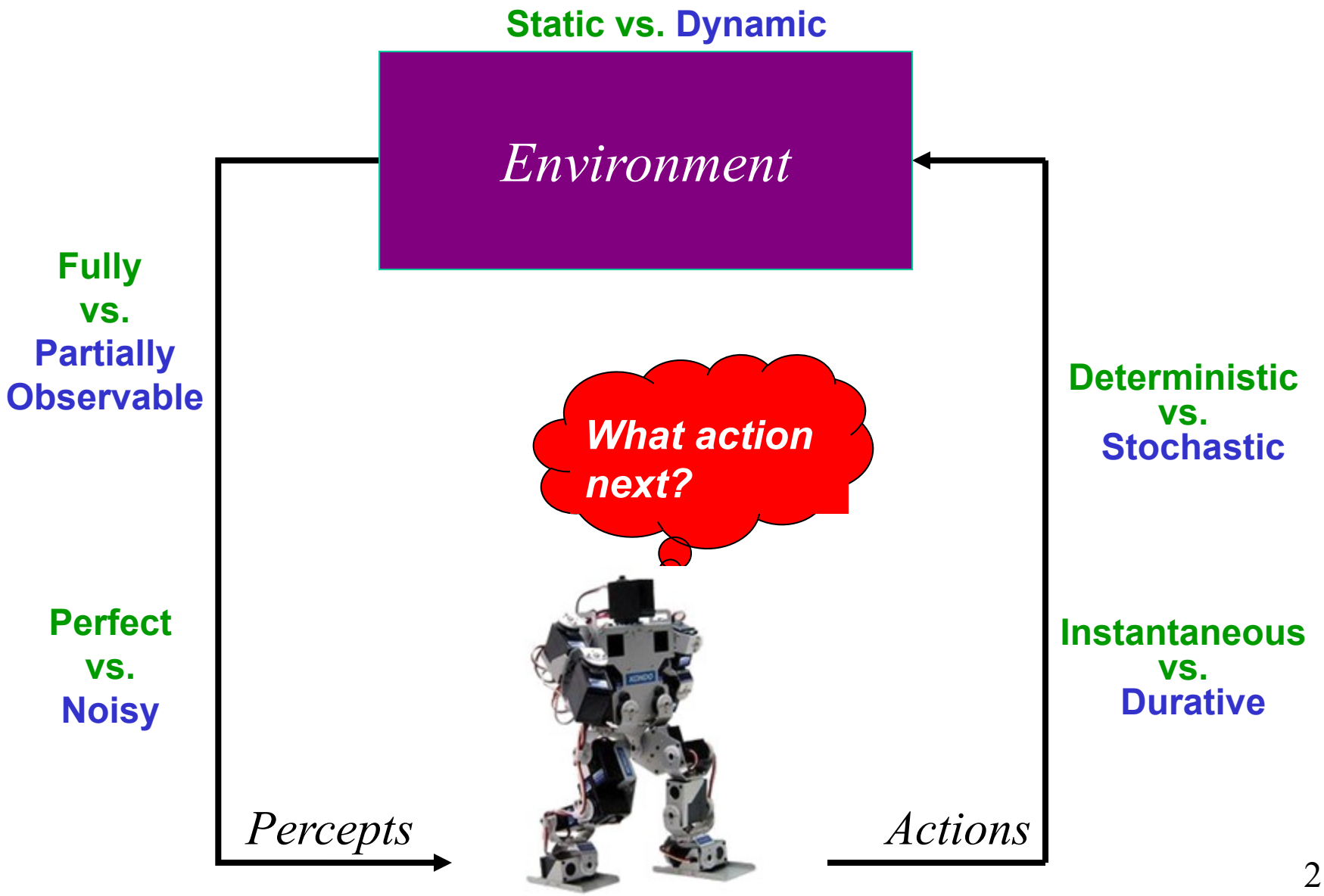
