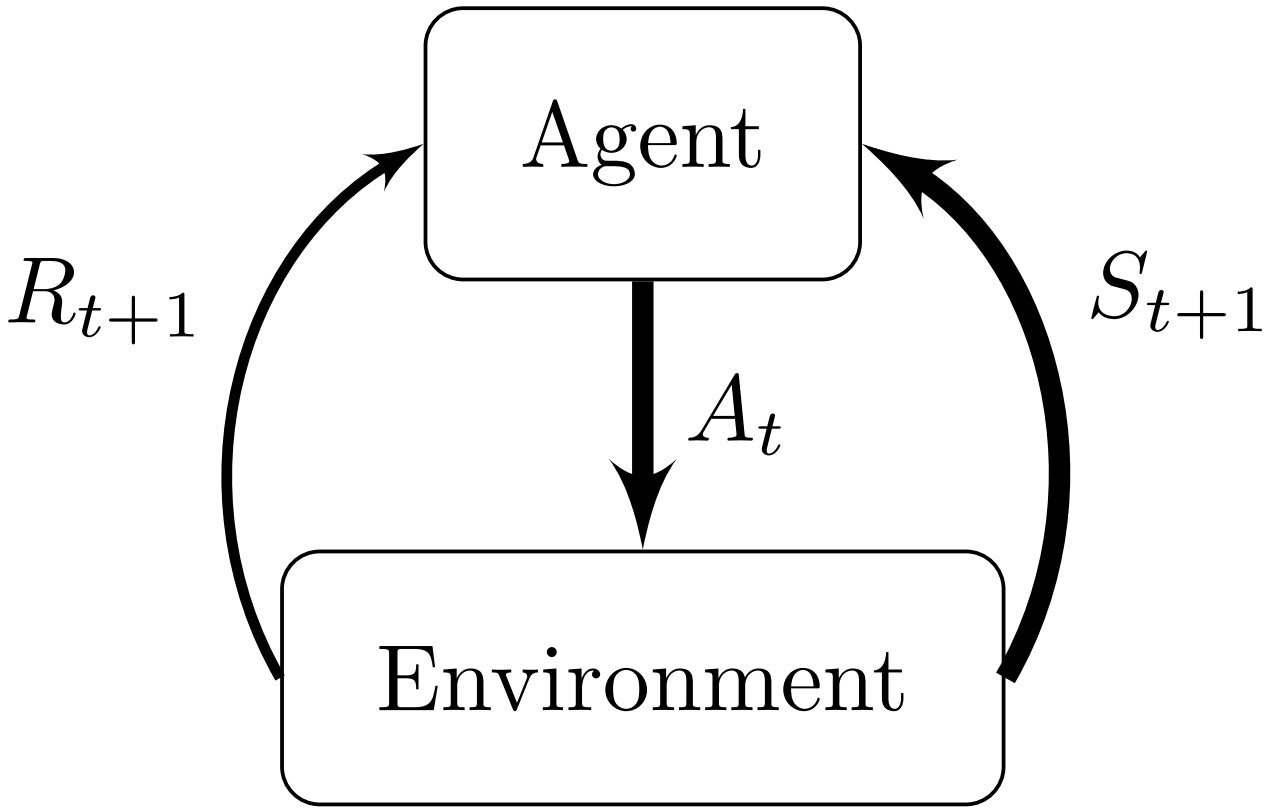
Reinforcement Learning

Finite Markov Decision Processes

Prof. James Brusey

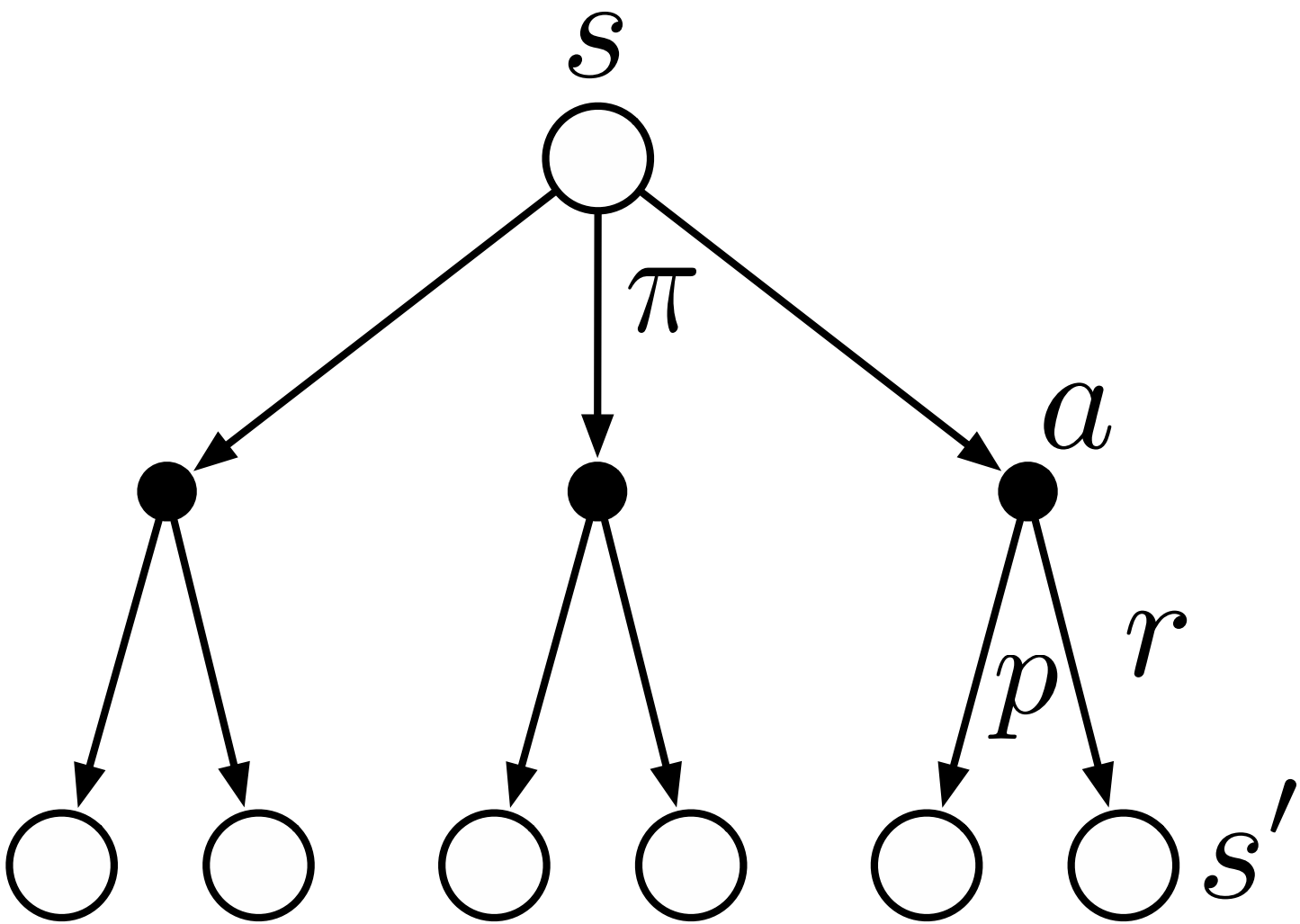
June 4, 2023

Agent-Environment Interface



$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$

Backup diagram



Some terms

- Each MDP comprises the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$
- $\mathcal{S}$ is the set of states
- $\mathcal{A}$ is the set of actions
- $\mathcal{R}$ is the set of rewards
- $\mathcal{P}$ is the transition model
- γ is a discount factor

What is a state?

- You might represent a state with a single number or a vector of numbers
- There are a finite number and so they can be enumerated
- State is a *compact* representation of all history
- If you could do better knowing the history, then the state does not have the *Markov property*
- We will later come to infinite or continuous states

What is an action?

- There are finite actions (infinite or continuous actions later)
- At each time step, some action must be taken but that can be a no-op
- The effect of the action is determined by the transition model

What is a transition model?

- A transition model is a (stochastic) mapping between states, actions, subsequent states, and rewards,

$$p(s', r|s, a)$$

- It represents how the environment "works"

The reward hypothesis

That all of what we mean by goals and purposes can be well thought of as a maximisation of the expected value of the cumulative sum of a received scalar signal (called reward). —Michael Littman (S&B)

Long term reward and Episodes

We aim to maximise our expected reward G_t, which is the sum of all future rewards,

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

ending in the final reward at time T.
An episode is everything up to a final time step T.
Note that if T were infinite, we would have the potential for infinite long-term reward.

Discounted reward

It is often natural to think of a gain in some distant future as being not so valuable as a gain right now.

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

for $\gamma < 1$.

$$G_t \doteq \sum_{0 \leq k < \infty} \gamma^k R_{t+k+1}$$

Note that it is now not necessary to place a finite limit on the episode length.

