Homework 3

Jingxi Xu
jingxixu@csail.mit.edu

Learning and Intelligent Systems Lab
Massachusetts Institute of Technology

1. (Bellman convergence) prove that the value iteration algorithm in a discrete MDP converges to a unique optimal value function. Does the policy converge uniquely? What would a counterexample be?

Convergence Using Bellman equation, the value function at state $S \in \mathcal{S}$ for policy π is defined as

$$V^{\pi}(s) := R(s) + \gamma \sum_{S' \in \mathcal{S}} \Pr[S' \mid s, \pi(s)] V^{\pi}(s')$$

Using value iteration, we update the value function for each state at each iteration by

$$\hat{V}(s) \leftarrow R(s) + \gamma \max_{a} \sum_{s' \in \mathcal{S}} \Pr[S' \mid s, a] \hat{V}(s')$$

Let B be the Bellman backup operator: $\mathbb{R} \mapsto \mathbb{R}$, which takes any arbitrary value function at some state $\hat{V}(S)$ and returns the its new value as updated by the value iteration

$$B\hat{V}(s) := R(s) + \gamma \max_{a} \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] \hat{V}(s')$$

We would like to first prove that the Bellman backup operator B is a contraction; in other words, for all $V_1(s)$ and $V_2(s)$

$$\max_{s \in \mathcal{S}} |BV_1(s) - BV_2(s)| \leq \gamma \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)|$$

The proof is as follows

$$
\begin{aligned}
& |BV_1(s) - BV_2(s)| \\
=& \gamma \left| \max_a \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] V_1(s') - \max_a \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] V_2(s') \right| \\
\leq& \gamma \max_a \left| \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] V_1(s') - \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] V_2(s') \right| \\
=& \gamma \max_a \left| \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] \left(V_1(s') - V_2(s') \right) \right| \\
=& \gamma \max_a \sum_{s' \in \mathcal{S}} \Pr[s' \mid s, a] \left| \left(V_1(s') - V_2(s') \right) \right| \\
=& \gamma \max_a \mathbb{E}_{s' \mid s, a} \left[\left| \left(V_1(s') - V_2(s') \right) \right| \right] \\
\leq& \gamma \max_{s' \in \mathcal{S}} \left| \left(V_1(s') - V_2(s') \right) \right| \\
=& \gamma \max_{s \in \mathcal{S}} \left| \left(V_1(s) - V_2(s) \right) \right|
\end{aligned}
$$

For the optimal value function $V^\star$, we have $BV^\star = V^\star$, therefore

$$\max_{s \in \mathcal{S}} |B\hat{V}(s) - V^\star(s)| \leq \gamma \max_{s \in \mathcal{S}} |\hat{V}(s) - V^\star(s)|$$

This means the updated values for the states are getting closer to the optimal values after each iteration. Thus, $\hat{V}(s)$ converges to $V^\star$ by value iteration algorithm.

2. Prove that if we look ahead k steps at the Q function, the value function converges to the same value function as the typical case when we look ahead 1 step at the Q function.

   Denote B_k to be the Bellman backup operator for looking k-step ahead for any value function V. Then

$$
\begin{aligned}
B_k V(s_1) := R(s_1) + \gamma \max_{a_1} \sum_{s_2 \in \mathcal{S}} \Pr[s_2 \mid s_1, a_1] \Bigg(R(s_2) \\
+ \gamma \max_{a_2} \sum_{s_3 \in \mathcal{S}} \Pr[s_3 \mid s_2, a_2] \Bigg(\cdots \Bigg(R(s_k) + \max_{a_k} \sum_{s_{k+1} \in \mathcal{S}} \Pr[s_{k+1} \mid s_k, a_k] V(s_{k+1}) \Bigg) \Bigg) \Bigg)
\end{aligned}
$$

   It is easy to show that for any two value functions V_1 and V_2

$$
\begin{aligned}
&|B_k V_1(s_1) - B_k V_2(s_1)| \\
&= \gamma^k \Bigg| \max_{a_1} \max_{a_2} \cdots \max_{a_k} \sum_{s_2 \in \mathcal{S}} \Pr[s_2 \mid s_1, a_1] \sum_{s_3 \in \mathcal{S}} \Pr[s_3 \mid s_2, a_2] \cdots \sum_{s_{k+1} \in \mathcal{S}} \Pr[s_{k+1} \mid s_k, a_k] V_1(s_{k+1}) \\
&\quad - \max_{a_1} \max_{a_2} \cdots \max_{a_k} \sum_{s_2 \in \mathcal{S}} \Pr[s_2 \mid s_1, a_1] \sum_{s_3 \in \mathcal{S}} \Pr[s_3 \mid s_2, a_2] \cdots \sum_{s_{k+1} \in \mathcal{S}} \Pr[s_{k+1} \mid s_k, a_k] V_2(s_{k+1}) \Bigg| \\
&= \gamma^k \max_{a_1} \max_{a_2} \cdots \max_{a_k} \sum_{s_2 \in \mathcal{S}} \Pr[s_2 \mid s_1, a_1] \sum_{s_3 \in \mathcal{S}} \Pr[s_3 \mid s_2, a_2] \cdots \sum_{s_{k+1} \in \mathcal{S}} \Pr[s_{k+1} \mid s_k, a_k] \\
&\qquad \Big| V_1(s_{k+1}) - V_2(s_{k+1}) \Big| \\
&\leq \gamma^k \max_{s_{k+1} \in \mathcal{S}} \Big| V_1(s_{k+1}) - V_2(s_{k+1}) \Big| \\
&= \gamma^k \max_{s \in \mathcal{S}} \Big| V_1(s) - V_2(s) \Big|
\end{aligned}
$$

Thus the operator for looking k-step ahead is also a contraction. Similarly, this algorithm will also converge and converge to the same value as looking one step ahead.