Example
POMDPs

More formally, a POMDP is:
- $S$, a set of states
- $A$, a set of actions
- $T$, transition function
- $R$, reward function
- $\gamma$, discount factor
- $\Omega$, set of observations
- $O$, observation function $O(\omega_t|s_t)$
POMDPs

$s$ changes

$a$

$O(s) \ r$
Belief State

Probability distribution over states $b(s)$:

- Estimate state using observations
- Update based on observations
- Distribution represents state uncertainty
- Take action based on distribution

$$b(s_t) = P(s_t | o_t, o_{t-1}, a_{t-1}, \ldots, o_0, a_0)$$

Use a filter to update $b(s)$ at every time step.
Belief State

Let's examine how we update $b(s)$:

$$b(s_{t+1}) \propto O(o_{t+1} | s_{t+1}) \sum_{s \in S} T(s_{t+1} | s, a_t) b(s)$$

... and also, what is the expected reward when taking an action?

$$\hat{r}(a_t) = \sum_{s \in S} b(s = s_t) \sum_{s' \in S} b(s' = s_{t+1} | s = s_t, a_t) R(s, a_t, s')$$

Notice anything?
The Belief MDP

Belief state updates are a *transition function*.

Expected rewards are a *reward function*.

Both of these, for time $t+1$, depend only on $b(s_t)$.

Therefore they satisfy the Markov property.

We can build an MDP, called the belief MDP,

- Markov even though underlying POMDP not
- Solving this as MDP solves the POMDP problem
Example

**Reward Function**
- Penalty for wrong opening: -100
- Reward for correct opening: +10
- Cost for listening action: -1

**Observations**
- to hear the tiger on the left (TL)
- to hear the tiger on the right (TR)

---

Actions = {0: listen, 1: open-left, 2: open-right}

S0
“tiger-left"
Pr(o=TL | S0, listen)=0.85
Pr(o=TR | S1, listen)=0.15

S1
“tiger-right"
Pr(o=TL | S0, listen)=0.15
Pr(o=TR | S1, listen)=0.85

(picture from Timmer, Intro to POMDPs)
More Formally

A belief MDP consists of a tuple \( (B, A, \tau, r, \gamma) \):

- \( B \) is the set of belief states.
- \( A \) is the action set
- \( \tau \) is the belief state transition function
- \( r \) is the belief reward function
- \( \gamma \) is the discount factor

Of these, \( A \) and \( \gamma \) are taken directly from the originating POMDP, and the definition of \( B \) follows from it.

Must define \( \tau \) and \( r \), which requires using stuff from the original POMDP.
Defining Transition and Reward

So:

\[ \tau : P(b'|b, a) = \sum_o P(b'|b, a, o) P(o|b, a) \]

belief state update

observation model

\[ \hat{r}(a_t) = \sum_{s \in S} b(s = s_t) \sum_{s' \in S} b(s' = s_{t+1}|s = s_t, a_t) R(s, a_t, s') \]
Planning in Belief Space

(images courtesy of Rob Platt)
Planning in Belief Space

"dark"

"light"

start

goal
Belief-Space Planning

We can use any MDP solver we like to solve the system.

[Platt et al., RSS 2010]
Belief-Space Planning

[Platt 2011]
Belief-Space Planning

Entropy: 8.98720

Create Plan