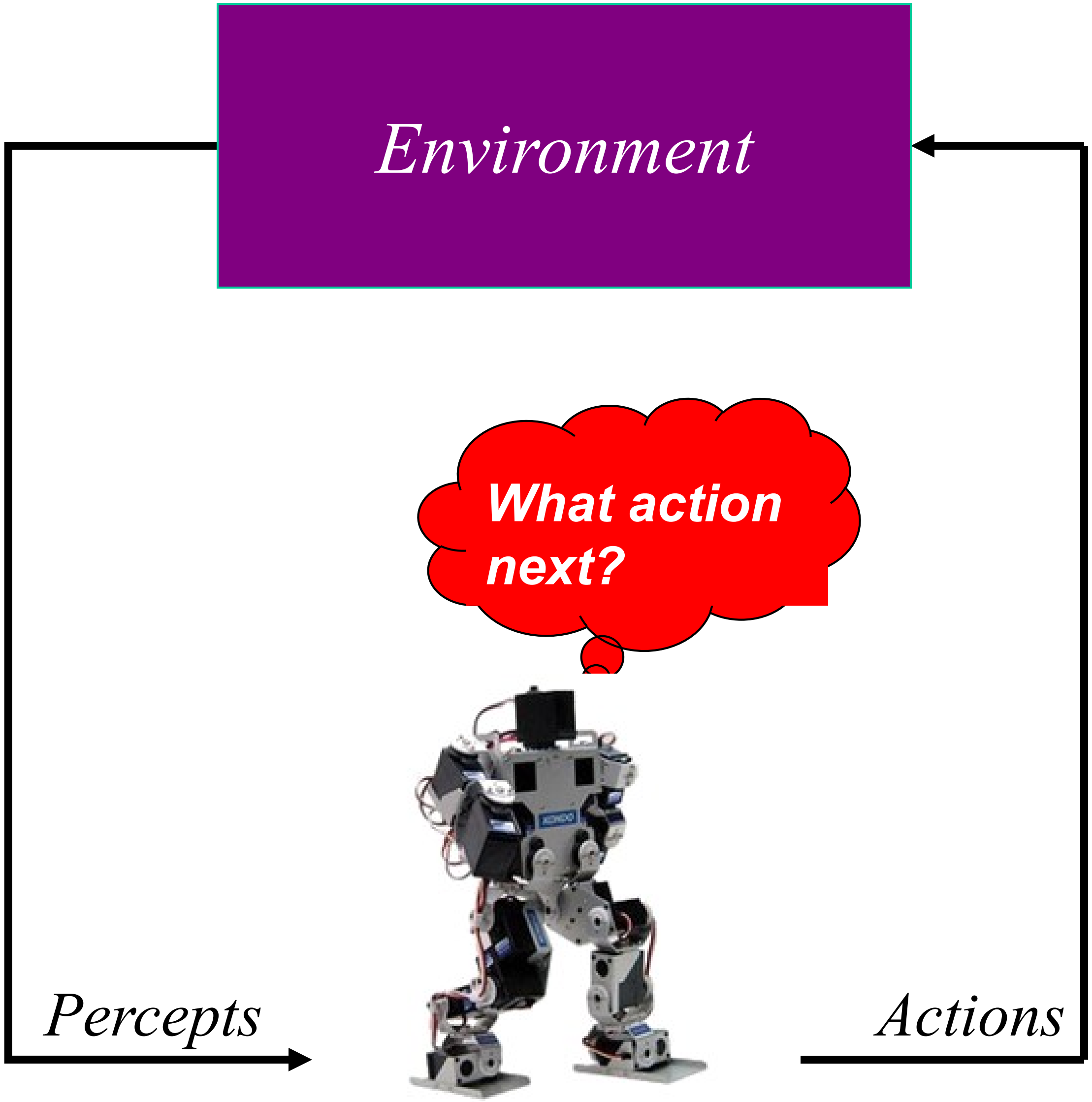
Markov Decision Processes

Chapter 17

Mausam

Planning Agent

Static vs. Dynamic



Environment

What action next?