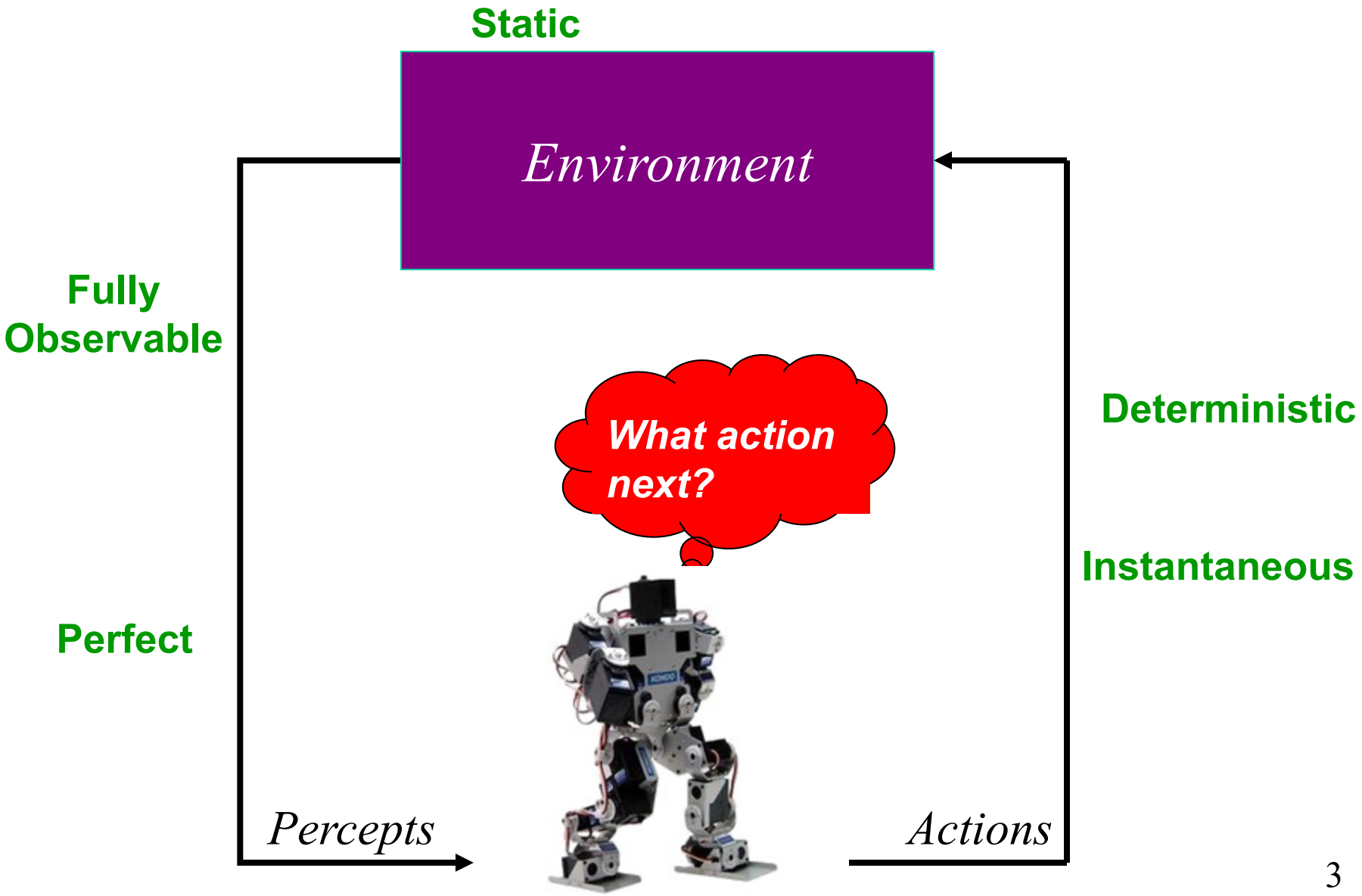
Markov Decision Processes
Chapter 17

