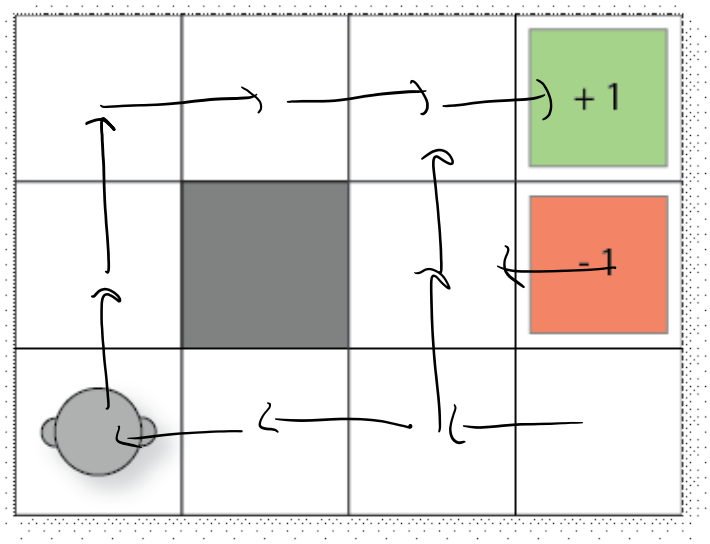
Markov Decision Processes

Tuesday, 4 September 2018 10:29 AM



The grey square is a wall.

The two labelled cells give a reward: 1 and -1 respectively. (Actually, we will assume V(s)=1 or -1)

But! Things can go wrong:

- If the agent tries to move north, 80% of the time, this works as planned (provided the wall is not in the way)
- 10% of the time, trying to move north takes the agent west (provided the wall is not in the way);
- 10% of the time, trying to move north takes the agent east (provided the wall is not in the way)
- If the wall is in the way of the cell that would have been taken, the agent stays put.
- Similar for all other directions

Classical Planning:

- Static environment
- Perfect knowledge
- Single actor
- ~~Deterministic action.~~

MDPs:

- State space S
- Initial state s0 in S
- Actions A --- A(s)
- Transition probabilities:
  - 0<= P_a(s'|s) <=1
- Rewards: r(s, a, s') numeric reward
- Discount factor 0 < gamma <=1

Discounted reward:

Probabilistic PDDL:

```
(define (domain bomb-and-toilet)
    (:requirements :conditional-effects :probabilistic-effects)
    (:predicates (bomb-in-package ?pkg) (toilet-clogged)
                 (bomb-defused))

    (:action dunk-package
        :parameters (?pkg)
        :effect (and (when (bomb-in-package ?pkg)
                     (bomb-defused))
                (probabilistic 0.05 (toilet-clogged))))
```

Solution for MDP is a *policy*:

```
at(0,0) => move_right
at(0,1) => move_right
at(0,2) => move_right
at(0,3) => stay
at(1,0) => move_up
at(1,2) => move_up
at(1,3) => move_up
at(2,0) => move_up
at(2,1) => move_left
at(2,2) => move_up
at(2,3) => move_left
```

Expected return exercise:

You can steal:

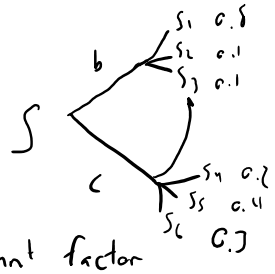
A) An iPhone, which you think you have a 20% chance of selling for $500, or an 80% chance of selling for $250.

B) A Samsung, which you think you have a 50% chance of selling for $500, or a 50% chance of

selling for $200.
Which do you steal?

A = 0.2*500 + 0.8*250 = 300
B = 0.5*500 + 0.5*200 = 350

Bellman equations:



$$V(s) = \max_{a \in A} \sum_{s' \in S} P_a(s'|s)\left[r(s,a,s') + \gamma \cdot V(s')\right]$$

discount factor (gamma)

value function

b, c