Percepts

Actions

Fully vs. Partially Observable

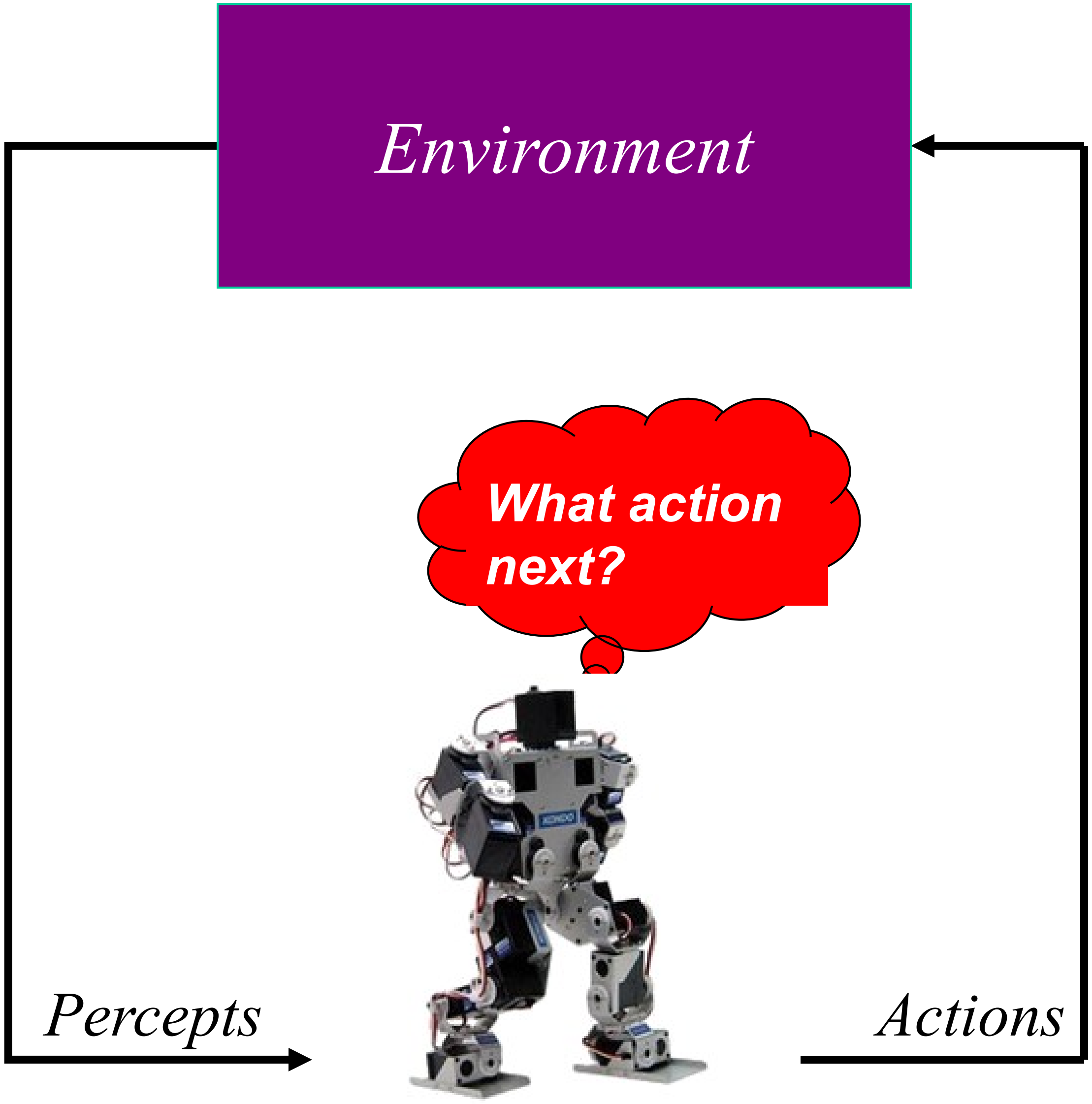
Perfect vs. Noisy

Deterministic vs. Stochastic

Instantaneous vs. Durative

Search Algorithms

Static



Environment

What action next?