Mausam

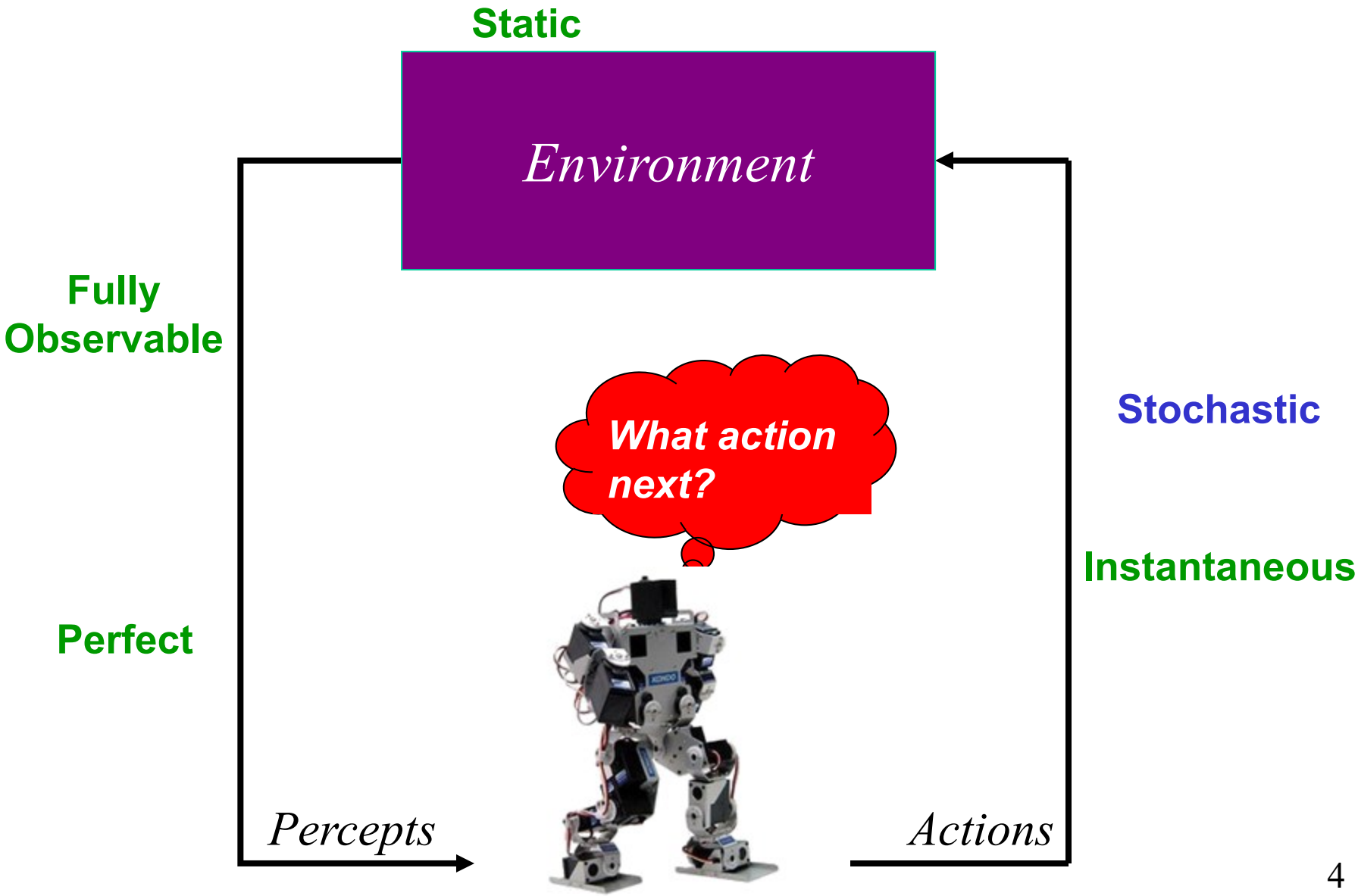
Planning Agent



Search Algorithms



Stochastic Planning: MDPs



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 π: $\mathcal{S} \rightarrow \mathcal{A}$
- which optimizes
  - minimizes
  - maximizes
  - maximizes

  [discounted / or / undiscount.] expected cost to reach a goal / expected reward / expected (reward-cost)
- 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: $c_1 + \gamma c_2 + \gamma^2 c_3 + \ldots$

