Introduction to RL

Reference

- FML Chapter 14
- Sutton and Barto - Reinforcement Learning

Outline

1. Introduction
2. Bellman Equations
3. Temporal Difference (TD) Methods
4. Function Approximation for Value Functions
5. Actor-critic Methods
6. Deep Reinforcement Learning

Outline

1. Introduction
2. Bellman Equations
3. Temporal Difference (TD) Methods
4. Function Approximation for Value Functions
5. Actor-critic Methods
6. Deep Reinforcement Learning

Difference Learning Frameworks

- Supervised:
  - learn from a training set of labelled examples
- Unsupervised:
  - find hidden structure in data, estimate density function
- Reinforcement:
  - learn from iterations, not from examples
  - goal is to maximize accumulated rewards, not to find hidden structure

Learning from Interactions

- Learn what to do: learn actions to maximize accumulated numerical reward
- The agent is not told what to do, but it must discover the best behavior
- The actions that it takes affect future outcome

Learning from Interactions In Practice

- Gives an approximation to a true solution
- Real problems might be continuous or high dimensional

Exploration and Exploitation Dilemma

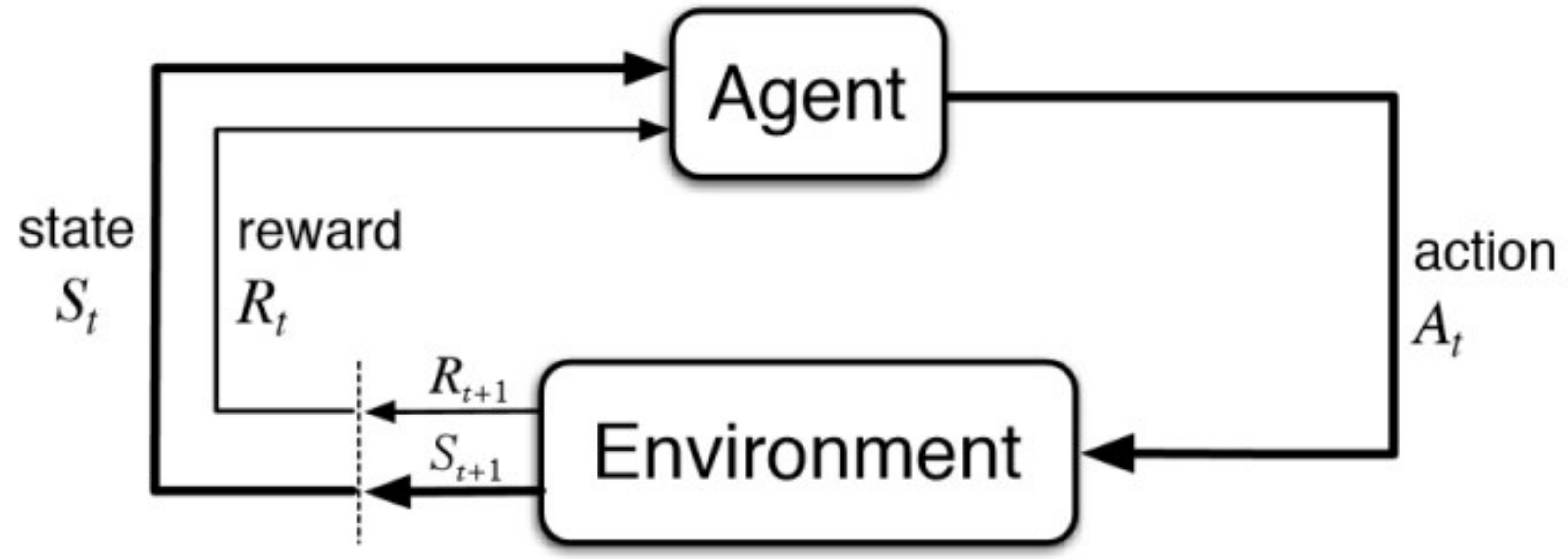
- In RL, a goal-seeking agent must simultaneously
  - exploit current knowledge
  - explore new actions

Abstraction

- RL offers an abstraction to the problem of goal-directed learning from iteration.
- It proposes that the sensory, memory and control apparatus and the objective can be reduced to states, actions and rewards passing back and forth between the agent and the environment.

