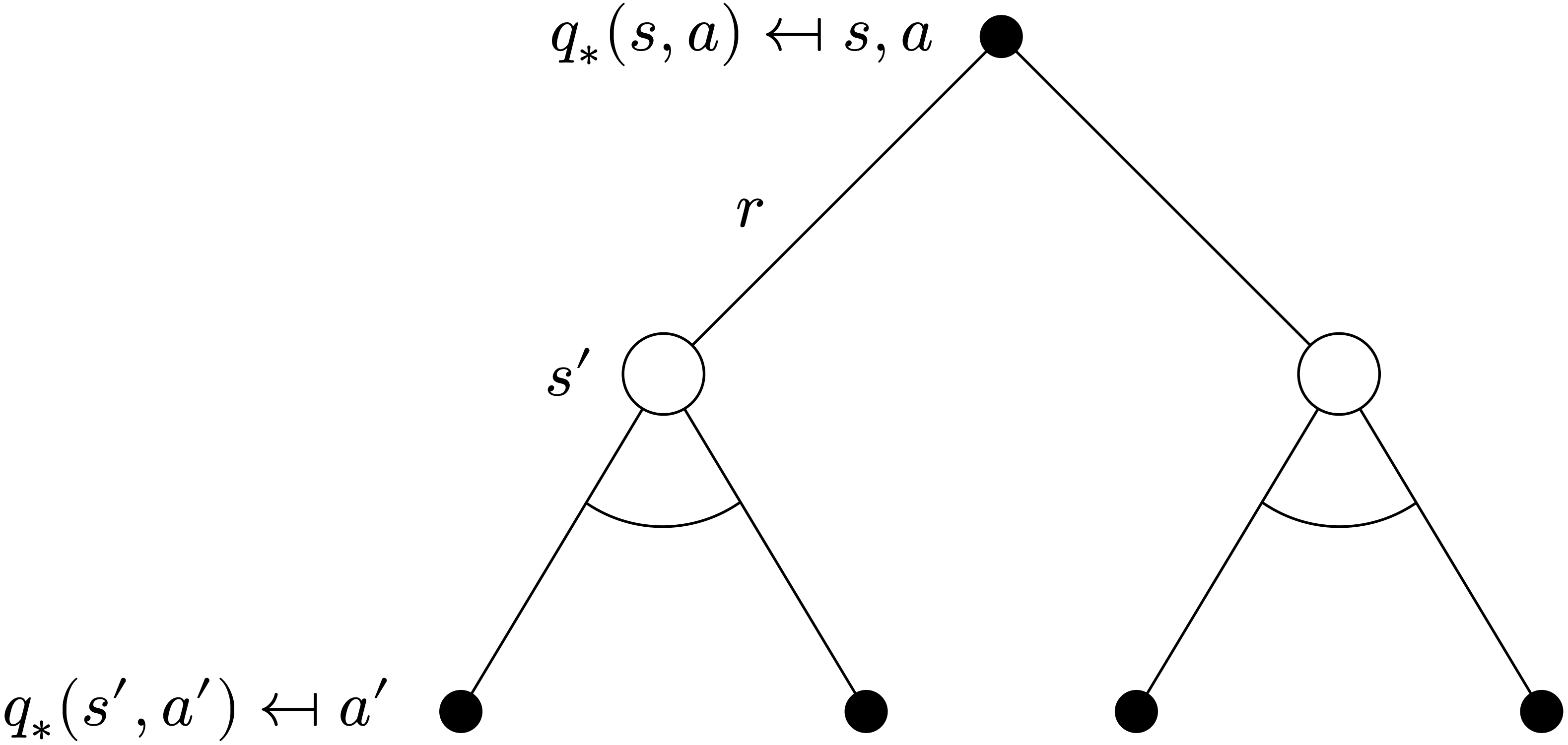
Proposition 3.9.2

$$q_*(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} q_*(s', a')$$

3.10 Backup diagrams



3.10 Backup diagrams



3.11 Onto the practice session

Now lets see what we can do with this theory.

Lets see some algorithms in the practice session…