Lecture 20 (Model Free Reinforcement Learning)

1 Idea

We skip the computation of the model $(\hat{T}, \hat{R})$ and just compute the value functions

2 Direct Evaluation

1. Approximates the value function: $V^{\pi}(s)$
2. We compute the average of the discounted rewards for each state

2.1 Pros

1. Saves model estimation cost
2. With large number of episodes, the correct value function is computed

2.2 Cons

1. Each state is independent of the other, which makes the algorithm inefficient
2. Bellman characterization is ignored

3 Policy Evaluation

1. We attempt to use the Bellman equation but skip computation of T and R
2. This can be done by taking samples of outcomes s' by doing the action and averaging it
3. Assumes that we can re-visit at the original state

4 Temporal Difference

1. Generalise the above to learn from every experience
2. Make updates for each transition instead of re-setting like it was done in Policy Evaluation

$$sample = R(s, \pi(s), s') + \gamma V^{\pi}(s')$$

$$V^{\pi}(s) \leftarrow (1 - \alpha)V^{\pi}(s) + (\alpha)sample$$

$$\implies V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha(sample - V^{\pi}(s))$$

3. We cannot find the optimal policy from this computation

5 Q-Learning

5.1 Q-Value Iteration

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} (Q_k(s',a')) \right]$$

6 Idea for Q-Learning

$$sample = R(s, \pi(s), s') + \gamma \max_{a'} (Q_k(s', a'))$$

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha) sample$$