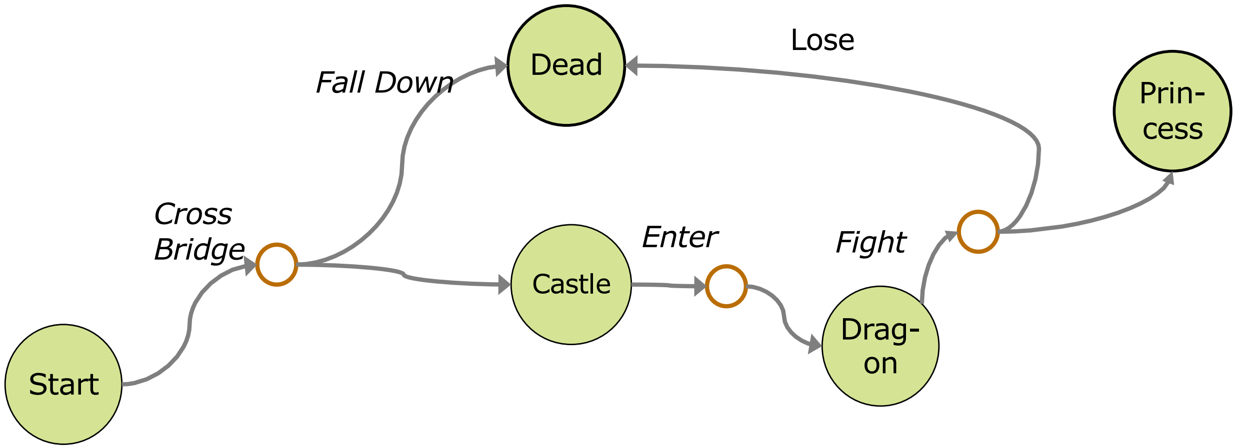
HSLU Lucerne University of Applied Sciences and Arts

