Markov Decision Processes
Chapter 17
Mausam

• Decision theory - episodic
• MDP -- sequential

Objective of an MDP
• Find a policy \( \pi: V \rightarrow D \)
• which optimizes
  - minimizes \( \gamma \) expected cost to reach a goal
  - maximizes \( \gamma \) expected reward
  - maximizes \( (1-\gamma) \) expected (reward-cost)
• given a horizon
  - finite
  - infinite
  - indefinite
• assuming full observability

Examples of MDPs
• Goal-directed, Indefinite Horizon, Cost Minimization MDP
  - \(<V, D, S, \pi, J, s_0>\)
  - Most often studied in planning, graph theory communities
• Infinite Horizon, Discounted Reward Maximization MDP
  - \(<V, D, S, R, J, \gamma, s_0>\)
  - Most often studied in machine learning, economics, operations research communities
• Oversubscription Planning: Non absorbing goals, Reward Max. MDP
  - \(<V, D, S, J, 0, s_0>\)
  - Relatively recent model

Role of Discount Factor (\( \gamma \))
• 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: \( c_1 + \gamma c_2 + \gamma^2 c_3 + \ldots \)
AND/OR Acyclic Graphs vs. 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(Pb) = 11
    - Q(Pa) = ????
    - suppose I decide to take a in P
      - Q(Pa) = 5 + 0.4*1 + 0.6Q(Pa)
      - = 13.5

Bellman Equations for MDP

- Define J*(s) {optimal cost} as the minimum expected cost to reach a goal from this state.
- J* should satisfy the following equation:

\[ J^*(s) = 0 \text{ if } s \in \mathcal{G} \]

Bellman Equations for MDP

- Define V*(s) {optimal value} as the maximum expected discounted reward from this state.
- V* should satisfy the following equation:

\[ V^*(s) = \max_{a \in A(s)} \sum_{s' \in S} P(s'|s, a) [R(s, a, s') + \gamma V^*(s')] \]