CS 234 Winter 2023
Assignment 1

Nguyen Minh Phuc

August 2023

1 Reward Choices

Consider a Tabular MDP with $0 < \gamma < 1$ and no terminal states. The agent will act forever. The original optimal value function for this problem is V_1^* and the optimal policy is π_1^*.

(a) Now someone decides to add a small reward bonus c to all transitions in the MDP. This results in a new reward function $\hat{r}(s,a) = r(s,a) + c; \forall s, a$ wherer $r(s,a)$ is the original reward function.What is an expression for the new optimal value function? Can the optimal policy in this new setting change ? Why or why not ?
Answer: Based on Bellman optimality, the new optimal value function is:

$$\begin{aligned}
V_2^*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi^*}[G_t | S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi^*}[\sum_{k=0}^{\infty} \hat{R}_{t+k+1} | S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi^*}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + c \sum_{k=0}^{\infty} \gamma^k | S_t = s, A_t = a] \\
&= V_1^*(s) + \frac{c}{1-\gamma}
\end{aligned}$$

The new optimal Policy π_2^* in this new setting is:

$$\begin{aligned}
\pi_2^*(s) &= \arg\max_a V_2^*(s) \\
&= \arg\max_a V_1^*(s) + \frac{c}{1-\gamma} \\
&= \arg\max_a V_1^*(s) = \pi_1^*(s)
\end{aligned}$$

So the new policy in the new reward function is unchanged.

(b) Instead of adding a small reward bonus, someone decides to multiply all rewards by an arbitrary constant $c \in \mathbb{R}$ This results in a new reward function $\hat{r}(s,a) = c \times r(s,a), \forall s,a$, where $r(s,a)$ is the original reward function. Are there cases in which the new optimal policy is still π_1^* and the resulting value function can be expressed as a function of c and V_1^* (if yes, give the resulting expression and explain for what value(s) of c and why? If not, explain why not) ? Are there cases where the optimal policy would change (if yes, provide a value of c and a description of why the optimal policy would change. If not, explain why not) ? Is there a choice of c such that all policies are optimal ?

Answer: The resulting value function $V_3^*(s)$ in terms of c and V_1^* is:

$$\begin{aligned} V_3^*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \mathbb{E}_{\pi^*}[G_t | S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi^*}[\sum_{k=0}^{\infty} \hat{R}_{t+k+1} | S_t = s, A_t = a] \\ &= cV_1^*(s) \end{aligned}$$

- When $c > 0$, the new optimal policy is still π_1^*, cause we just scale up the optimal Value function $V_1^*(s)$.
- When $c < 0$, the optimal policy would change and the new optimal policy will choosing the action with the lowest value of $V_1^*(s)$.
- When $c = 0$, all policies are optimal, and the optimal value function is always 0.

(c) If the MDP instead has terminal states that can end an episode, does that change your answer to part (a) ? If yes, provide an example MDP where your answers to part (a) and this part would differ.

Answer: In episodic setting, the new value function can be expressed:

$$\begin{aligned} V^*(s) &= \max_a \mathbb{E}_{\pi^*}[\sum_{k=0}^{T} \gamma^k R_{t+k+1} + c \sum_{k=0}^{T} \gamma^k | S_t = s, A_t = a] \\ &= V_1^*(s) + c \sum_{k=0}^{T} \gamma^k \end{aligned}$$

The optimal policy in episodic setting will change, if $c > 0$, the new optimal policy will take longer time to reach the terminal state than original, cause the new optimal value function is increase when the time step increase and vice versa.

2 Bellman Residuals and performance bounds

Definitions: From lecture, we know that the Bellman backup operator B, defined below is a contraction with the fixed point as V^*, the optimal value

function of the MDP. The symbols have their usual meanings. γ is the discount factor and $0 \leq \gamma < 1$. In all parts, $||v||$ is the infinity norm of the vector.

$$(BV)(s) = \max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')]$$

We also saw the contraction operator B^π with the fixed point V^π, which is the Bellman backup operator for a particular policy given below:

$$(B^\pi V)(s) = \mathbb{E}_{a \sim \pi}[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')]$$

In class, we showed that $||BV - BV'|| \leq \gamma ||V - V'||$ for two arbitrarily function V and V'. For the rest of this question, you can also assume that $||B^\pi V - B^\pi V'|| \leq \gamma ||V - V'||$.