Markov Decision Processes Summary



Reinforcement Learning
October 13, 2022

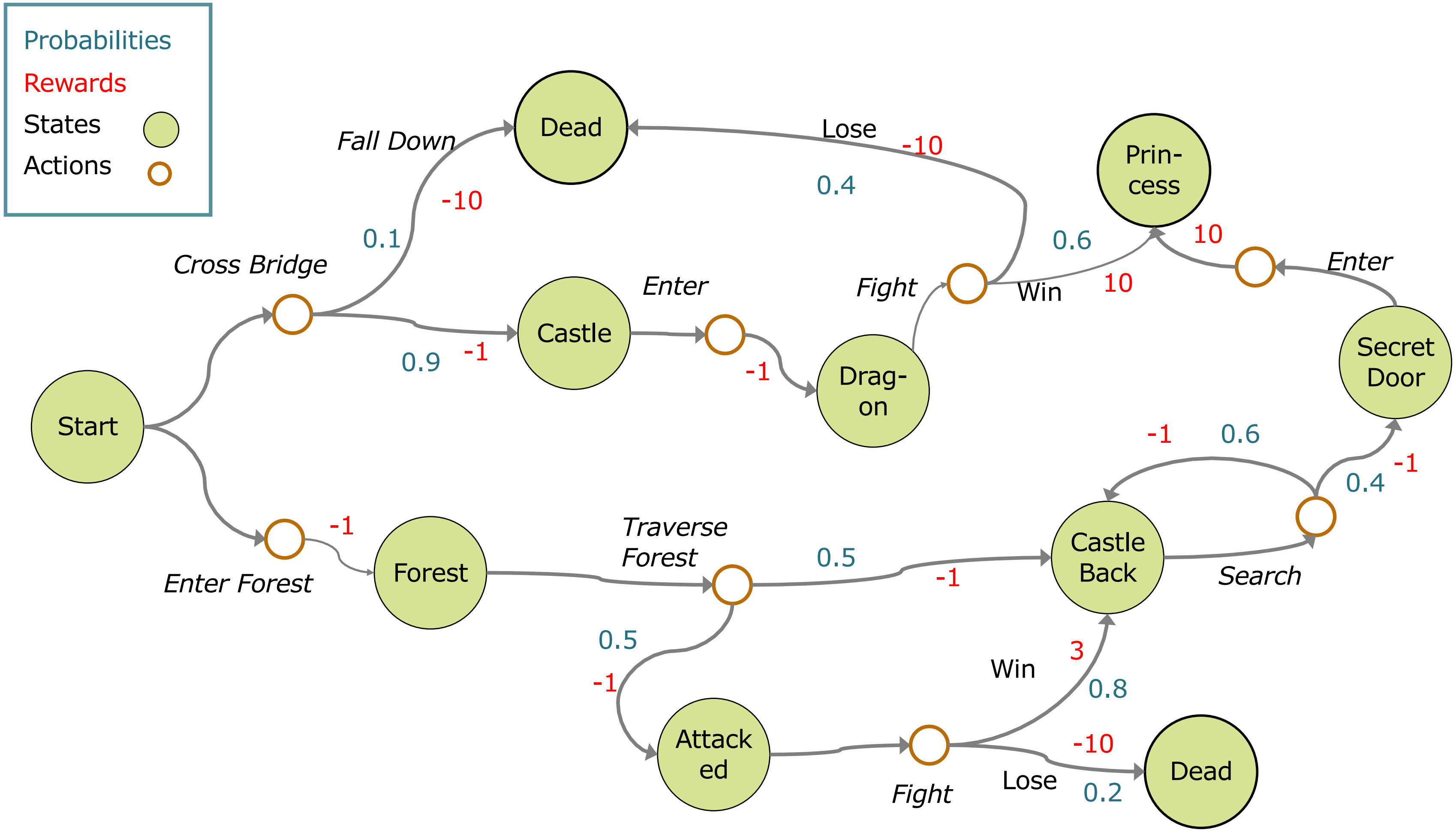
Definition of MDP

The dynamics of an MDP is defined as

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

(this can be viewed as a function of 4 parameters)

Markov Decision Process



Policy and Value Functions

A **policy** is a mapping from states to probabilities of selecting each possible action:

$$\pi(a|s) \doteq \Pr\{A_t = a|S_t = s\}$$

The **state-value function** of a state s under a policy π is the expected return by following π from s:

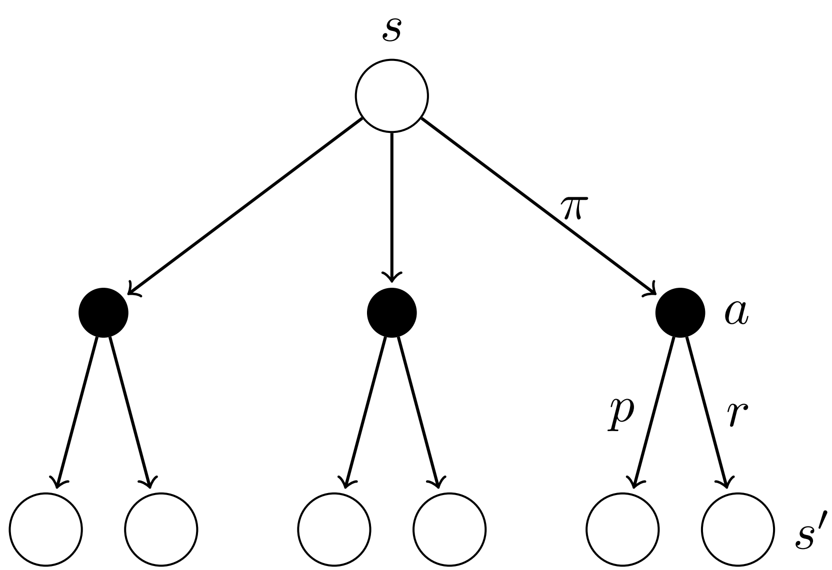
$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s], \text{ for all } s \in \mathcal{S}$$

The **action-value function** is the expected return by taking action a in state s and then following π:

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

Bellman Equation

$$v_\pi(s) \;=\; \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_\pi(s')\big], \;\; \text{for all } s \in \mathcal{S}$$



Optimal State-Value Function

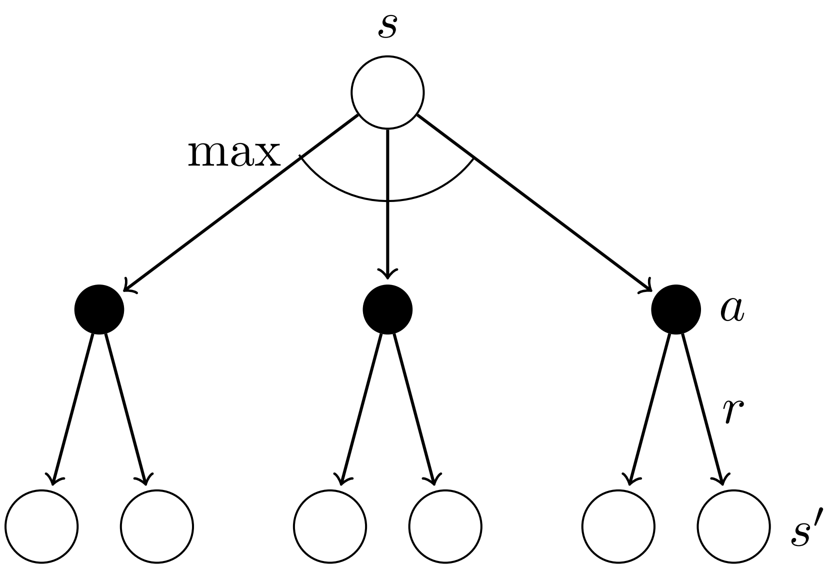
- the optimal state-value function is defined as

