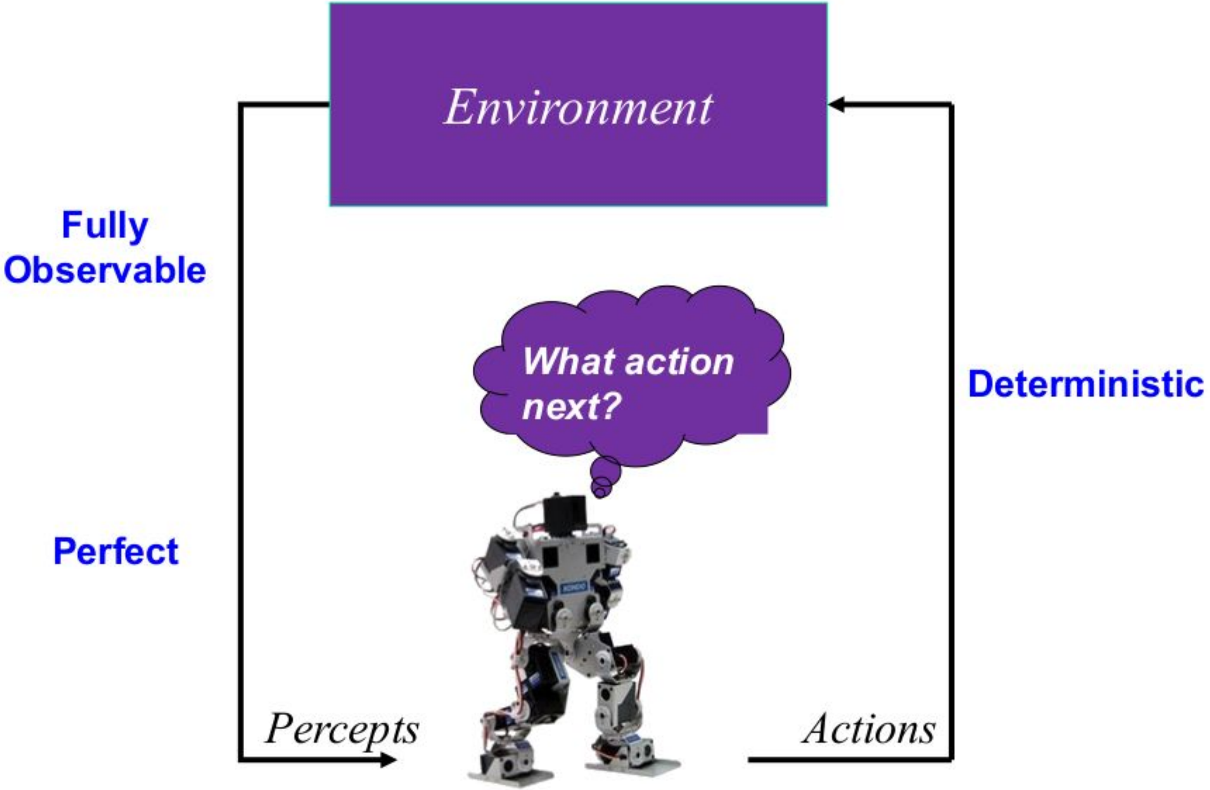
Partially Observable Markov decision process (POMDP)

