# KI2 - 12

# Partially Observable Markov Decision Processes POMDP

***Marius Bulacu***

Kunstmatige Intelligentie / RuG



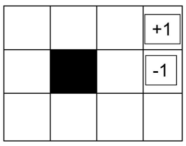
## Motivation

- MDP
  - Stochastic environment
  - States are fully observable
- Partially Observable MDP
  - Like MDP
    but…
  - Missing knowledge of the state where the agent is in:
    - May not be complete
    - May not be correct



## Example

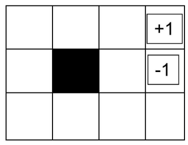
- Agent doesn't know where it is
  - Has no sensors
  - Being put in an unknown square
- What should the agent do?
  - If it knew that it is in (3,3) it would do *Right* action
  - But it doesn't know






## Solutions

- First solution:
  - Step 1: reduces uncertainty
  - Step 2: try heading to the +1 exit
- Reduce uncertainty
  - move 5 times *Left* so it is quite likely to be at the left wall
  - Then move 5 times *Up* so it is quite likely to be at left top wall
  - Then move 5 time *Right* to goal state
  - Continue moving right to increase chance to get to +1
- Quality of solution
  - Chance of 81.8% to get to +1
  - Expected utility of about 0.08
- Second solution: POMDP






## MDP

- Components:
  - States - S
  - Actions - A
  - Transitions - T(s, a, s')
  - Rewards – R(s)
- Problem:
  - choose the action that makes the right tradeoffs between the immediate rewards and the future gains, to yield the best possible solution
- Solution:
  - Policy



## PO-MDP

- Components:
  - States - S
  - Actions - A
  - Transitions – T(s,a,s')
  - Rewards – R(s)
  - Observations – O(s,o)



## Policy Mapping: MDP vs POMDP

- In MDP, mapping is from states to actions.
- In POMDP, mapping is from probability distributions (over states) to actions.



## Observation Model

- O(s, o) – probability of getting observation o in state s
- Example:
  - for robot without sensors:
    - O(s, o) = 1 for every s in S
    - Only one observation o exists

## Belief State

- The agent may be in any state
- b(s) = probability distribution over all states
- Example:
  – in the 4x3 example for robot without sensors the initial state:

b=<1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 0, 0>



## Belief State (contd.)

- b(s) = probability assigned to actual state s by belief b
  – $0 < b(s) < 1$ for every $s$ in $S$
  – sum of $b(s)$ over all states = 1
- We will define b' = FORWARD(b,a,o)

$$b'(s') = \alpha O(s', o) \sum_s T(s, a, s') b(s)$$

α is normalizing constant that makes the belief state sum to 1



## Finding the Best Action

- The optimal action depend only on the current belief state
- Choosing the action:
  – Given belief state b, execute π*(b)
  – Received observation o
  – Calculate new state belief b′=FORWARD(b,a,o)
  – Set b← b′



## π*(b) vs π*(s)

- POMDP belief state is continuous
- In the 4x3 example, b is in an 11-dimensional continuous space
  – b(s) is a point on a line between 0 to 1
  – b is point in n dimensional space



## Transition Model & Rewards

- Need different view of transition model and reward
- Transition model as τ(b, a, b′)
  – Instead T(s, a, s′)
  – Actual state s is not known
- Reward as R(b)
  – Instead R(s)
  – Have to model the uncertainty in the belief state



## Transition Model

- Probability of getting the observation o

$$P(o \mid a,b) = \sum_{s'} P(o \mid a, s', b) P(s' \mid a, b)$$

$$= \sum_{s'} O(s', o) P(s' \mid a, b) = \sum_{s'} O(s', o) \sum_s T(s, a, s') b(s)$$

- Transition model

$$t(b, a, b') = P(b' \mid a, b) = \sum_o P(b' \mid o, a, b) P(o \mid a, b)$$

$$= \sum_o P(b' \mid o, a, b) \sum_{s'} O(s', o) \sum_s T(s, a, s') b(s)$$

*where* $P(b' \mid o, a, b)$ *is* 1 *if* $b' = FORWARD(b, a, o)$

*and* 0 *otherwise*



## Reward Function

- Expected reward of all the states the agent might be in

$$R(b) = \sum_s b(s) R(s)$$



## From POMDP to MDP

- τ(b, a, b′) and R(b) define an MDP
  – π*(b) for this MDP is optimal policy for original POMDP
  – Belief state is observable to the agent
- Need new versions of Value / Policy Iteration
  – for the continuous belief state

# Back to the Example

- POMDP solution for the 4x3 environment:
  [*Left, Up, Up, Right, Up, Up, Right, Up, Up, Right, Up, Right, Up, Right, Up ...*]
- The policy is a sequence since the problem is deterministic in beliefs space
- The agent gets to the goal 86.6% of times
  - Expected utility is 0.38



# Overview of Markov Models

| Markov Models | | Do we have control over the state transitons? NO | Do we have control over the state transitons? YES |
|---|---|---|---|
| Are the states completely observable? | YES | **Markov Chain** | **MDP** Markov Decision Process |
| | NO | **HMM** Hidden Markov Model | **POMDP** Partially Observable Markov Decision Process |