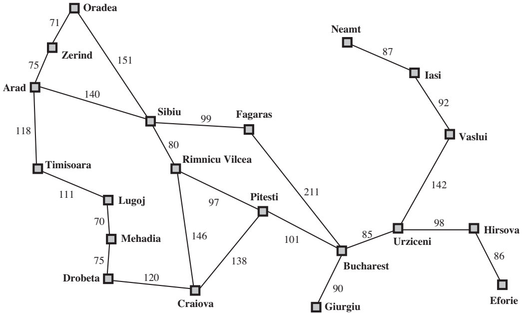
Partially Observable MDP

3007/7059 Artificial Intelligence

School of Computer Science
The University of Adelaide

Romania again...



Previously we introduced **stochastic transition models** to account for the fact that actions may not lead to desired outcomes.

Realistically, the environment is only **partially observable** — we are not 100% sure in which city/town we currently are, since we've never been to those places, and we may misread traffic signs.

Partially Observable MDP

Compared to MDP, another source of uncertainty in **Partially Observable MDP (POMDP)** is the **current state of the agent**.

If we don't know what the current state s is, we can't simply calculate the optimal action using $\pi^*(s)$.

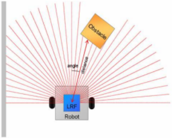
The right action to take depends not only on s, but how much the agent knows that it is in s.

POMDPs are much more difficult to solve than MDPs — we cannot avoid solving POMDPs, since the real world is one.

Example:
A sophisticated positioning sensor like GPS can only pin-point the coordinate up to 1 to 10 meters error — **no sensors are perfect!**

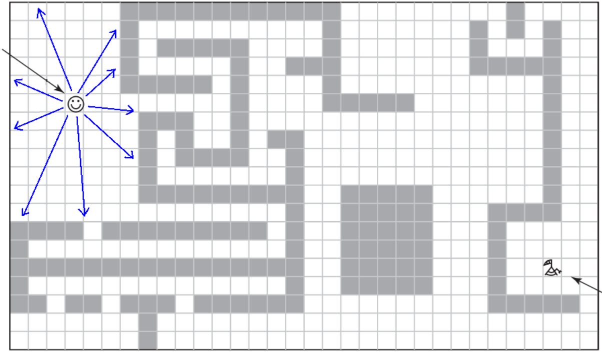
Robotic path planning

Sensors report the current position. Sensors are imperfect --> model the error using a **sensor model**



Wheels not totally reliable --> stochastic transition models

Robot

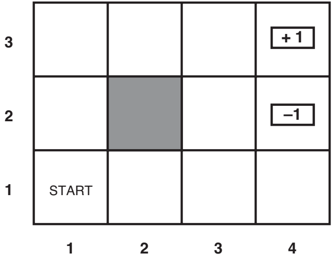


Goal

Sensor model

Apart from a transition model $P(s'|s,a)$ and a reward function $R(s)$, a POMDP has a **sensor model** $P(e|s)$, which gives the probability of perceiving **evidence** or **measurement** e in state s.

Example:
Using the 4×3 world again, a sensor might measure the **number of adjacent walls** at the current (unknown) position. The noisy sensor gives the wrong value with probability 0.1.



Note that sensors in general may provide only **indirect measurements** of the current state.

Belief states

To accommodate our uncertainty about the current **physical state**, we maintain a probability distribution over the states we can possibly be in. We call this distribution a **belief state** b.

$b(s)$ gives the probability that we are currently in physical state s.

Note that $b(s)$ is a valid probability distribution

$$\sum_s b(s) = 1$$

Example:
The initial belief state in the 4×3 world is

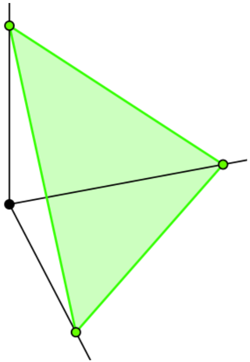
$$b = \left\langle \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, 0, 0 \right\rangle$$

i.e., we are equally probable to be in one of the non-terminal states.

Belief states (cont.)

A belief state b over N physical states is a point on an $(N-1)$-dimensional simplex in N-dimensional space.

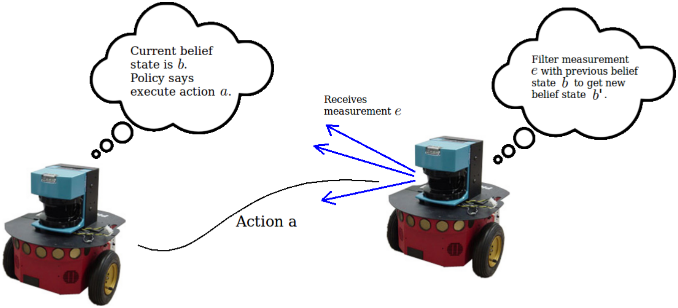
Example:
A belief state b over 3 physical states is a point on the 2D simplex.



Filtering

At each time instance, given the current belief state b, we perform action a (as suggested by a policy) and receive evidence e.

This information can be combined with our previous belief state b to obtain a new belief state b'. The task is called **filtering**.