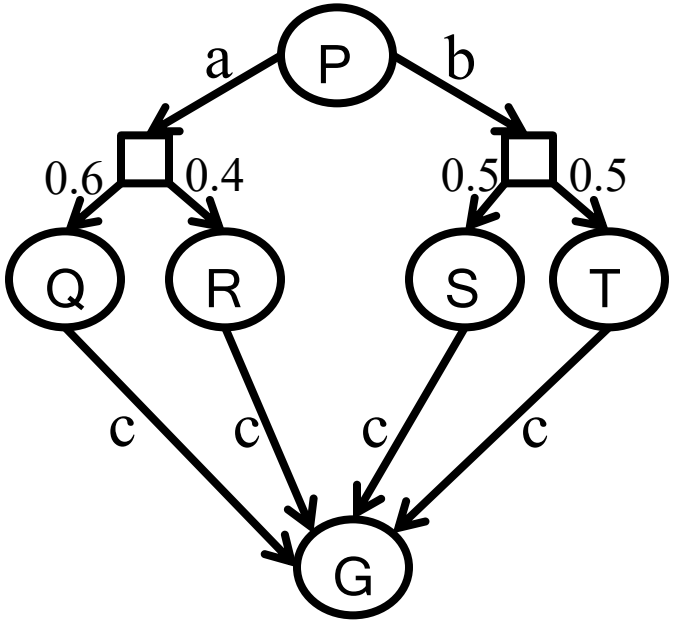
Examples of MDPs

- 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

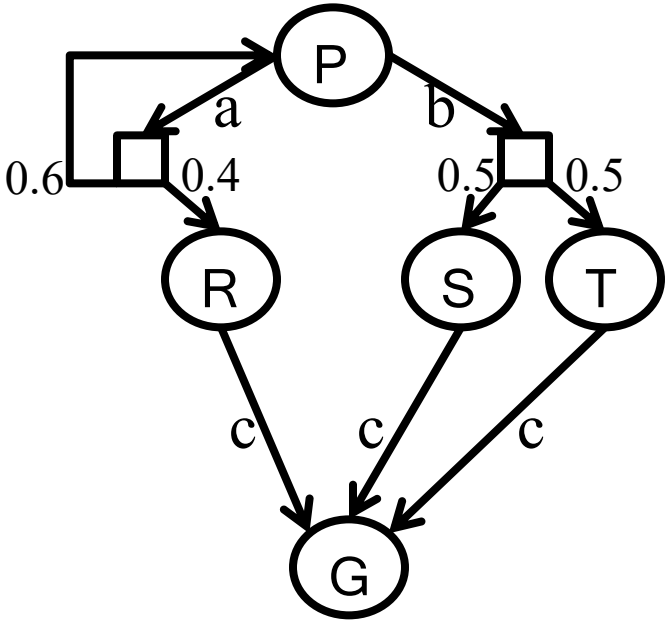
Acyclic vs. Cyclic MDPs



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

Brute force Algorithm

- Go over all policies π
  - How many? $|A|^{|S|}$ ← finite
- Evaluate each policy ← how to evaluate?
  - $V^{\pi}(s) \leftarrow$ expected cost of reaching goal from s
- Choose the best
  - We know that best exists (SSP optimality principle)
  - $V^{\pi *}(s) \leq V^{\pi}(s)$

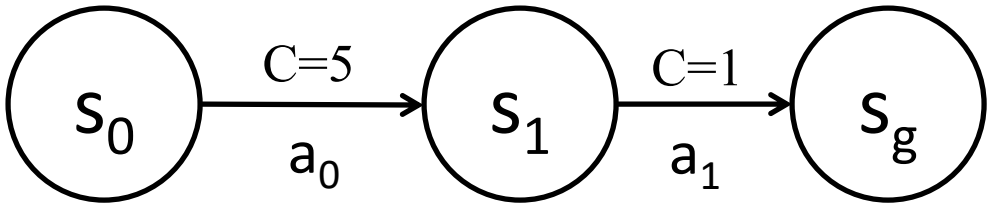
Policy Evaluation

- Given a policy π: compute V^{π}
  - V^{π} : cost of reaching goal while following π

