Partially observable MDPs

Introduction

- Agent cannot directly observe its current state. E.g: Robot navigation with noisy actuators, medical applications. Need observations for determining the probable current state.

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

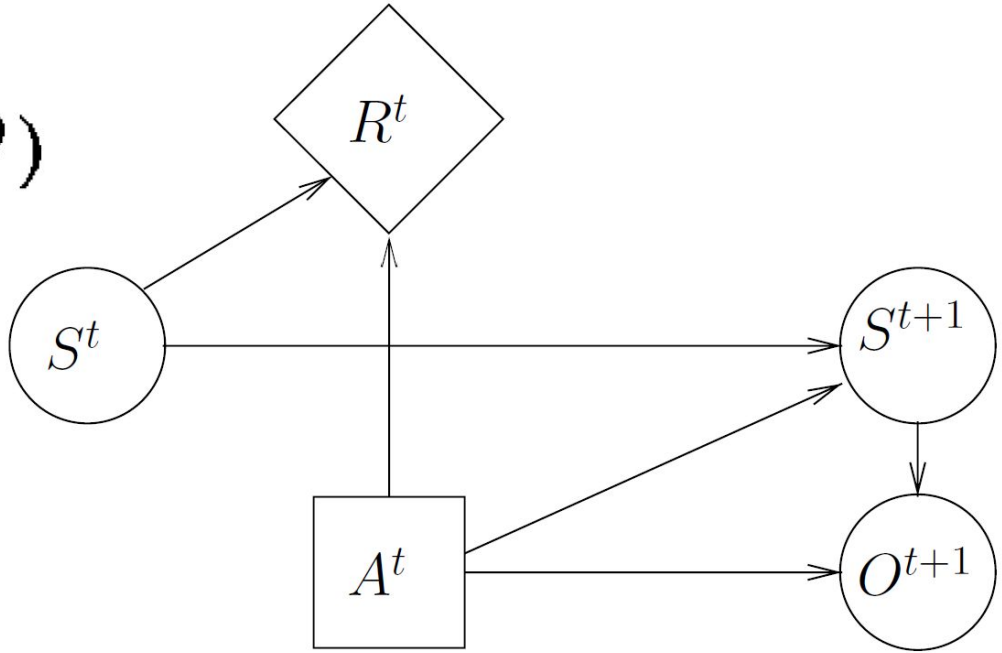
Formal model

$$O : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\Omega)$$

Observation space

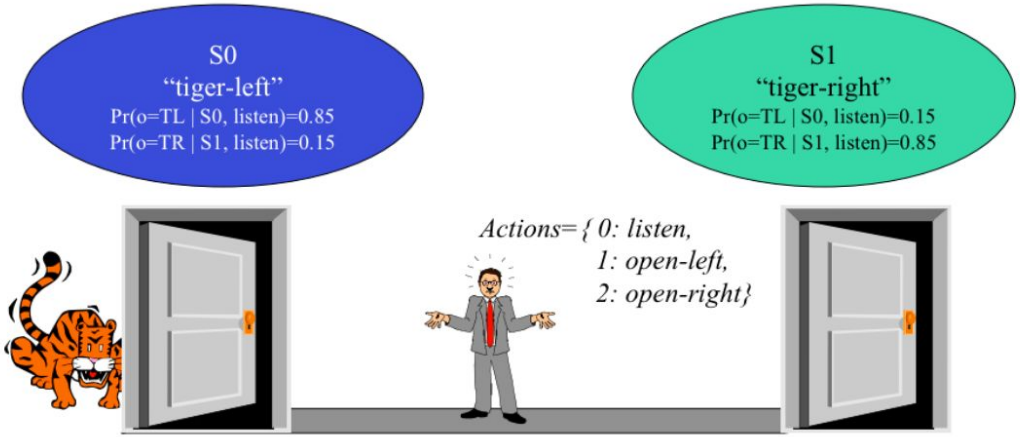
$$O(s', a, o)$$

Pr(Observing o | agent took action a and ended up in state s')