Percepts

Actions

Fully Observable

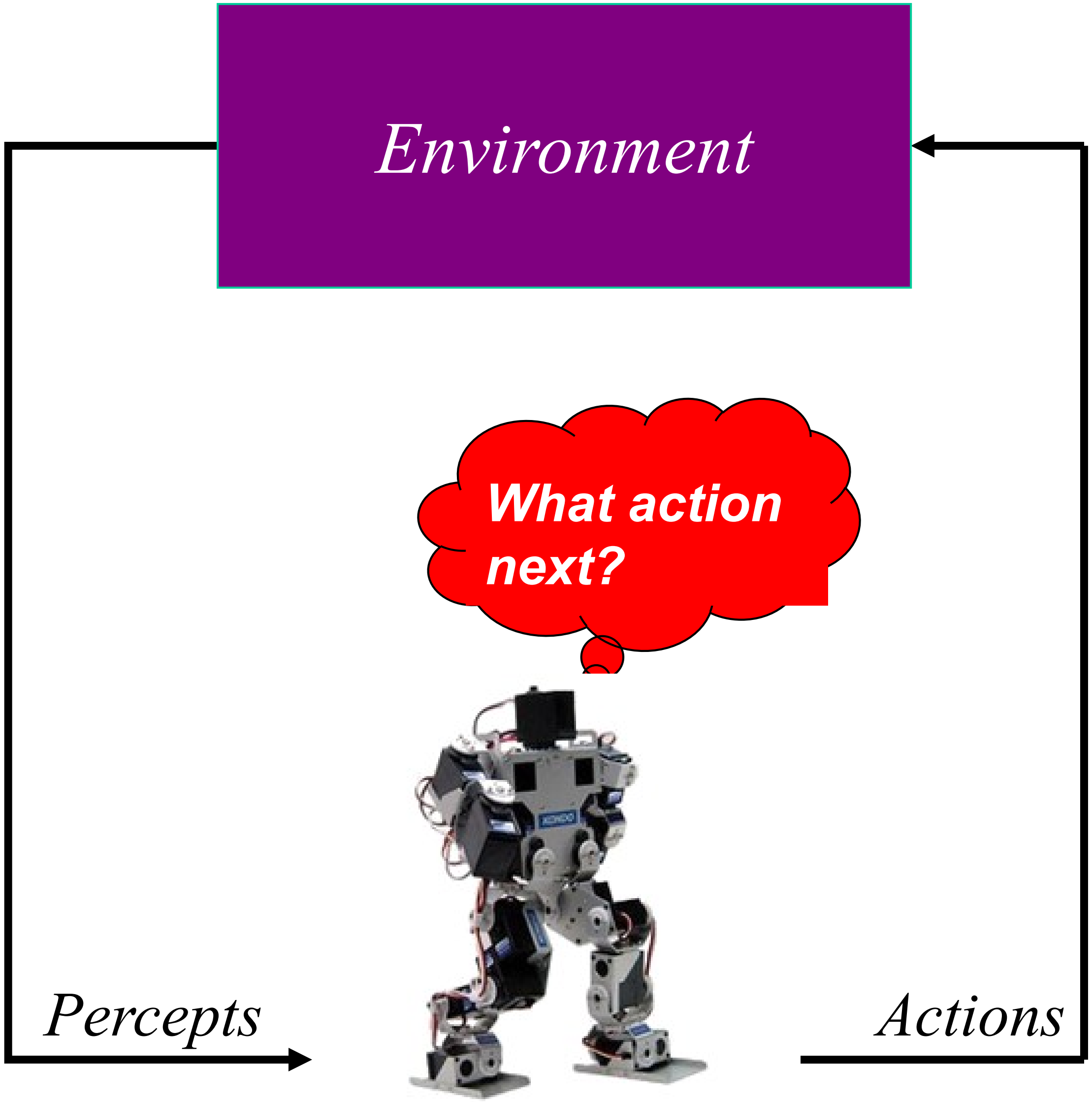
Perfect

Deterministic

Instantaneous

Stochastic Planning: MDPs

Static



Environment

Fully Observable

What action next?

Stochastic

Instantaneous

Perfect

Percepts

Actions

MDP vs. Decision Theory

- Decision theory - episodic

- MDP -- sequential

Markov Decision Process (MDP)

- $\mathcal{S}$: A set of states
- $\mathcal{A}$: A set of actions
- $\mathcal{T}$(s,a,s'): transition model
- $\mathcal{C}$(s,a,s'): cost model
- $\mathcal{G}$: set of goals
- s_0: start state
- γ: discount factor
- $\mathcal{R}$(s,a,s'): reward model

factored

Factored MDP

absorbing/
non-absorbing

Objective of an MDP

- Find a policy $\pi: \mathcal{S} \rightarrow \mathcal{A}$

- which optimizes

$$\begin{array}{l} \bullet\ \text{minimizes} \\ \bullet\ \text{maximizes} \\ \bullet\ \text{maximizes} \end{array} \begin{pmatrix} \text{discounted} \\ \text{or} \\ \text{undiscount.} \end{pmatrix} \begin{array}{l} \text{expected cost to reach a goal} \\ \text{expected reward} \\ \text{expected (reward-cost)} \end{array}$$

- given a _____ horizon
  - finite
  - infinite
  - indefinite

- assuming full observability

Role of Discount Factor (γ)

- Keep the total reward/total cost finite
  - useful for infinite horizon problems

- Intuition (economics):
  - Money today is worth more than money tomorrow.