Deterministic MDPs

- Policy Graph for π

  $\pi(s_0) = a_0$; $\pi(s_1) = a_1$

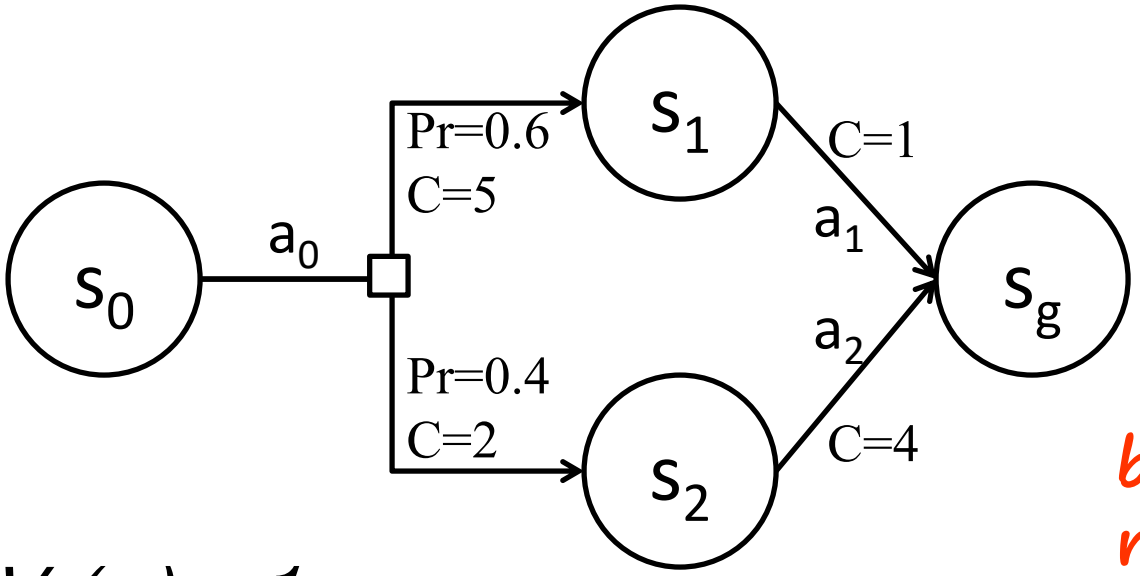


- $V^{\pi}(s_1) = 1$
- $V^{\pi}(s_0) = 6$

add costs on **path** to goal

Acyclic MDPs

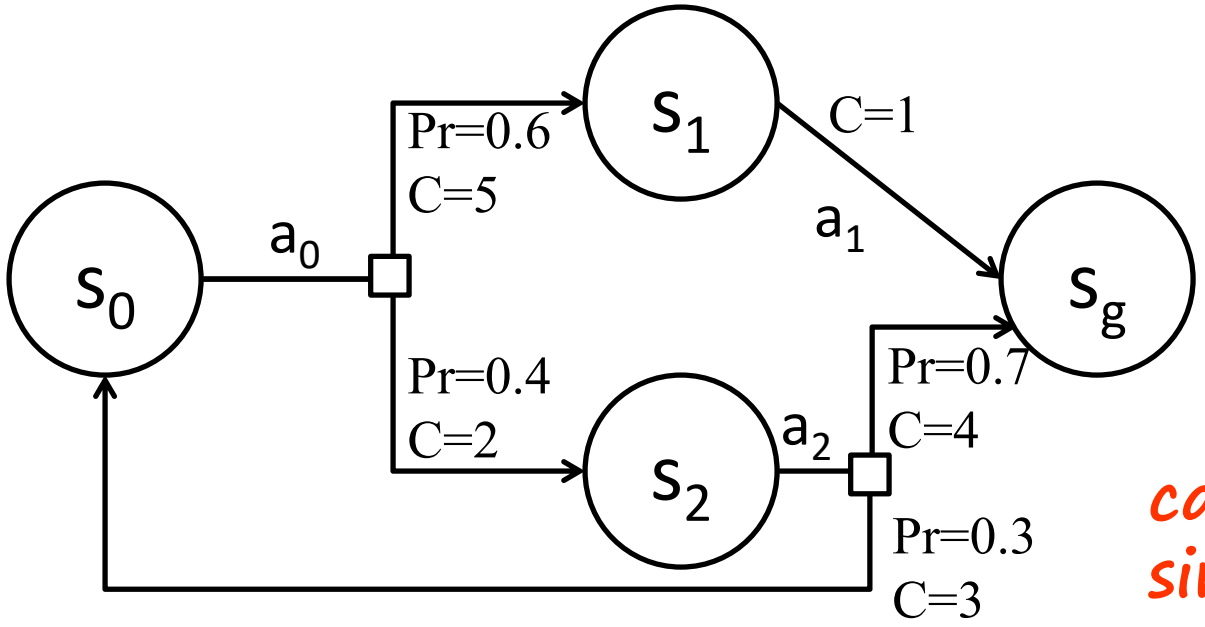
- Policy Graph for π



- $V^{\pi}(s_1) = 1$
- $V^{\pi}(s_2) = 4$
- $V^{\pi}(s_0) = 0.6(5+1) + 0.4(2+4) = 6$

