Venkateswar Reddy M.
CTO, Brillium Technologies
"…dare to dream; care to win…"

Date: **4-May-24**

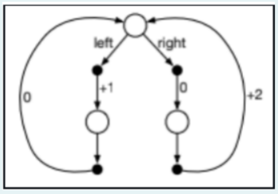
EE932 Assignment-2 Solution

---

eMasters in Communication Systems, IITK
EE932: Introduction to Reinforcement Learning
Instructor: Prof. Subrahmanya Swamy Peruru
Student Name: Venkateswar Reddy Melachervu
Roll No: 23156022

Question 8:

Consider the continuing MDP shown. The only decision to be made is in the top state, where two actions are available, left and right. In the other two states, only one action is available, and hence there is nothing to decide. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policies are optimal for the three cases given below? Show your calculations and upload an image. Case 1: $\gamma = 0$, Case 2: $\gamma = 0.9$, Case 3: $\gamma = 0.5$.



Solution:
Total rewards (Returns) for Continuing Task

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

Case $\boldsymbol{\gamma = 0}$:
For this case, the delayed/future rewards do not count, so the total reward from top state
$\therefore G_{t-\pi_{left}} = 1$ and $G_{t=\pi_{right}} = 0 \Rightarrow \boldsymbol{\pi_{left}}$ **is optimal**

Case $\boldsymbol{\gamma = 0.9}$
For this case, delayed/future regards would have substantial weightage as $\gamma = 0.9$
The total rewards from the top state,
$G_{t-\pi_{left}} = 1 + 0 * 0.9 + 1 * 0.9^2 + 0 * 0.9^3 + 1 * 0.9^4 + \cdots$
$= 1 + 0.9^2 + 0.9^4 + \cdots$
$= \frac{1}{(1-0.9^2)} = 5.2632$

$G_{t-\pi_{right}} = 0 + 2 * 0.9 + 0 * 0.9^2 + 2 * 0.9^3 + 0 * 0.9^4 + 2 * 0.9^5 \ldots$
$= 2 * 0.9 + 2 * 0.9^3 + 2 * 0.9^5 + \cdots$
$= 2 * 0.9(1 + 0.9^2 + 0.9^4 + \cdots)$
$= 2 * 0.9 * \dfrac{1}{(1 - 0.9^2)} = 9.4737$
$\therefore \boldsymbol{\pi_{right}}$ **is optimal**

Case $\boldsymbol{\gamma = 0.5}$
For this case, delayed/future regards would have substantial weightage as $\gamma = 0.5$
The total rewards from the top state,
$G_{t-\pi_{left}} = 1 + 0 * 0.5 + 1 * 0.5^2 + 0 * 0.5^3 + 1 * 0.5^4 + \cdots$
$= 1 + 0.5^2 + 0.5^4 + \cdots$
$= \frac{1}{(1-0.5^2)} = 1.3333$

Venkateswar Reddy M.
CTO, Brillium Technologies
"…dare to dream; care to win…"

$$G_{t-\pi_{right}} = 0 + 2 * 0.5 + 0 * 0.5^2 + 2 * 0.5^3 + 0 * 0.5^4 + 2 * 0.5^5 \ldots$$
$$= 2 * 0.5 + 2 * 0.5^3 + 2 * 0.5^5 + \cdots$$
$$= 2 * 0.5(1 + 0.5^2 + 0.5^4 + \cdots)$$
$$= 2 * 0.5 * \frac{1}{(1 - 0.5^2)} = 1 * \frac{1}{(1 - 0.5^2)} = 1.3333$$

∴ **Both π_{left} and π_{right} policies are optimal**

-- End of the Document --

C-501, Salarpuria Serenity, 5th Main, Sector 7, HSR Layout, Bengaluru 560102 KA India
Mobile: +91 97012 22130, Email: vmelachervu@gmail.com, vmela23@iitk.ac.in, Website: www.linkedin.com/in/vmelachervu