Assignment 2

Gandharv Patil-260727335
COMP767

Monday 5^{th} February, 2018

Theorem 1. Bellman operator is a contraction mapping in the non linear case.

Preliminaries:

1. **Bellman update:**

$$V(s) = \underset{a}{max} \left(R(s,a) + \sum_{s'} \gamma P_a^{ss'} V(s') \right) \tag{1}$$

2. **Infinity Norm:** $\|u - v\|_\infty = \underset{s \in S}{max} |u(s) - v(s)|$

3. **To Prove:** $\underset{s \in S}{max} |B^*(u)(s) - B^*(v)(s)| \leqslant \underset{s \in S}{max} |u(s) - v(s)|$ i.e. B^* operator is a γ-contraction, i.e. it makes value functions closer by at least γ

Proof.

$$\begin{aligned}
&\underset{s}{max} |B^*(u)(s) - B^*(v)(s)| \\
&= \underset{s \in S}{max} \left| \underset{a}{max} \left(R(s,a) + \sum_{s'} \gamma P_a^{ss'} u(s') \right) - \underset{a}{max} \left(R(s,a) + \sum_{s'} \gamma P_a^{ss'} v(s') \right) \right| \\
&\leqslant \underset{s \in S}{max}\ \underset{a}{max} \left| R(s,a) + \gamma \sum_{s'} P_a^{ss'} u(s') - R(s,a) - \gamma \sum_{s'} P_a^{ss'} v(s') \right| \\
&\because max(a-b) \geqslant max(a) - max(b) \quad \forall a, b \\
&= \underset{s \in S}{max}\ \gamma \left| \sum_{s'} P_a^{ss'} u(s') - \sum_{s'} P_a^{ss'} v(s') \right| \\
&\leqslant \underset{s \in S}{max}\ \gamma \left| \sum_{s} P_a^{ss'} \underset{s \in S}{max} |(u(s) - v(s)| \right| \\
&\leqslant \gamma \underset{s \in S}{max} |u(s) - v(s)|
\end{aligned}$$

□

Question 2. Show that the values of two successive policies generated by policy iteration are nondecreasing. Assume a finite MDP and conclude (explain why) that policy iteration must terminate under a finite number of steps. Finally, show that upon termination, policy iteration must have found an optimal policy (i.e. one which satisfies the optimality equations).

Solution. Let V_n and V_{n+1} be the successive iterations of the policy iteration algorithm.

1. We have to prove that : $V_n \leqslant V_{n+1} \leqslant V^*$ where V^* = Optimal value function.
2. Let State-Space(S) and Action-space(A) be finite, then the Policy Iteration algorithm converges to the optimal policy after at most after $\left|A^{|S|}\right|$ iterations.

□

Proof. Let π_{n+1} be the policy in the policy improvement step which is chosen greedily with respected to V_n, then:

$$\begin{aligned} R_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} V_n^{\pi_n} &\geqslant R_{\pi_n} + \gamma P_{\pi_n} V_n^{\pi_n} = V_n^{\pi_n} \\ R_{\pi_{n+1}} &\geqslant \left(I - \gamma\, P_{\pi_{n+1}}\right) V_n \\ V_{n+1} &= \left(I - \gamma\, P_{\pi_{n+1}}\right)^{-1} \geqslant V_n \\ V_{n+1} &\geqslant V_n \end{aligned}$$

Finally, for every iteration since $V_{n+1} \geqslant V_n$ a policy $\pi \in \prod^{MD}$ is unique in every iteration except when $V_{n+1} = V_n = V^*$ and the ties are broken consistently while using the *argmax* operator.

$$\therefore N \leqslant \left|\prod^{MD}\right| \leqslant \left|A^{|S|}\right|$$

. □