(a) Proof that the fixed point for B^π i.e $(B^\pi V)(s) = V(s), \forall s$ is unique.
Answer: Suppose there are at least 2 fixed point value function V and V' such that $\exists s \in S, V'(s) \neq V(s)$. We have:

$$||B^\pi V - B^\pi V'|| \leq \gamma ||V - V'||$$
$$||V - V'|| \leq \gamma ||V - V'||$$

Since $0 \leq \gamma < 1$, then $||V - V'|| = 0 \Rightarrow V = V'$. So the fixed point for B^π is unique.

We can recover a greedy policy π from an arbitrary value function V using the equation below:

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

(b) When π is the greedy policy, explain what is the relationship between B and B^π ?
Answer: Given the greedy policy $\pi(s)$ defined above, we can expressed it as:

$$\pi(a|s) = \begin{cases} a, & \text{if } a = \arg\max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \\ 0 & \text{otherwise} \end{cases}$$

The Bellman backup operator for a given greedy policy:

$$\begin{aligned}(B^\pi V)(s) &= \mathbb{E}_{a \sim \pi}[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \\ &= \sum_a \pi(a|s)[r + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \\ &= \max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \\ &= (BV)(s)\end{aligned}$$

When π is the greedy policy, B and B^π are identical.

Motivation: It is often helpful to know what the performance will be if we extract a greedy policy from a value function. In the rest of this problem, we will prove a bound on this performance.
Recall that a value function is a $|S|$-dimensional vector where $|S|$ is the number of states of the MDP. When we use the term V in these expression as an "arbitrary value function", we mean that V is an arbitrary $|S|$-dimensional vector which need not be aligned with the definition of the MDP at all. On the other hand, V^π is a value function that is achieved by some policy π in the MDP. For example, say the MDP has 2 states and only negative immediate rewards. $V = [1, 1]$ would be a valid choice for V even though this value function can never be achieved by any policy π, but we can never have a $V^\pi = [1, 1]$. This distinction between V and V^π is important for this question and more broadly in reinforcement learning.

Why do we care about setting V to vectors than can never be achieved in the MDP ? Sometimes algorithms, such as Deep Q-networks, return such vectors. In such situations we may still want to extract a greedy policy π from a provided V and bound the performance of the policy we extracted aka V^π.
Bellman Residuals: Define the Bellman residual to be $(BV - V)$ and the Bellman error magnitude to be $||BV - V||$. As we will see through the course, this Bellman residual is an important component of several popular RL algorithms such as the Deep Q-networks, referenced above. Intuitively, you can think of it as the difference between what the value function V specifies at a state and the one-step look-ahead along the seemingly best action at the state using the given value function V to evaluate all future states (the BV term).

(c) For what V does the Bellman error magnitude equal 0 ?
Answer: The Bellman error magnitude equal 0 if and only if:

$$\begin{aligned} V &= BV \\ &= \max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s')] \end{aligned}$$

Which satisfy the Bellman optimality condition, thus V is the optimal value function.

(d) Prove the following statements for an arbitrary value function V and any policy π.
Hint: Try leveraging the triangle inequality by inserting a zero term.

$$||V - V^\pi|| \leq \frac{||V - B^\pi V||}{1 - \gamma}$$
$$||V - V^*|| \leq \frac{||V - BV||}{1 - \gamma}$$

Answer: From triangle inequality:

$$\begin{aligned}
||V - V^\pi|| &= ||V - B^\pi V + B^\pi V - V^\pi|| \\
&\leq ||V - B^\pi V|| + ||B^\pi V - V^\pi|| \\
&= ||V - B^\pi V|| + ||B^\pi V - B^\pi V^\pi|| \\
&\leq ||V - B^\pi V|| + \gamma ||V - V^\pi||
\end{aligned}$$

From the above inequality:

$$\begin{aligned}
(1-\gamma)||V - V^\pi|| &\leq ||V - B^\pi V|| \\
\Leftrightarrow ||V - V^\pi|| &\leq \frac{||V - B^\pi V||}{1-\gamma}
\end{aligned}$$

