Policies

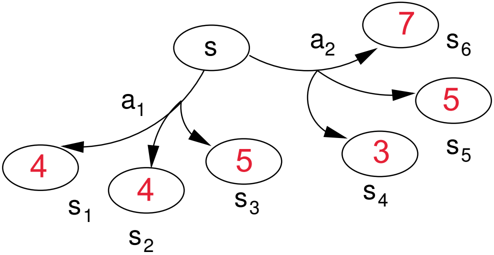
Frederik Mallmann-Trenn
6CCS3AIN

Other notions of "rational"

- There are other criteria for decision-making than maximising expected utility.
- One approach is to look at the option which has the least-bad worst outcome.
- This maximin criterion can be formalised in the same framework as MEU (Maximum Expected Utility), making the rational (in this sense) action:

$$a^* = \arg\max_{a \in A}\{\min_{s' \in s_a} u(s')\}$$

- Its effect is to ignore the probability of outcomes and concentrate on optimising the worst case outcome.
- Example (utilities are in red):



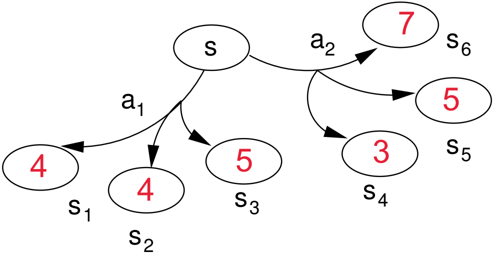
- Here we would pick a_1

Other notions of “rational”

- The opposite attitude, that of optimistic risk-seeker, is captured by the maximax criterion:

$$a^* = \arg\max_{a \in A}\{\max_{s' \in s_a} u(s')\}$$

- This will ignore possible bad outcomes and just focus on the best outcome of each action.
- Example:



- Here we would pick a_2

Sequential decision problems

- These approaches give us a battery of techniques to apply to individual decisions by agents.
- However, they aren't really sufficient.
- Agents aren't usually in the business of taking single decisions
  - Life is a series of decisions.

  The best overall result is not necessarily obtained by a greedy approach to a series of decisions.
- The current best option isn't the best thing in the long-run.

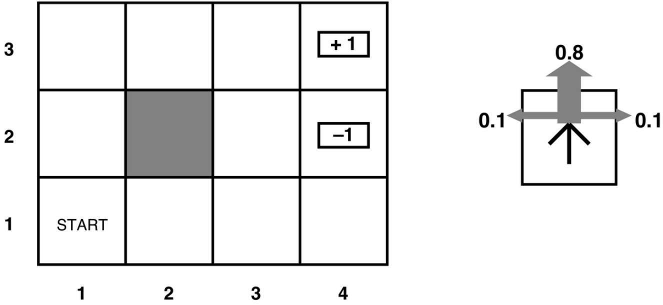
Sequential decision problems

- Otherwise I'd only ever eat cherry pie

(pillsbury.com)

(Damn fine pie.)

An example



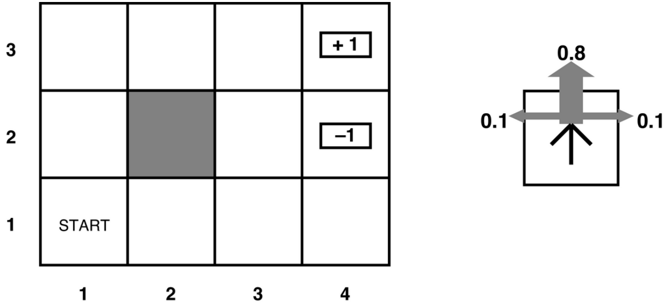
(a) (b)

- The agent has to pick a sequence of actions.

$$A(s) = \{Up, Down, Left, Right\}$$

for all states s.

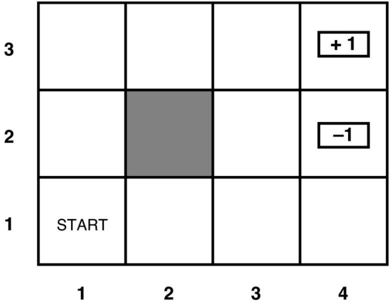
An example



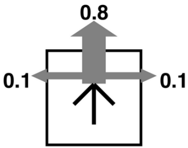
(a) (b)

- The world is fully observable.
- End states have values $+1$ or -1.

