Lecture 16 (Markov Decision Processes)

1 Intuition for Formalisation

1. Start state
2. Good and bad goal states
3. Small "living" rewards are given (+ve or -ve)
4. Large rewards on reaching a goal state (+ve or -ve)
5. Agent's goal is to maximise sum of rewards
6. Rewards are instantaneous

2 Markov Decision Process (MDP)

1. Set of states S
2. Set of actions A
3. Transition function $T(s, a, s')$ gives $P(s'|s, a)$
4. Reward function $R(s, a, s')(= R(s')$ sometimes), small negative values for irrelevant states as "cost of breathing"
5. Start state S_0
6. Possibly ternimation states

3 Policies

1. For each state, we need an optimal policy $\pi^* : S \to A$
2. Optimal policy maximises expected utility
3. Agent looks up policy on arriving at state

3.1 Policies in terms of Reward

1. Invalid transitions can be penalised more
2. This creates urgency to reach goal state soon
3. Large negative reward can be harmful since agent will then just reach a terminal state (irrespective of good/bad)

4 Markov Assumption in MDPs

1. Next state only depends on current state and current action
2. Past states and actions are irrelevant