Informatics 2D: Reasoning and Agents

Alex Lascarides

School of informatics

The University of Edinburgh

Lecture 30a: Markov Decision Processes Representation

Where are we?

Last time ...

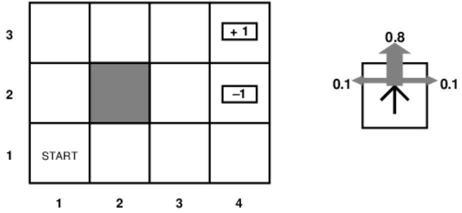
- Talked about decision making under uncertainty
- Looked at utility theory
- Discussed axioms of utility theory
- Described different utility functions
- Introduced decision networks

Today ...

- **Markov Decision Processes**

Sequential decision problems

- So far we have only looked at one-shot decisions, but decision process are often sequential
- Example scenario: a 4x3-grid in which agent moves around (fully observable) and obtains utility of +1 or -1 in terminal states



(a) (b)

- Actions are somewhat unreliable (in deterministic world, solution would be trivial)

Markov decision processes

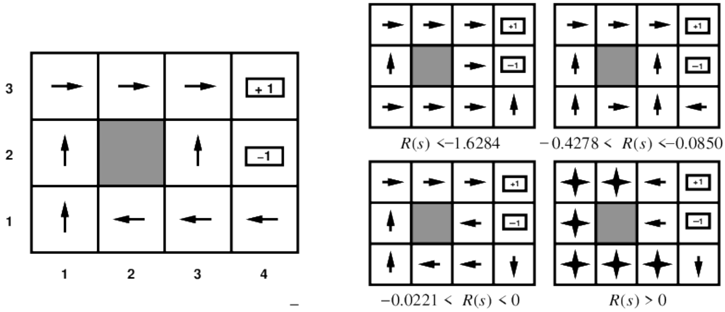
- To describe such worlds, we can use a **(transition) model** $T(s, a, s')$ denoting the probability that action a in s will lead to state s'
- Model is Markovian: probability of reaching s' depends only on s and not on history of earlier states
- Think of T as big three-dimensional table (actually a DBN)
- Utility function now depends on **environment history**
  - agent receives a reward $R(s)$ in each state s (e.g. -0.04 apart from terminal states in our example)
  - (for now) utility of environment history is the sum of state rewards
- In a sense, stochastic generalisation of search algorithms!

Markov decision processes

- Definition of a **Markov Decision Process (MDP)**:
  Initial state: S_0
  Transition model: $T(s, a, s')$
  Utility function: $R(s)$
- Solution should describe what agent does in every state
- This is called **policy**, written as π
- $\pi(s)$ for an individual state describes which action should be taken in s
- **Optimal policy** is one that yields the highest expected utility (denoted by π^*)

Example

- Optimal policies in the 4x3-grid environment
  (a) With cost of -0.04 per intermediate state π^* is conservative for (3,1)
  (b) Different cost induces direct run to terminal state/shortcut at (3,1)/no risk/avoid both exits



(a) (b)

Summary

- Sequential decision making
- Defined Markov Decision Processes
- Defined policies, and optimal policies
- Next time: **Computing optimal policies from MDPs**