Example:
We were equally likely to be in any non-terminal state in the 4×3 map. We performed $a = MoveLeft$, and received measurement of 1 adjacent wall. Where are we likely to be in now?

Filtering (cont.)

The new belief state b' is calculated as

$$b'(s) = \alpha P(e|s') \sum_{s} P(s'|s,a)b(s)$$

where α is simply a normalisation constant. Intuititively,

- $b(s)$ — the probability that we could have been in s.
- $P(s'|s,a)$ — the probability that we have moved to s' if we were in s.
- $P(e|s')$ — the probability of seeing measurement e if we are now in s'.

Henceforth we summarise this equation as

$$b' = Forward(b,a,e)$$

POMDP sequential decision making

The fundamental insight is this:
The optimal action depends only on the agent's current belief state, i.e., the optimal policy $\pi^*(b)$ *maps belief state* b *to an action. It does not depend on the actual physical state the agent is in.*

Thus sequential decision making becomes

1. Given the current belief state b, execute action $a = \pi^*(b)$.
2. Receive percept or measurement e.
3. Update the current belief state as $Forward(b, a, e)$.
4. Repeat from Step 1.

With this strategy, an action a not only changes the true underlying physical state (which is unobserved), but also the belief state.

Transition model for belief states

Like MDPs, POMDPs have a transition model for physical states $P(s'|s,a)$.

But we also want to derive a **transition model for belief states** $P(b'|b,a)$.

If we knew the action a and the resulting evidence e, this is just the **deterministic update** $Forward(b,a,e)$.

If we don't have e yet, all we can say is that we may arrive at several possible belief states b', some more likely that others, by virture of knowing b and a.

The probability of being in one of them after executing a is just $P(b'|b,a)$.

Transition model for belief states (cont.)

Before we can get $P(b'|b, a)$, we need to compute

$$
\begin{aligned}
P(e|b,a) &= \sum_{s'} P(e|b,a,s')P(s'|b,a) && \text{(Product rule and marginalisation)} \\
&= \sum_{s'} P(e|s')P(s'|b,a) && \text{(Cond. ind. between } e \text{ and } b,a \text{ given } s'\text{)} \\
&= \sum_{s'} P(e|s') \sum_{s} P(s'|s,a)b(s) && \text{(Product rule and marginalisation)}
\end{aligned}
$$

Transition model for belief states (cont.)

Before we can get $P(b'|b,a)$, we need to compute

$$
\begin{aligned}
P(e|b,a) &= \sum_{s'} P(e|b,a,s')P(s'|b,a) && \text{(Product rule and marginalisation)} \\
&= \sum_{s'} P(e|s')P(s'|b,a) && \text{(Cond. ind. between } e \text{ and } b,a \text{ given } s'\text{)} \\
&= \sum_{s'} P(e|s') \underbrace{\sum_{s} P(s'|s,a)b(s)}_{} && \text{(Product rule and marginalisation)}
\end{aligned}
$$

Where we could have been and where we could have gone after performing a.

Transition model for belief states (cont.)

Before we can get $P(b'|b,a)$, we need to compute

$$\begin{aligned}
P(e|b,a) &= \sum_{s'} P(e|b,a,s')P(s'|b,a) && \text{(Product rule and marginalisation)} \\
&= \sum_{s'} P(e|s')P(s'|b,a) && \text{(Cond. ind. between } e \text{ and } b,a \text{ given } s'\text{)} \\
&= \underbrace{\sum_{s'} P(e|s')}\sum_{s} P(s'|s,a)b(s) && \text{(Product rule and marginalisation)}
\end{aligned}$$

What's the probability of seeing e in the places we could have landed in.

Transition model for belief states (cont.)

The transition model is then

$$P(b'|b,a) = \sum_e P(b'|e,a,b)P(e|a,b) \qquad \text{(Product rule and marginalisation)}$$
$$= \sum_e P(b'|e,a,b) \sum_{s'} P(e|s') \sum_s P(s'|s,a)b(s)$$

where $P(b'|e,a,b)$ is 1 if $b' = Forward(b,a,e)$ and 0 otherwise.

Value iteration for POMDP

Completing the list of things we need is a **reward function for belief states**

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

where $R(s)$ is the reward function for physical states.

Together, $P(b'|b,a)$ and $\rho(b)$ define an **observable MDP over belief states**. The optimal policy for this MDP is also the optimal policy for the original POMDP.

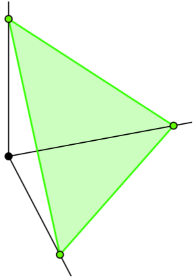
So we can just run Value Iteration using $P(b'|b,a)$ and $\rho(b)$ to get the optimal policy...

Value iteration for POMDP (cont.)

... except that we cannot really do that feasibly.

In Value Iteration for MDPs, we have a list of accumulators (one for each physical state) which are iteratively updated.

It might appear that we can do the same thing — have one accumulator for each belief state — but there are an **infinite number of belief states**! Recall the simplex:



Approximate Value Iteration for POMDP is required — this is beyond the scope of this course. The underlying concept is similar.