Lecture 17 (Solving MDPs)

1 MDP as Search Tree

1. We use an algorithm similar to expectimax search
2. The probability comes from environment

2 Utility of Reward Sequences

1. Different sequences of rewards might have the same total outcome
2. To decide on the ordering, we add a *discount* factor
3. On descending a level, we multiply in the discount for all rewards
4. This helps in convergence of our algorithm too
5. Utility of an infinite sequence is finite

3 Optimal Quantities

1. $V^*(s)$ = expected utility starting in s and acting optimally
2. $Q^*(s,a)$ = expected utility starting in s taking action a
3. $\pi^*(s)$ = optimal action from s

3.1 Formulating - Bellman Equation

1. $V^*(s) = \max_a Q^*(s,a)$
2. $Q^*(s,a) = \sum_{s'}^{a} T(s,a,s')\left(R(s,a,s') + \gamma V^*(s')\right)$
3. $V^*(s) = \max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^*(s')\right)$

3.2 Computing - Value Iteration

$$V_{k+1}(s) = \max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V_k(s')\right)$$

Complexity of each iteration is $O(S^2 A)$

3.3 Policy Function

$\pi^*(s) = \underset{a}{argmax} \sum_{s'} T(s,a,s') \left(R(s,a,s') + \gamma V^*(s')\right)$