CS 5100 – PS 6

Yijing Xiao

Problem 1: Long term reward

For what values of the discount factor γ should the agent choose Up and for which Down?

Additive utility: $U([r_0, r_1, r_2, \cdots]) \ = \ r_0 \ + \ r_1 \ + \ r_2 \ + \ \cdots$

Discounted utility: $U([r_0, r_1, r_2, \cdots]) \ = \ r_0 \ + \ \gamma r_1 \ + \ \gamma^2 r_2 \ + \ \cdots$

Therefore, more generally: For $0 \ \leq \ \gamma \ \leq \ 1$,

$$U([s_0, s_1, s_2, \cdots]) \ = \ R(s_0) \ + \ \gamma R(s_1) \ + \ \gamma^2 R(s_2) \ + \ \cdots \ = \ \sum_{t=0}^{\infty} \gamma^t R(s_t)$$

If the agent chooses Up, the $R(s_0)$ = 0, the $R(s_1)$ = + 50 and we will have:

$$50\gamma \ + \ \sum_{t=2}^{101} \gamma^t \ \times \ (-1)$$

If the agent chooses Down, the $R(s_0)$ = 0, the $R(s_1)$ = - 50 and we will have:

$$-50\gamma \ + \ \sum_{t=2}^{101} \gamma^t \ \times \ 1$$

We could use the sum of geometric series formula to express them:

$$50\gamma \ - \ \frac{\gamma^2(1-\gamma^{100})}{1-\gamma} \ and \ - 50\gamma \ + \ \frac{\gamma^2(1-\gamma^{100})}{1-\gamma}$$

If we solve them numerically by using Wolfram Alpha, we could get $\gamma \ \approx \ 0.984506306285 \cdots$. As long as γ which we pick bigger than this value, the agent would be going Down; Otherwise, the agent would be going Up.

Problem 2: Bellman Equation

a) **Explain, in words, what the general version of the Bellman equation means. Additionally, show that it reduces to the simpler version when using deterministic policies π(s) and state-only rewards R(s).**

Bellman equation for V(π) expresses a relationship between the value of a state and the values of its successor states, for any policy π and any state s, the consistency conditions hold between the value of s and the value of its possible successor states.

In the case of given state s and action a, the subsequent state s' and state-only reward R(s) are determined (using deterministic policies π(s), so the transfer probability T in the equation can be ignored and can be simplified as:

$$\mathrm{V}^{\pi}(\mathrm{s}) = \sum_{\mathrm{a}} \pi(\mathrm{a}|\mathrm{s}) \left[\mathrm{R}(\mathrm{s}) + \gamma \mathrm{V}_{\pi}(\mathrm{s}')\right]$$

b) **Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, −0.4,+0.7. The discount factor is γ = 0.9.**

Since the value function for the equiprobable random policy, we have $\pi(\cdot\,|s) = 0.25$ for all 4 actions and 0 reward for other actions, we have:

0.25 * (0 + 0.9 * 2.3) + 0.25 * (0 + 0.9 * 0.4) + 0.25 * (0 + 0.9 * (-0.4)) + 0.25 * (0 + 0.9 * 0.7) = 0.25 * 0.9 * (2.3 + 0.7) = 0.675 ≈ 0.7

c) **Similar to the previous part, show numerically that the Bellman equation holds for the center state, valued at +17.8, with respect to its four neighboring states, for the optimal policy $\pi*$ shown in the figure on the next page (right). Also show numerically that the Bellman optimality equation holds for the same center state, valued at +17.8, and verify that the optimal actions at that state are indeed as shown.**

Bellman equation:

0.5 * (0 + 0.9 * 19.8) + 0.5 * (0 + 0.9 * 19.8) = 17.82 ≈ 17.8

Bellman optimality equation:

Max((0 + 0.9 * 19.8), (0 + 0.9 * 19.8), (0 + 0.9 * 16), (0 + 0.9 * 16)) = 0.9 * 19.8 = 17.82 ≈ 17.8