Unified notation for episodic and continuing tasks

- We can deal with the difference between episodic and continuing tasks using the concept of *absorbing* states
- Absorbing states yield zero reward and always transition to the same state

$$G_t \doteq \sum_{t+1 \leq k \leq T} \gamma^{k-t-1} R_k$$

where $T = \infty$ or $\gamma = 1$ (but not both).

Policies

- A policy represents how to act in each possible state
- Policies are a distribution over actions

$$\pi(a|s) \rightarrow [0, 1]$$

For all $s \in \mathcal{S}$,

$$\sum_{a} \pi(a|s) = 1$$

Value functions

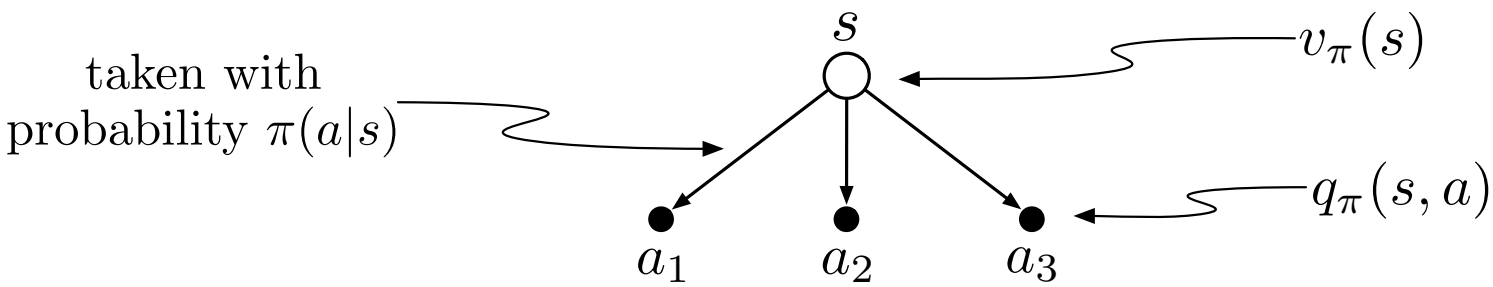
- A state value function $v_\pi(s, a)$ is the long term value of being in state s assuming that you follow policy π

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s]$$

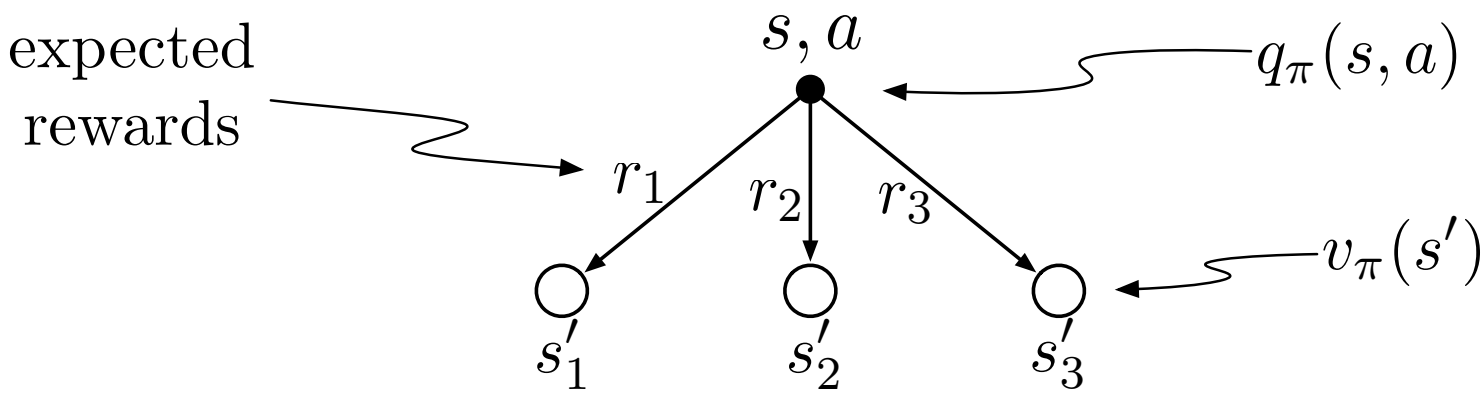
- A state action value function $q_\pi(s, a)$ is the long term value of being in state s, taking action a, and then following π from then on.

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

How do we select an action?



What is the consequence of taking that action?



Optimal policies and value functions

- Given some MDP, what is the best value we can achieve?

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

  for all $s \in \mathcal{S}$

- What is the best state action value achievable?

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

  for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$

Exercise

- Develop a recursive expression for $v_*(s)$ and $q_*(s, a)$ from what we know so far
- Feel free to look at the book for help