Example

POMDPs: Tiger 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)*

Assumptions for this talk

- T, O are known
- Finite horizon case
- Offline policies

Formal model

- Belief state $b(s)$: Belief of the agent that it is in state s

$$
\begin{aligned}
b'(s') &= \Pr(s' \mid o, a, b) \\
&= \frac{\Pr(o \mid s', a, b) \Pr(s' \mid a, b)}{\Pr(o \mid a, b)} \\
&= \frac{\Pr(o \mid s', a) \sum_{s \in \mathcal{S}} \Pr(s' \mid a, b, s) \Pr(s \mid a, b)}{\Pr(o \mid a, b)} \\
&= \frac{O(s', a, o) \sum_{s \in \mathcal{S}} T(s, a, s') b(s)}{\Pr(o \mid a, b)}.
\end{aligned}
$$

Belief MDP

- $\mathcal{B}$, the set of belief states, comprise the state space;
- $\tau(b, a, b')$ is the state-transition function, which is defined as

$$\tau(b, a, b') = \Pr(b' \mid a, b) = \sum_{o \in \Omega} \Pr(b' \mid a, b, o) \Pr(o \mid a, b),$$

where

$$\Pr(b' \mid b, a, o) = \begin{cases} 1 & \text{if } \mathrm{SE}(b, a, o) = b' \\ 0 & \text{otherwise;} \end{cases}$$

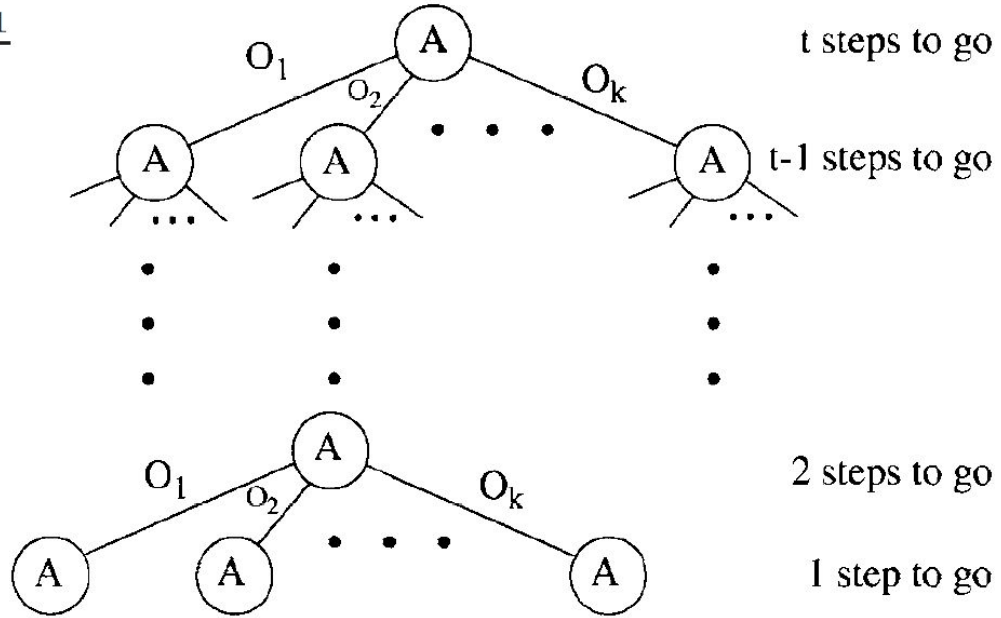
- $\rho(b, a)$ is the reward function $\rho(b, a) = \sum_{s \in \mathcal{S}} b(s) R(s, a)$

Policy trees

- Size of representation for a horizon H policy tree: $|\mathcal{A}|^{\frac{|\mathcal{O}|^H - 1}{|\mathcal{O}| - 1}}$