An example



(a)



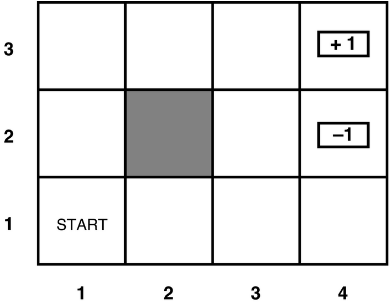
(b)

- If the world were deterministic, the choice of actions would be easy here.

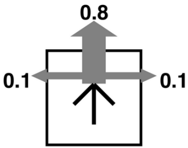
$$Up, Up, Right, Right, Right$$

- But actions are stochastic.

An example



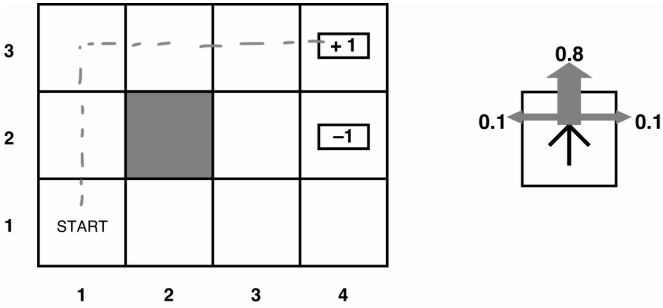
(a)



(b)

- 80% of the time the agent moves as intended.
- 20% of the time the agent moves perpendicular to the intended direction. Half the time to the left, half the time to the right.
- The agent doesn't move if it hits a wall.

An example

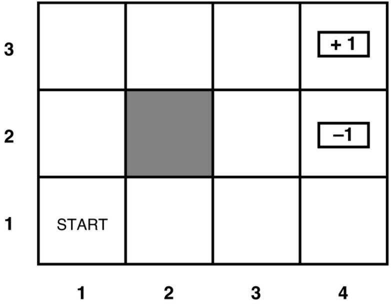


(a) (b)

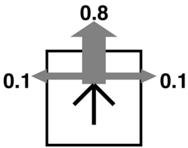
- So $Up, Up, Right, Right, Right$ succeeds with probability:

$$0.8^5 = 0.32768$$

An example



(a)



(b)

- Also a small chance of going around the obstacle the other way.

An example

- We can write a transition model to describe these actions.
- Since the actions are stochastic, the model looks like:

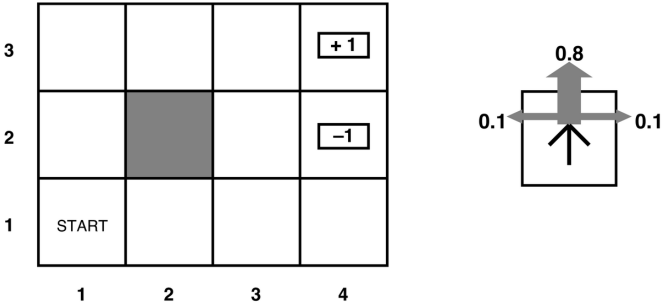
$$P(s'|s, a)$$

  where a is the action that takes the agent from s to s'.
- Transitions are assumed to be first order Markovian.
- That is, they only depend on the current and next states.
- So, we could write a large set of probability tables that would describe all the possible actions executed in all the possible states.
  This would completely specify the actions.

An example

- The full description of the problem also has to include the utility function.
- This is defined over sequences of states — runs in the terminology of the first lecture.
- We will assume that in each state s the agent receives a reward $R(s)$.
- This may be positive or negative.

