Lecture 19 (Reinforcement Learning)

1 Setup

1. Similar to MDP but reward function R and transition probability T isn't known
2. Thus, we explore different actions and *learn* details about the environment and find the optimal policy
3. This algorithm is an *online* algorithm unlike MDP which is *offline*
4. Environment gives the reward after the action - action-reward loop
5. Humans learn from experiences in life

2 Difference from other Learning Problems

1. Supervised learning is like feeding data of both x, y; goal is to find a mapping
2. Unsupervised learning is feeding data of only x; goal is to find the structure
3. Reinforcement learning is given state-action pairs; goal is to maximise reward
4. Unlike the other two, RL is evaluative feedback

3 RL Agents

1. Utility-based agent
2. Q-learning
3. Reflex agent

4 RL Approaches

1. Passive learning
2. Active learning

4.1 Passive Learning

1. Input is the policy
2. No information about T or R
3. Run multiple instances (episodes/trials) and estimate the value function

4.1.1 Model-Based RL

1. From experiences *normalise* to estimate $\hat{T}(s, a, s')$ and then discover $\hat{R}(s, a, s')$
2. Now, use these values to estimate the optimal policy using value/policy iteration