- Finding optimal policies for finite POMDPs is PSPACE-complete

- Existence of optimal solution to infinite horizon POMDPs is undecidable



Value function

$$V_p(b) = \sum_{s \in S} b(s) V_p(s)$$

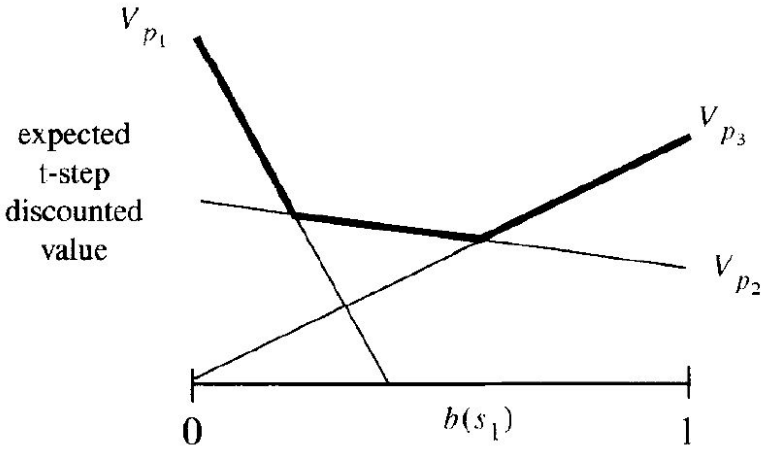
← Value function for state s corresponding to policy tree p

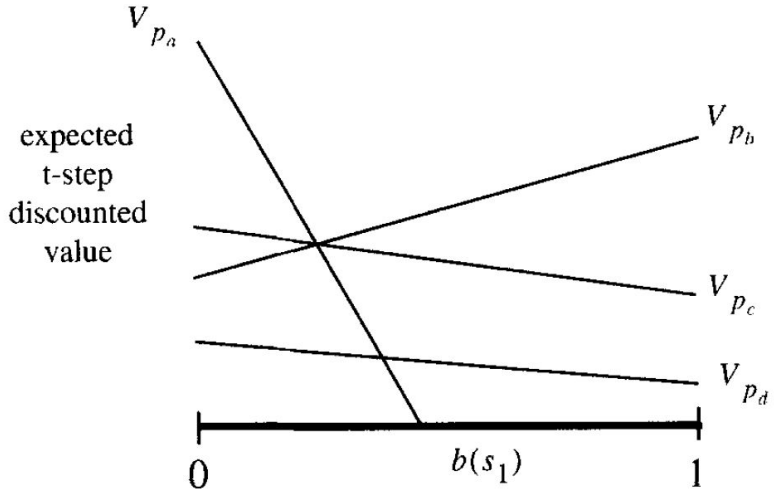
$$\alpha_p = \langle V_p(s_1), \ldots, V_p(s_n) \rangle, \text{ then } V_p(b) = b \cdot \alpha_p$$

$$V_t(b) = \max_{p \in \mathcal{P}} b \cdot \alpha_p$$

Value function

Value function is convex and piecewise linear in b





Some policies are completely dominated by others => Can prune the space of policies

Value iteration

- Initialize $t = 0$ and $V_0(b) = 0$ for all $b \in \mathcal{B}$.
- While $\sup_{b \in \mathcal{B}} |V_{t+1}(b) - V_t(b)| > \epsilon$, calculate $V_{t+1}(b)$ for all states $b \in \mathcal{B}$ according to the following equation, and then increment t:

$$V_{t+1}(b) = \max_{a \in \mathcal{A}} \left[R^b(b, a) + \gamma \sum_{b' \in \mathcal{B}} T^b(b, a, b') \, V_t(b') \right]$$

To address scalability: Do Value iteration + pruning at every t

Approximate solvers

- Point based algorithms
  - Maintain a fixed set of candidate beliefs (either random or reachable after few steps) and only update those

- QMDP
  - Ignore the observations, and update only on the basis of states and transitions

- SARSOP
  - Maintain and continuously refine the set of candidate beliefs

- Finite State Controller (FSC)