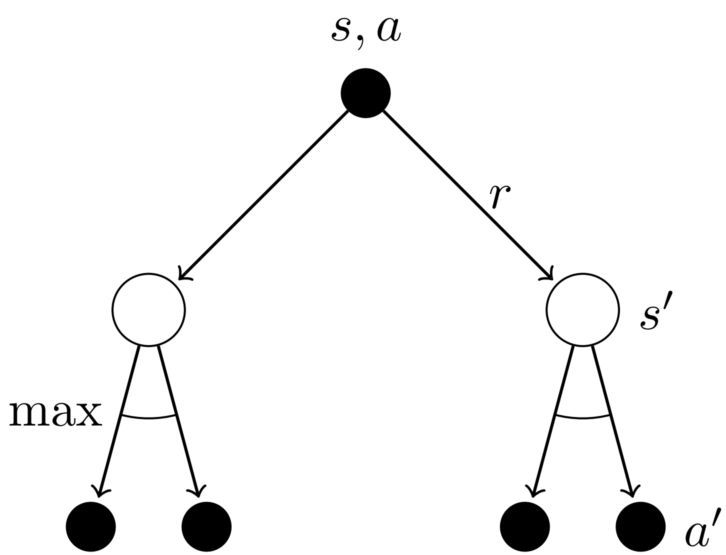
$$v_*(s) \doteq \max_{\pi} v_\pi(s)$$

- and the optimal action-value function as

$$q_*(s, a) \doteq \max_{\pi} q_\pi(s, a)$$

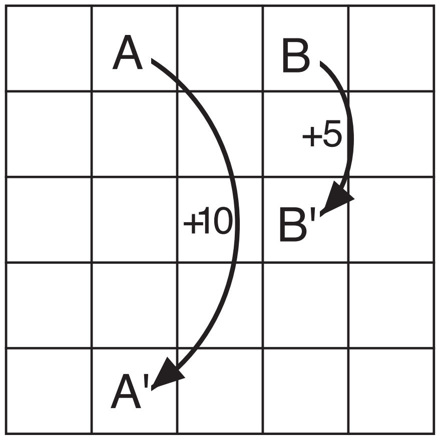
Backup diagrams for the optimal functions

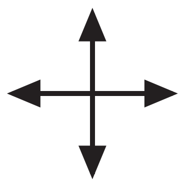




Examples of MDPs

- Grid world environment
- In state A (or B), the agent is transferred to the state A' (or B') with the indicated reward by any action
- Action that take the agent off the grid have reward -1, other actions 0



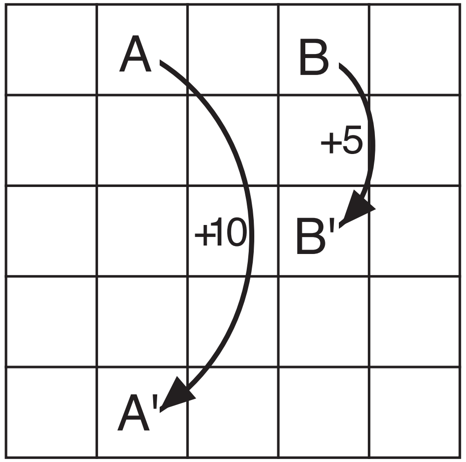


3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

value function for random policy and discount factor 0.9

Example: Gridworld

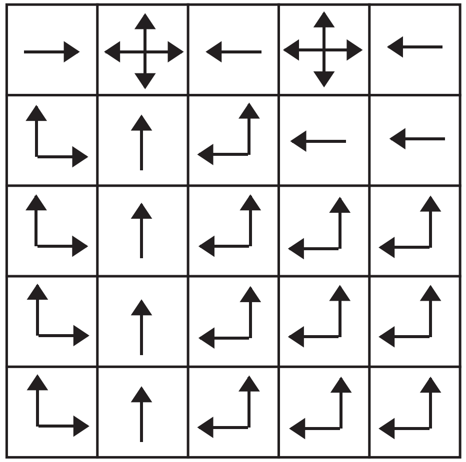
Optimal value function (and policy) for the gridworld problem, using the Bellman Equation



Gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



π_*