backward pass in reverse topological order

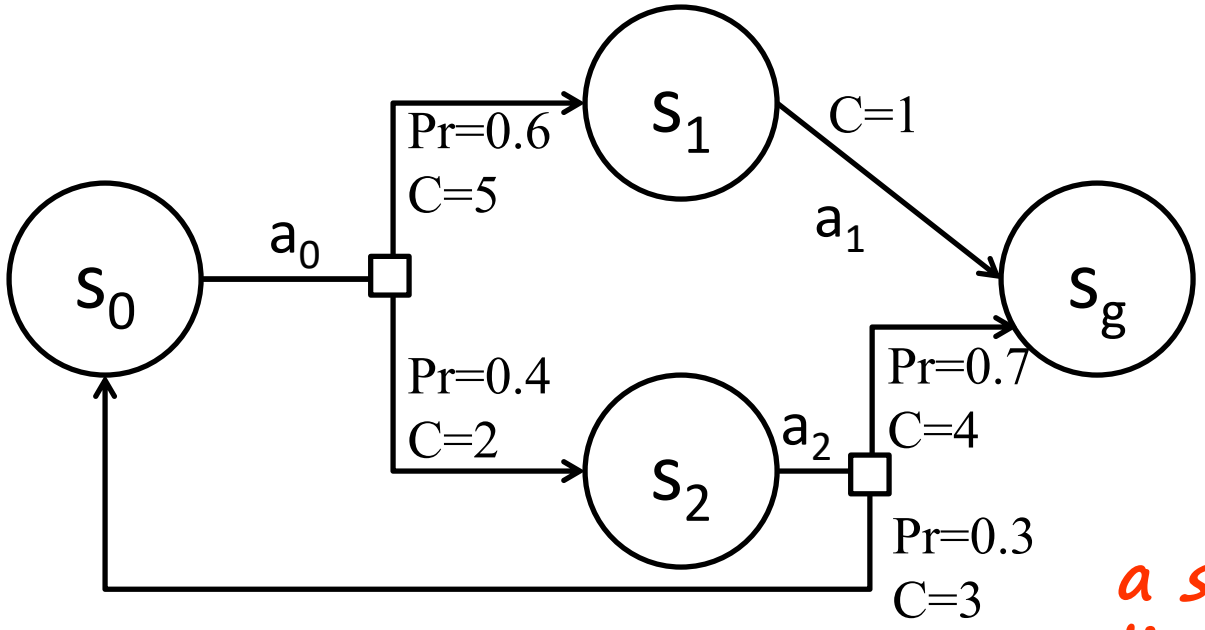
General MDPs can be cyclic!



cannot do a simple single pass

- $V^{\pi}(s_1) = 1$
- $V^{\pi}(s_2)$ = ?? (depends on $V^{\pi}(s_0)$)
- $V^{\pi}(s_0)$ = ?? (depends on $V^{\pi}(s_2)$)

General SSPs can be cyclic!



a simple system of linear equations

- $V^{\pi}(g) = 0$
- $V^{\pi}(s_1) = 1 + V^{\pi}(s_g) = 1$
- $V^{\pi}(s_2) = 0.7(4 + V^{\pi}(s_g)) + 0.3(3 + V^{\pi}(s_0))$
- $V^{\pi}(s_0) = 0.6(5 + V^{\pi}(s_1)) + 0.4(2 + V^{\pi}(s_2))$

Policy Evaluation (Approach 1)