Classical

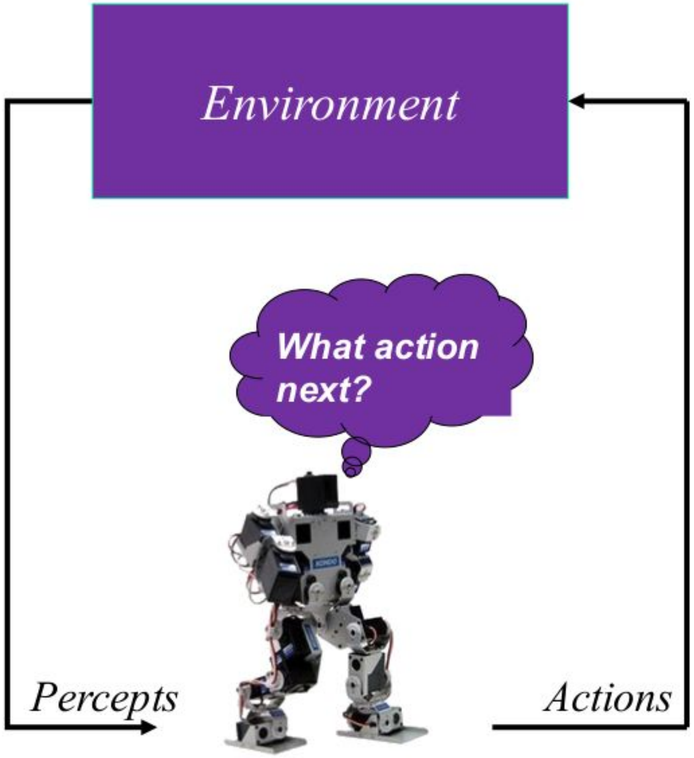

Environment
Fully
Observable
What action
next?
Deterministic
Perfect
Percepts
Actions

Stochastic (MDP)

- **S**: set of states
- **A**: set of actions
- **Pr**(s' | s,a): transition model
- **R**(s,a,s'): reward model
- γ: discount factor
- s_0: start state

Fully Observable

Perfect



Stochastic T(s, a, s')

Objective of a Fully Observable MDP

- Find a policy

  $\pi: \mathbf{S} \rightarrow \mathbf{A}$

- which maximizes expected discounted reward

  - given an infinite horizon
  - assuming full observability

Partially-Observable Stochastic (POMDP)

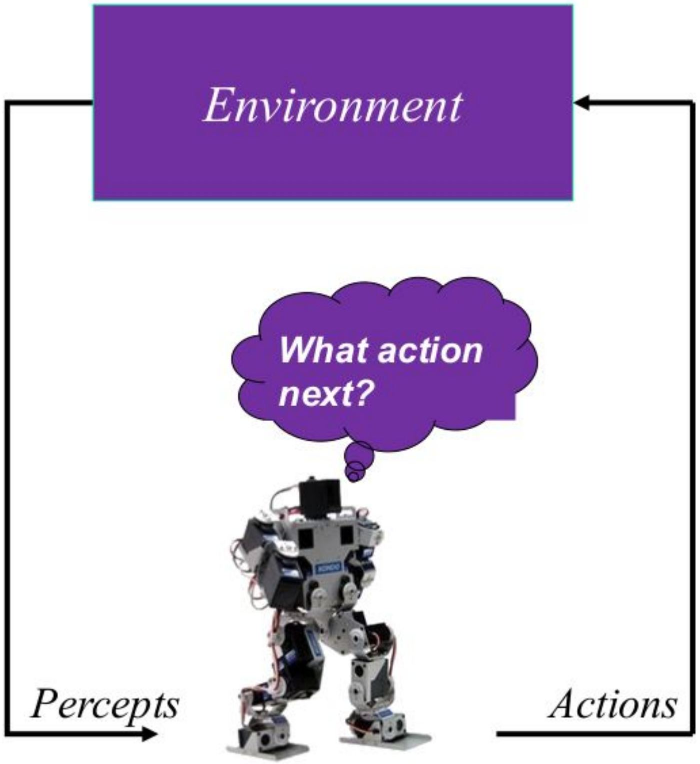
- **S**: set of states
- **A**: set of actions
- **Pr**(s'|s,a): transition model
- **R**(s,a,s'): reward model
- γ: discount factor
- s_0: start state

O set of observation

O(o| a, s')* = *Pr(o| a, s', b)

Partially Observable

Noisy



Stochastic
T(s, a, s')

Objective of a POMDP

- Find a policy

  π: BeliefStates($\mathbf{S}$) → $\mathbf{A}$

  A belief state is a ***probability distribution*** over states

- which maximizes expected discounted reward

  - given an infinite horizon

  - assuming ***partial*** & ***noisy*** observability

Markov decision process

- MDP adds a set of possible actions at each time step.
- The action a_t changes the transition probabilities, written as $\Pr(s_{t+1} \mid s_t, a_t)$.
- The rewards can also depend on the action and written as $\Pr(r_{t+1} \mid s_t, a_t)$.
- An MDP produces a sequence $s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, r_4 \ldots$ of states, actions, and rewards

MDP

Algorithm 1: Value iteration (Bellman, 1957)

input : MDP problem, convergence parameter ε

output: A policy that is ε-optimal for all states

begin

Initialize V'

repeat

$V \leftarrow V'$

for *each state* s **do**

$V'(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} p(s'|s,a)V(s')$

until *CloseEnough*(V, V')

return Greedy policy with respect to V'

end

POMDP

- A partially observable Markov decision process (POMDP) is a generalization of a Markov decision process (MDP).
- When the environment is only partially observable, the agent does not necessarily know which state it is in, so it cannot execute the action $\boldsymbol{\pi}(\mathbf{s})$ recommended for that state.
- POMDPs are difficult as utility of state **s** and optimal action in **s** depends on s.

