Lecture 12

Lai Zehua 2014012668

2017??12??12??

For $\gamma \in (0, 1)$, define

$$d(v, v') = \max_{s \in S} |v(s) - v'(s)|$$

then Bellman Expectation Operator is a contraction mapping with respect to l_∞-norm. So by iteration we can simply calculate $v_\pi(s)$ by Bellman Operator. Also, q_π can be computed by the equation.

$$q_\pi(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) v_\pi(s')$$

We would like to find a policy π^*, such that $v_{\pi^*}(s) \geq v_\pi(s), \forall \pi, s$, or $v_{\pi^*} \succ v_\pi$. From now on, assume $R(s, a)$ is bounded.

Algorithm:

$$v'(s) = \max_a [R(s, a) + \gamma \sum P(s'|s, a) v(s')]$$

Exercise 1: Prove: $\phi : R^{|S|} \to R^{|S|}, \forall s \in S, \phi(v(s)) = v'(s)$ is a contraction mapping with respect to l_∞-norm (but its value might not be a value of any policy). It is called the Bellman Optimality Operator.

Exercise 2: Prove: $v_{\pi'}(s) \geq v_\pi(s)$.

The fixed point for $\phi(v_\pi)$ must be the value function of a policy π^*.

It is clear that $\phi(v_\pi) \geq v_\pi$. And if $v, v' \in R^{|S|}$(not necessarily a value function), $v \succ v'$, then $\phi(v) \succ \phi(v')$. Thus, $\phi^{(n+1)}(v_\pi) \geq \phi^{(n)}(v_\pi)$. So the policy π^* is indeed the optimal policy.

In reinforce learning, this algorithm is called the value iteration. Another algorithm is called the policy iteration. For a initial policy π_0. $\pi_{n+1}(s) = \arg\max_a [R(s, a) + \gamma \sum P(s'|s, a) v_\pi(s')]$. $v_\pi(s)$ can be evaluated by using Bellman expectation operator.