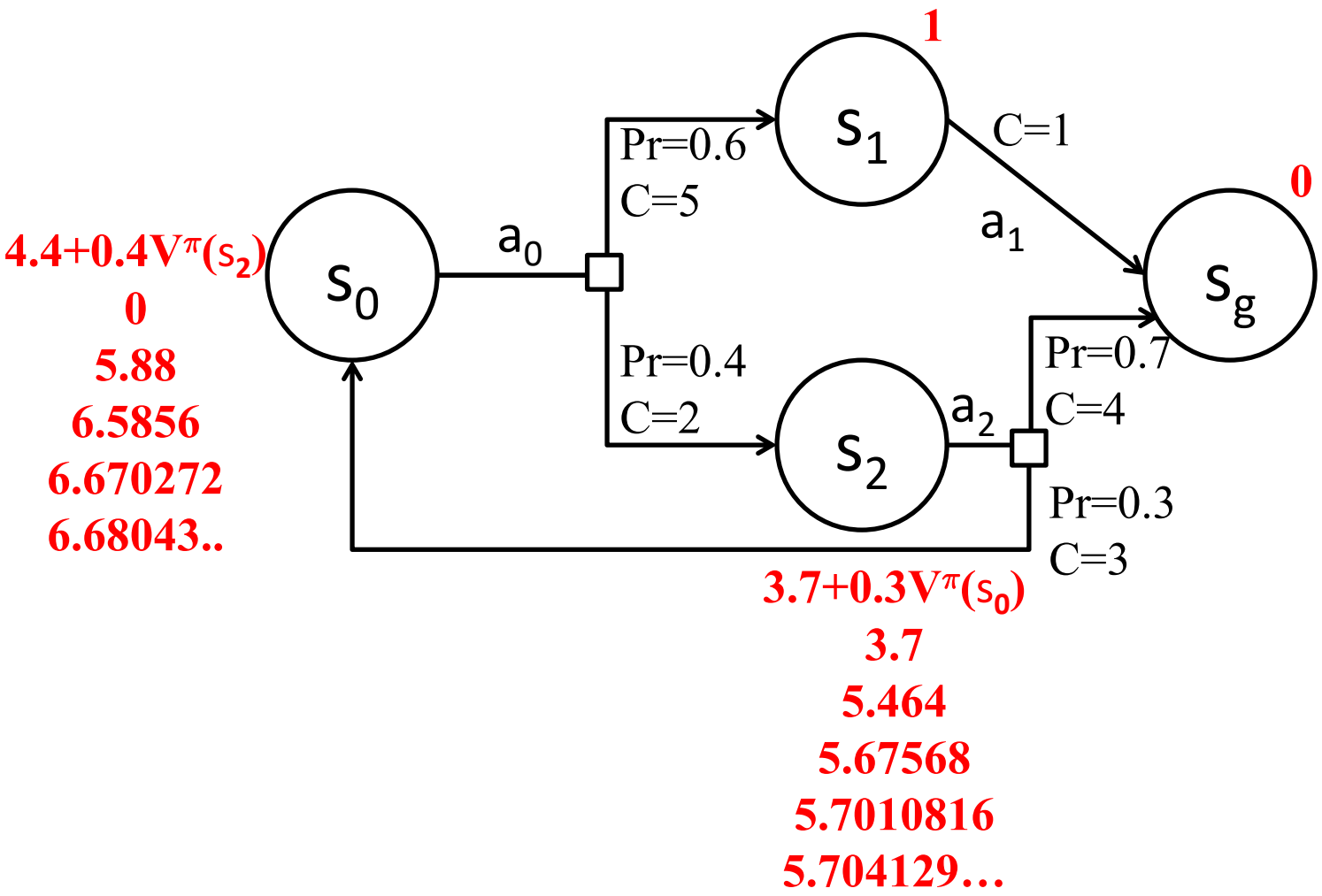
- Solving the System of Linear Equations

$$V^{\pi}(s) = 0 \quad \text{if } s \in \mathcal{G}$$

$$= $$

- $|S|$ variables.
- $O(|S|^3)$ running time

Iterative Policy Evaluation



Policy Evaluation (Approach 2)

$$V^{\pi}(s) = \sum_{s' \in \mathcal{S}} \mathcal{T}(s, \pi(s), s') \left[\mathcal{C}(s, \pi(s), s') + V^{\pi}(s') \right]$$

iterative refinement

$$V_n^{\pi}(s) \leftarrow \sum_{s' \in \mathcal{S}} \mathcal{T}(s, \pi(s), s') \left[\mathcal{C}(s, \pi(s), s') + V_{n-1}^{\pi}(s') \right]$$

Iterative Policy Evaluation

1 *//Assumption: π is proper*
2 initialize V_0^π arbitrarily for each state
3
4
5
6
7
8
9
10
11

iteration n

ϵ-consistency

termination condition

Convergence & Optimality

For a proper policy π

Iterative policy evaluation

converges to the true value of the policy, i.e.

$\lim_{n \to \infty} V_n^\pi = V^\pi$

irrespective of the initialization V_0