The agent-environment Interface

RL - abstraction

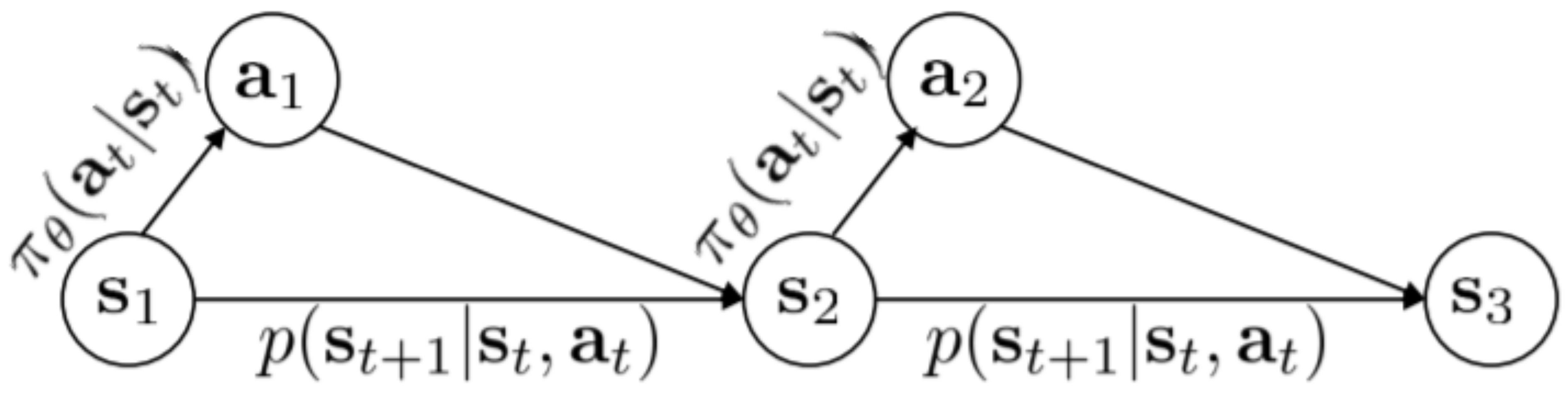


- State space $\mathsf{S} = \{s^1, \ldots, s^{|\mathsf{S}|}\}$
- Action space $\mathsf{A} = \{a^1, \ldots, a^{|\mathsf{A}|}\}$
- Reward space $\mathbb{R}$
- History $h_t = \{s_0, a_0, r_1, \ldots, s_{t-1}, a_{t-1}, r_t, s_t, a_t\}$

Transition model $\Pr(s_{t+1} = s', r_{t+1} = r \mid h_t)$

- Markov Property: s_{t+1} only depends on s_t and a_t

$$\Pr(s_{t+1} = s', r_{t+1} = r \mid h_t) \overset{\text{markov}}{=} \Pr(s_{t+1} = s', r_{t+1} = r \mid s_t, a_t)$$



- Expected reward of taking action a at a state s

$$\mathbb{E}[r_{t+1} \mid s_t = s, a_t = a] = \sum_{r,s'} r \ \Pr(s_{t+1} = s', r_{t+1} = r \mid s_t = s, a_t = a) := \sum_{r,s'} r \ p(s', r \mid s, a)$$

1. state transition probability $p(s' \mid s, a)$
2. expected reward $r(s, a, s') = \mathbb{E}[r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s']$

$$r(s, a, s') = \sum_r r p(r \mid s, a, s') = \sum_r r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Value Functions

- Policy $\Pr(a_{t+1} | s_{t+1}) = \Pr(a_t | s_t) = \pi(a_t | s_t)$
- Return or Accumulated future reward $R_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$
- State-value function for policy π

$$V^\pi(s) = \mathbb{E}_\pi[R_t | s_t = s]$$

- State-action-value function for policy π

$$Q^\pi(s, a) = \mathbb{E}_\pi[R_t | s_t = s, a_t = a]$$

$$\boxed{V^\pi(s) = \sum_a \pi(a | s) Q^\pi(s, a)}$$