Assignment-1

Pradeep Mundlik (AI21BTECH11022)

September 25, 2023

Problem 1: Value Iteration

a. **<u>Solution</u>:** To prove that the Bellman optimality operator is a contraction under the max-norm, we need to show that it satisfies the Lipschitz condition with a Lipschitz coefficient l less than 1.

   $L : \vartheta \to \vartheta$ is contraction mapping if $\forall u, v \in \vartheta$

$$||L(v) - L(u)|| < l\,||v - u||$$

   Let's start by defining the Bellman optimality operator. Given a state s, it is defined as:

$$L[V](s) = \max_{a \in A} \left\{ \sum_{s' \in S} P(s'|s, a) \left(R(s, a, s') + \gamma V(s')\right) \right\}$$

   Now, We need to prove that there exists a $\gamma \in (0, 1]$ such that for any two value funtions V_1 and V_2:

$$||L[V_1](s) - L[V_2](s)||_\infty \leq l\,||V_1 - V_2||_\infty$$

   where, $||.||_\infty$ represents the max-norm, which is the maximum absolute value of the elements of a vector.

$$L[V_1](s) = \max_{a \in A} \left\{ \sum_{s' \in S} P(s'|s, a) \left(R(s, a, s') + \gamma V_1(s')\right) \right\}$$
$$L[V_2](s) = \max_{a \in A} \left\{ \sum_{s' \in S} P(s'|s, a) \left(R(s, a, s') + \gamma V_2(s')\right) \right\}$$

$$
\begin{aligned}
&||L[V_1] - L[V_2]||_\infty = \\
&= \left\| \max_{a\in A} \left\{ \sum_{s'\in S} P(s'|s,a) \left(R(s,a,s') + \gamma V_1(s') \right) \right\} - \max_{a\in A} \left\{ \sum_{s'\in S} P(s'|s,a) \left(R(s,a,s') + \gamma V_2(s') \right) \right\} \right\|_\infty \\
&\le \max_{a\in A} \left\| \sum_{s'\in S} P(s'|s,a) \left(R(s,a,s') + \gamma V_1(s') \right) - \sum_{s'\in S} P(s'|s,a) \left(R(s,a,s') + \gamma V_2(s') \right) \right\| \\
&\le \max_{a\in A} \left\| \sum_{s'\in S} P(s'|s,a) \left(\gamma \left(V_1(s') - V_2(s') \right) \right) \right\| \\
&\le \gamma \max_{a\in A} \left\| \sum_{s'\in S} P(s'|s,a) \left(V_1(s') - V_2(s') \right) \right\| \\
&\le \gamma \max_{a\in A} \left\| \sum_{s'\in S} P(s'|s,a) \right\| ||V_1 - V_2|| \ldots \text{(Cauchy Schewartz Inequality)} \\
&\le \gamma \, ||V_1 - V_2|| \ldots \left(\left\| \max_{a\in A} \sum_{s'\in S} P(s'|s,a) \right\| \le 1 \right)
\end{aligned}
$$

This proves that the Bellman optimality operator is a γ-contraction under the max-norm.

b. **<u>Solution</u>:** Let V_k represent the estimated value function after k iterations, and let V represent the true value function for the policy π. We can define the error at each state s as $|V_k(s) - V(s)|$.
Update Equation:

$$V_{k+1}(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + \gamma V_k(s')]$$

Bellman Expectation Equation:

$$V(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + \gamma V(s')]$$

$$
\begin{aligned}
|V_{k+1}(s) - V(s)| &= \left| \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + \gamma V_k(s')] - \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + \gamma V(s')] \right| \\
&= \left| \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \gamma (V_k(s') - V(s')) \right|
\end{aligned}
$$

Take max norm,

$$||V_{k+1} - V|| = \gamma \max_s \left| \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)(V_k(s') - V(s')) \right|$$

$$||V_{k+1} - V|| \le \gamma \left(\max_s \left| \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \right| \right) ||V_k - V||$$

Let, $M = \max_s \left|\sum_a \pi(a|s) \sum_{s'} p(s'|s, a)\right|$

$$\|V_{k+1} - V\| \le \gamma M \|V_k - V\|$$
$$\|V_{k+1} - V\| \le (\gamma M)^k \|V_0 - V\|$$

Let, $C = \max_s |V_0(s) - V(0)| = \|V_0 - V\|$

$$\|V_{k+1} - V\| \le (\gamma M)^k C$$

Here, $\gamma M < 1$, as k increases, the error decreases exponentially. This implies that the error is decreasing geometrically with each iteration.

c. **Solution:** We can use the triangle inequality to bound the difference between V_{k+1} and V as follows:

$$\|V_{k+1} - V_*\| \le \|V_{k+1} - V_k\| + \|V_k - V_*\|$$

Now,

$$V_k(s) = \max_a Q_k(s, a)$$
$$V_*(s) = \max_a Q_*(s, a)$$
$$\|V_k(s) - V_*(s)\| \le \gamma \max_a \|Q_k(s, a) - V_*(s)\|$$

The Q-value of taking action a in state s under policy π_k is the expected value of the immediate reward and the discounted value of the future rewards that the agent can achieve by taking action a in state s and then following policy π_k.

$$Q_k(s, a) = E[r_k + \gamma V_k(s')]$$
$$Q_*(s, a) = E[r_* + \gamma V_*(s')]$$
$$\|Q_k(s, a) - Q_*(s, a)\| = \|E[\gamma(V_k(s') - V_*(s'))]\|$$
$$\|Q_k(s, a) - Q_*(s, a)\| \le \gamma \max_s \|V_k(s') - V_*(s')\|$$

from above two inequalities,

$$\|V_k(s) - V_*(s)\| \le \gamma^2 \max_{s',a} \|V_k(s') - V_*(s')\|$$

Now we have, $\|V_{k+1}(s) - V_k(s)\| \le \epsilon$ and $\|V_{k+1} - V_*\| \le \|V_{k+1} - V_k\| + \|V_k - V_*\|$

$$\|V_{k+1}(s) - V_*(s)\| \le \epsilon + \gamma^2 \max_{s',a} \|V_k(s') - V_*(s')\|$$
$$\|V_{k+1}(s) - V_*(s)\| \le \epsilon(1 + \gamma^2 + \gamma^4 + \dots)$$
$$\|V_{k+1}(s) - V_*(s)\| \le \epsilon * \frac{1}{1 - \gamma^2}$$
$$\|V_{k+1}(s) - V_*(s)\| \le \frac{\epsilon}{1 - \gamma^2}$$

This bound shows how far the estimate V_{k+1} is from the optimal value function V_*.