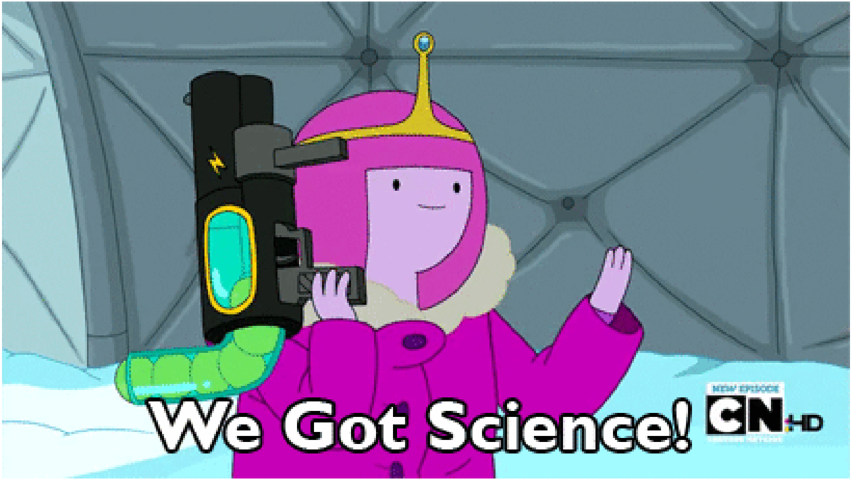
An example



(a) (b)

- The reward for non-terminal states is -0.04.
- We will assume that the utility of a run is the sum of the utilities of states, so the -0.04 is an incentive to take fewer steps to get to the terminal state. (You can also think of it as the cost of an action).

How do we tackle this?



(Pendleton Ward/Cartoon Network)

Markov decision process

- The overall problem the agent faces here is a Markov decision process (MDP)
- Mathematically we have
  - a set of states $s \in S$ with an initial state s_0.
  - A set of actions $A(s)$ in each state.
  - A transition model $P(s'|s, a)$; and
  - A reward function $R(s)$.
- Captures any fully observable non-deterministic environment with a Markovian transition model and additive rewards.

Leslie Pack Kaelbling

Markov decision process

- What does a solution to an MDP look like?

Markov decision process

- A solution is a policy, which we write as π.
- This is a choice of action for every state.
  - that way if we get off track, we still know what to do.
- In any state s, $\pi(s)$ identifies what action to take.
- Example policy π: $\pi(s_0) = left, \pi(s_1) = left, \pi(s_2) = right, \ldots$
- Another example policy π': For all states s, $\pi'(s) = left$

Markov decision process

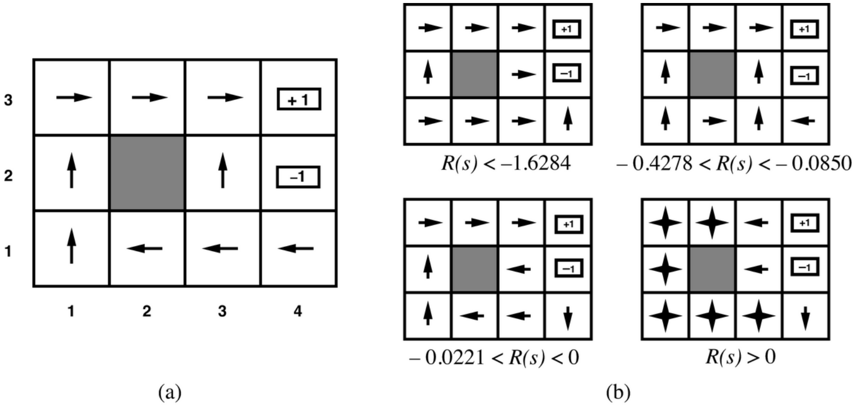
- Naturally we'd prefer not just any policy but the optimum policy.
  - But how to find it?
- Need to compare policies by the reward they generate.
- Since actions are stochastic, policies won't give the same reward every time.
  - So compare the expected value.
- The optimum policy π^* is the policy with the highest expected value.
- At every stage the agent should perform $\pi^*(s)$.

Markov decision process

(40 Acres and a Mule Filmworks/Universal Pictures)

- $\pi^*(s)$ is the right thing.

An example



(a) An optimal policy for the stochastic environment with $R(s) = -0.04$.

(b) Optimal policies for different values of $R(s)$.

An example

- $R(s) \leqslant -1.6284$, life is painful so the agent heads for the exit, even if it is a bad state.
- $-0.4278 \leqslant R(s) \leqslant -0.0850$, life is unpleasant so the agent heads for the $+1$ state and is prepared to risk falling into the -1 state.
- $-0.0221 < R(s) < 0$, life isn't so bad, and the optimal policy doesn't take any risks.
- $R(s) > 0$, the agent doesn't want to leave.