- Total reward: $r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots$
- Total cost: $c_1 + \gamma c_2 + \gamma^2 c_3 + \ldots$

Examples of MDPs

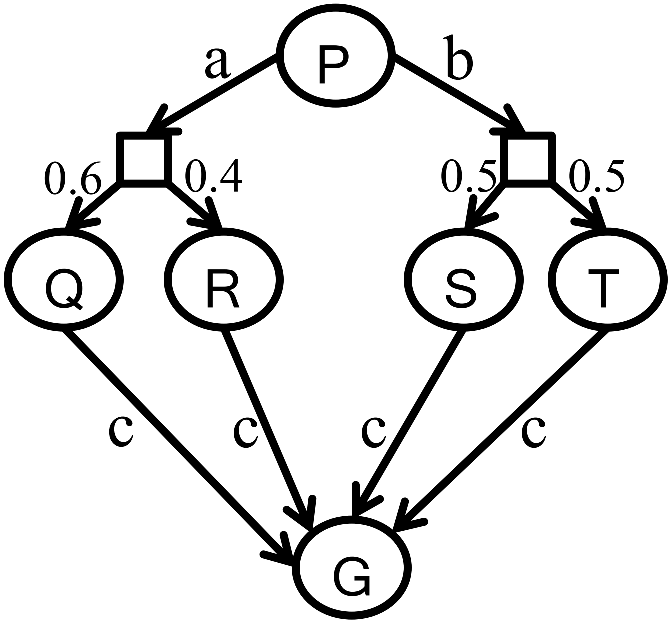
- Goal-directed, Indefinite Horizon, Cost Minimization MDP
  - $<\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{C}, \mathcal{G}, s_0>$
  - Most often studied in planning, graph theory communities

- Infinite Horizon, Discounted Reward Maximization MDP
  - $<\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma>$
  - Most often studied in machine learning, economics, operations research communities

most popular

- Oversubscription Planning: Non absorbing goals, Reward Max. MDP
  - $<\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{G}, \mathcal{R}, s_0>$
  - Relatively recent model

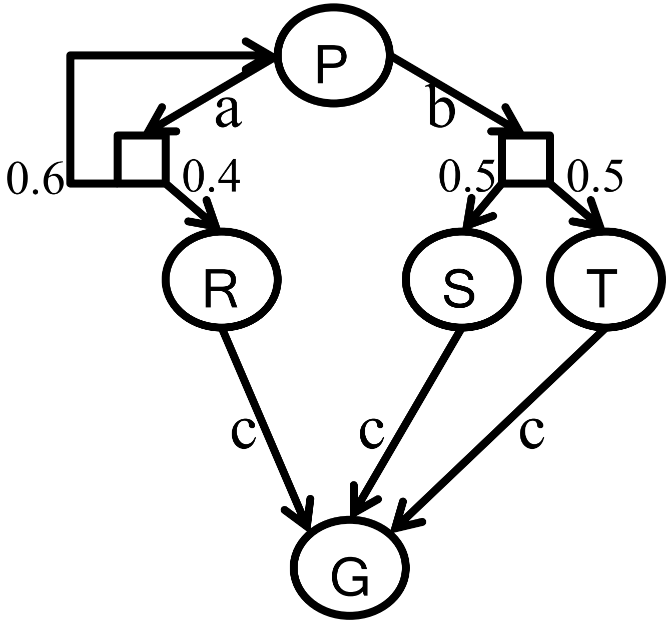
Acyclic vs. Cyclic MDPs



C(a) = 5, C(b) = 10, C(c) =1

Expectimin works

- V(Q/R/S/T) = 1
- V(P) = 6 – action a



Expectimin doesn't work

- •infinite loop

- V(R/S/T) = 1
- Q(P,b) = 11
- Q(P,a) = ????
- suppose I decide to take a in P
- Q(P,a) = 5+ 0.4*1 + 0.6Q(P,a)
- ➔ = 13.5