For the second inequality, we can prove the same way:

$$\begin{aligned}
||V - V^*|| &= |V - BV + BV - V^*|| \\
&\leq ||V - BV|| + ||BV - V^*|| \\
&= ||V - BV|| + ||BV - BV^*|| \\
&\leq ||V - BV|| + \gamma ||V - V^*||
\end{aligned}$$

Thus,

$$\begin{aligned}
(1-\gamma)||V - V^*|| &\leq ||V - BV|| \\
\Leftrightarrow ||V - V^*|| &\leq \frac{||V - BV||}{1-\gamma}
\end{aligned}$$

(e) Let V be an arbitrary value function and π be the greedy policy extracted from V . Let $\epsilon = ||BV - V||$ be the Bellman error magnitude for V. Prove the following for any state s.
Hint: Try to use the results from part (d).

$$V^\pi(s) \geq V^*(s) - \frac{2\epsilon}{1-\gamma}$$

Answer:

$$\begin{aligned}
||V^\pi - V^*|| &= ||V^\pi + V - V - V^*|| \\
&\leq ||V^\pi - V|| + ||V^* - V^\pi|| \\
&\leq \frac{||V - B^\pi V|| + ||V - BV||}{1-\gamma} \text{ (from part (d))} \\
&= \frac{2||V - BV||}{1-\gamma} \text{ (π is the greedy policy)} \\
&= \frac{2\epsilon}{1-\gamma}
\end{aligned}$$

Since $\forall \pi, V^{\pi}(s) \leq V^{*}(s)$:

$$\begin{aligned} ||V^{\pi} - V^{*}|| &= \max_{s} |V^{\pi}(s) - V^{*}(s)| \\ &= \max_{s} V^{*}(s) - V^{\pi}(s) \\ &\leq \frac{2\epsilon}{1-\gamma} \end{aligned}$$

Thus, $\forall s \in S, V^{\pi}(s) \geq V^{*}(s) - \frac{2\epsilon}{1-\gamma}$

A little bit more notation: define $V \leq V'$ if $\forall s, V(s) \leq V'(s)$. What if our algorithm returns a V that satisfies $V^{*} \leq V$, i.e., it returns a value function that is better than the optimal value function of the MDP. Once again, remember that V can be any vector, not necessarily achievable in the MDP but we would still like to bound the performance of V^{π} where π is extracted from said V . We will show that if this condition is met, then we can achieve an even tighter bound on policy performance.

(f) Using the same notation and setup as part (e), if $V \leq V^{*}$, show the following holds for any state s.
Hint: Recall that $\forall \pi, V^{\pi} \leq V^{*}$. (why ?)

$$V^{\pi} \geq V^{*} - \frac{\epsilon}{1-\gamma}$$

Answer: Still working =))

(g) It's not easy to show that the condition $V^{*} \leq V$ holds because we often don't know V^{*} of the MDP. Show that if $BV \leq V$ then $V^{*} \leq V$. Note that this sufficient condition is much easier to check and does not require knowledge of V.
Answer: We have $BV \leq V$, Assuming from some $k \geq 1$, the inequality $B^{k}V \leq V$ holds, then:

$$\begin{aligned} (B^{k+1}V)(s) &= \max_{a}[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)(B^{k}V)(s')] \\ &\leq \max_{a}[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \\ &= BV \leq V \end{aligned}$$

From induction, $\forall k \geq 1, B^{k}V \leq V$. (1)
For some k,

$$\begin{aligned} ||B^{k}V - V^{*}|| &= ||B^{k}V - BV^{*}|| \\ ||B^{k}V - BV^{*}|| &\leq \gamma ||B^{k-1}V - V^{*}|| \\ &\leq \gamma^{2} ||B^{k-2}V - V^{*}|| \\ &\cdots \\ &\leq \gamma^{k} ||V - V^{*}|| \end{aligned}$$

Let $\lim_{k\to\infty}$ and we have $||B^k V - V^*|| \to 0$. Thus, $\lim_{k\to\infty} B^k V = V^*$ (2).
From (1) and (2), we can conclude $V^* \leq V$.