Homework -Interpretability

Shaun Toh 1002012

July 15, 2019

1 a proof

Bellman equation for a given deterministic policy

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s')$$

s' is the next state after having taken an action A.
$V^{\pi}(s)$ is the value function and hence takes the shape of $(1 \times s_n)$ where s_n is the total number of states that are possible.
Since $R(s, \pi(s))$ is added together with the second term, $R(s, \pi(s))$ must hence be of the exact same shape as V^{π}, i.e $(1 \times s_n)$
gamma is a constant, so we ignore its shape as it is a scalar.
The summation can also be safely ignored as it should not affect the output dimension of that term in any way.

$P(s'|s, \pi(s)) \times V^{\pi}(s')$ consists of 2 different terms.
$P(s'|s, \pi(s))$ is defined as the probability of a certain new state being arrived in given a current state and the action. Since this is applied for each of the various states, to each of the states, it is a $(1 \times s_n)$ shaped matrix.
$V^{\pi}(s')$ is the reward values at new states, and is $(1 \times s_n)$, since only one state is considered at a time.

To summarise,

1. V^{π} -$(1 \times s_n)$
2. $R(s, \pi(s))$ - $(1 \times s_n)$
3. γ is a scalar.
4. $P(s'|s, \pi(s))$ is a $(1 \times s_n)$ shaped vector.
5. $V^{\pi}(s')$ is a scalar.

1.1 Condition

In order for this to be solvable, we need $0< \gamma <1$ and one more condition below. We know that $V^{\pi}(s') = V^{\pi}(s)$ for one particular state.

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s')$$

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s' \neq s} [P(s'|s, \pi(s)) V^{\pi}(s')] + \gamma \times P(s|s, \pi(s)) V^{\pi}(s')$$

$$V^{\pi}(s) + \gamma \times P(s'|s, \pi(s)) V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s' \neq s} [P(s|s, \pi(s)) V^{\pi}(s')]$$

$$V^{\pi}(s)[1 - \gamma \times P(s|s, \pi(s))] = R(s, \pi(s)) + \gamma \sum_{s' \neq s} [P(s'|s, \pi(s)) V^{\pi}(s')]$$

$$V^{\pi}(s) = \frac{R(s, \pi(s)) + \gamma \sum_{s' \neq s} [P(s'|s, \pi(s)) V^{\pi}(s')]}{[1 - \gamma \times P(s|s, \pi(s))]}$$

$[1 - \gamma \times P(s'|s, \pi(s))]$ must be invertible.

2 MDP per hand

Since

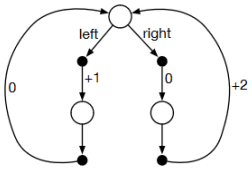
$$V_{k+1}(s) = \max_{a} \sum_{s'} P(s'|s, a) R(s, a, s') + \gamma \sum_{s'} P(s'|s, a) V_k(s')$$

We first note that s' in this case is actually the next state in this case.
There are hence 27 possibilities for the policy valuation since it is based off what states were landed on for this particular case. Please see the attached second latex document for code and outputs for this question. This question was done in collaboration with Chang Jun Qing, 1002088.

3 one small exercise

Exercise 3.22 Consider the continuing MDP shown on to the right. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? □

Exercise 3.23 Give an equation for v_* in terms of q_*. □



The answer is contingent on $P(s'|s, a)$ being =1 for all. If it varies, the answer is indeterminate.
If the model is uncompressed,

γ measures the weightage of your choice
if γ =0, always taking left is the optimal policy.
if γ =0.9, always taking right is the optimal policy.
if γ =0.5, there is no preference between both sides
If the model is compressed so the instant reward is +1 or +2 instantly, the answer will always be to take the right side, even if the probabilities are skewed towards ending up on the left. (unless picking left and right both give a 100% probability of going left, which gives the same expected value of 1, so neither is preferable.)