POMDP

- POMDP: the state is not directly visible. Instead, the agent receives an observation o_t drawn from $\Pr(o_t \mid s_t)$.
- A POMDP generates a sequence $s_1, o_1, a_1, r_2, s_2, o_2, a_2, r_3, o_3, a_3, s_3, r_4, \ldots$ of states, observations, actions, and rewards.
- Each observation will be more compatible with some states but insufficient to identify the state uniquely

POMDP

Transition model for belief-states: Let's calculate the probability that an agent in belief state b reaches belief state b' after executing action a.

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

Sensor model: Probability of perceiving o, given that $\boldsymbol{a}$ was performed starting in belief state b, is given by summing over all the actual states s' that the agent might reach:

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

POMDP (belief state)

When action a is taken in belief state $b(s)$ and o is observed, the new belief $b'(s')$ can be calculated as follows

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

we can remove $\Pr(a, b)$

POMDP (belief state)

When action a is taken in belief state $b(s)$ and o is observed, the new belief $b'(s')$ can be calculated using Bayes' rule:

$$b'(s') = \Pr(s' \mid b, a, o) = \frac{\Pr(o \mid a, s', b) \Pr(s' \mid a, b)}{\Pr(o \mid a, b)}$$

POMDP (Sensor Model)

When action a is taken in belief state $b(s)$ and o is observed, the new belief $b'(s')$ can be calculated using Bayes' rule:

$$b'(s') = \Pr(s' \mid b, a, o) = \frac{O(o \mid a, s') \sum_s b(s) \Pr(s' \mid s, a)}{\Pr(o \mid a, b)}$$

Probability of the observation can be computed by summing over all possible s'

$$\begin{aligned} \Pr(o \mid a, b) &= \sum_{s'} \Pr(o \mid a, s', b) \Pr(s' \mid a, b) \\ &= \sum_{s'} O(o \mid a, s') \Pr(s' \mid a, b) \end{aligned}$$

POMDP (Sensor model)

When action a is taken in belief state $b(s)$ and o is observed, the new belief $b'(s')$ can be calculated using Bayes' rule:

$$b'(s') = \Pr(s' \mid b, a, o) = \frac{O(o \mid a, s') \sum_s b(s) \Pr(s' \mid s, a)}{\Pr(o \mid a, b)}$$

Probability of the observation can be computed by summing over all possible s'

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

POMDPs (Transition Model)

We can now define a new "belief-state MDP" with the following transition model:

$$\begin{aligned} Pr(b'|b,a) &= \sum_{o} Pr(b'|o,a,b)Pr(o|a,b) \\ &= \sum_{o} Pr(b'|o,a,b)\sum_{s'} O(o|a,s')\sum_{s} b(s)P(s'|s,a) \end{aligned}$$

POMDPs (Reward function)

We can now define a new “belief-state MDP” with the following transition model:

$$
\begin{aligned}
Pr(b'|b,a) &= \sum_o Pr(b'|o,a,b) Pr(o|a,b) \\
&= \sum_o Pr(b'|o,a,b) \sum_{s'} O(o|a,s') \sum_s b(s) P(s'|s,a)
\end{aligned}
$$

And the following reward function:

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

POMDPs (Policy)

- In POMDPs, an optimal policy $\pi^*(b)$ maps beliefs to actions. The policy $\pi(b)$ is a function over a continuous set of probability distributions over S.

POMDPs (Policy and Value function)

- In POMDPs, an optimal policy $\pi^*(b)$ maps beliefs to actions. The policy $\pi(b)$ is a function over a continuous set of probability distributions over S.

- A policy π can be characterized by a value function $V^\pi : \Delta(S) \rightarrow \mathbb{R}$ which is defined as the expected future discounted reward $V^\pi(b)$ the agent can gather by following π starting from belief b :

$$V^\pi(b) = E_\pi\left[\sum_{t=0}^{h} \gamma^t R\left(b_t, \pi\left(b_t\right)\right) \mid b_0 = b\right]$$

where $R\left(b_t, \pi\left(b_t\right)\right) = \sum_{s \in S} R\left(s, \pi\left(b_t\right)\right) b_t(s)$.

POMDPs (value iteration)

- The value of an optimal policy π^* is identified by the optimal value function V^*. Considering the Bellman optimality equation we have:

$$V^*(b) = \max_{a \in A} \left[\sum_{s \in S} R(s,a) b(s) + \gamma \sum_{o \in O} p(o \mid a, b) \; V^*(b^{ao}) \right]$$

with b^{ao} given by Bayes' rule for $b'(s')$, and $p(o \mid a, b)$ is the sensor model.

- When above equation holds for every $b \in \Delta(S)$ we are ensured the solution is optimal.

POMDPs

- Computing value functions over a continuous belief space fortunately is possible as the value function has a particular structure that we can exploit (Sondik, 1971).

- Value function can be parameterized by a finite number of vectors and has a convex shape.

Dempster–Shafer theory (DST)

Introduction

- Dempster–Shafer theory (DST), is a general framework for reasoning with uncertainty.
- Arthur P. Dempster first introduced DST in the context of statistical inference, later Glenn Shafer developed a general framework.
- In DST, a degree of belief is referred to as a mass and formulates a belief function rather than a Bayesian probability distribution.

Introduction

- Let X be the universe: the set representing all possible states of a system under consideration. The power set 2^X is the set of all subsets of X, including the empty set $\emptyset$.
- For example, if: $X = \{a, b\}$ then $2^X = \{\emptyset, \{a\}, \{b\}, X\}$.
- DST assigns a belief mass to each element of the power set.

Belief Assignment

- Formally, a function $m : 2^X \rightarrow [0, 1]$ is called a basic belief assignment (BBA), when it has two properties.
  - First, the mass of the empty set is zero: $m(\emptyset) = 0$.
  - Second, the masses of all the members of the power set add up to a total of 1: $\sum_{A \in 2^X} m(A) = 1$.

Mass function

- The mass m(A) of A, a given member of the power set, expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to A but to no particular subset of A.
- The value of m(A) pertains only to the set A and makes no additional claims about any subsets of A, each of which have, by definition, their own mass.

Belief and Plausibility

- From the mass assignments, the upper and lower bounds of a probability interval can be defined.

- This interval is bounded by two non-additive continuous measures called belief (or support) and plausibility:

  - $\text{bel}(A) \leq P(A) \leq \text{pl}(A)$.

  - The belief bel(A) for a set A is defined as the sum of all the masses of subsets of the set of interest:
    $\text{bel}(A) = \sum_{B|B \subseteq A} m(B)$

Belief and Plausibility

- From the mass assignments, the upper and lower bounds of a probability interval can be defined.

- This interval is bounded by two non-additive continuous measures called belief (or support) and plausibility:

  - $\mathrm{bel}(A) \leq P(A) \leq \mathrm{pl}(A)$.

  - The belief bel(A) for a set A is defined as the sum of all the masses of subsets of the set of interest:
    $\mathrm{bel}(A) = \sum_{B|B\subseteq A} m(B)$

  - The plausibility pl(A) is the sum of all the masses of the sets B that intersect the set of interest A: $\mathrm{pl}(A) = \sum_{B|B\cap A\neq\emptyset} m(B)$.

Belief and Plausibility

- The two measures are related to each other as follows: $\mathrm{pl}(A) = 1 - \mathrm{bel}(\overline{A})$.

Example

Hypothesis	Mass	Belief	Plausibility
Neither (alive nor dead)	0	0	0
Alive	0.2	0.2	0.5
Dead	0.5	0.5	0.8
Either (alive or dead)	0.3	1.0	1.0

Belief and Plausibility

- The two measures are related to each other as follows: $\mathrm{pl}(A) = 1 - \mathrm{bel}(\overline{A})$.
- For finite A, given the belief measure bel(B) for all subsets B of A, we can find the masses m(A) with the following inverse function:

$$m(A) = \sum_{B | B \subseteq A} (-1)^{|A-B|} \,\mathrm{bel}(B) \tag{1}$$

where |A - B| is the difference of the cardinalities of the two sets.

DST and Bayesian Theory

- In the generalized probability view of DS theory, belief and plausibility are regarded as lower and upper bounds respectively for an underlying probability which is unknown.

$$\mathrm{bel}(A) \leq P(A) \leq \mathrm{pl}(A)$$

- Bayesian framework assign probabilities to a single event. In DST, probability values are assigned to a set of possibilities.
- the generalization of DST allows one to compute the posterior belief/plausibility given the likelihood beliefs/plausibilities and prior beliefs/plausibilities.