Quiz_7

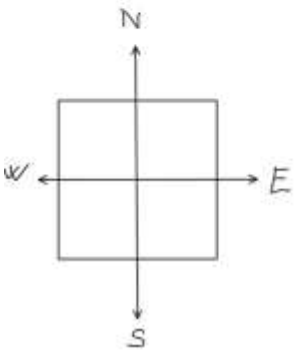
Policy Evaluation A

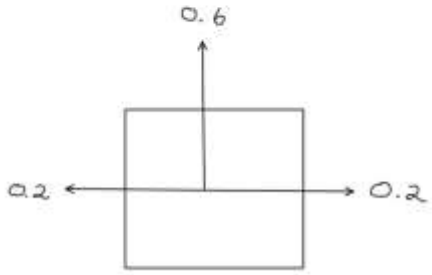
Policy evaluation quizz

-150	160	-40
-150		-40
-150		-40

Numbers represent rewards. Red states are terminal.

Agent will follow the order with probability of 0.6 and with probability of 0.2 will go in ±90° direction. Discount factor is 0.8. For non-terminal states, find values $V(s)$ for policy: " jeď na západ (go west) " . The reward for the nonterminal states is 0, i.e. $r(s) = 0$.





$r(s) = 0$

$\gamma = 0.8$

Value iteration:

$$v^*(s) = \max_{a \in A(s)} \sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma v^*(s')\right] \longrightarrow$$

If we only have $r(s)$ we get

$$V(s) = r(s) + \gamma \max_a \sum_{s'} p(s'|s,a)\, V(s')$$

-150	160	-40
-150	1	-40
-150	2	-40

I start by setting the two values to 0

- V(A)
- V(B)

Here the agent will go west, meaning that the goal is not to maximize V. This in turn means that we can further simplify the equation into:

$$V(s) = r(s) + \gamma \cdot P(s'|s,a)\, V(s')$$

a: Chosen by π (policy = go west)

$$V_1(A) = 0 + 0.8 \cdot [0.6 \cdot (-150) + 0.2 \cdot 160 + 0.2 \cdot V_0(B)]$$
$$= -46.4$$

$$V_1(B) = 0 + 0.8\,[0.6\,(-150) + 0.2 \cdot V_0(A) + 0.2 \cdot V_0(B)]$$
$$= -72$$

$$V_2(A) = 0 + 0.8 \cdot [0.6 \cdot (-150) + 0.2 \cdot 160 + 0.2 \cdot V_1(B)]$$
$$= -57.92$$

$$V_2(B) = 0 + 0.8 \cdot [0.6\,(-150) + 0.2 \cdot V_1(A) + 0.2 \cdot V_1(B)]$$
$$= -90.944$$

$$V_3(A) = 0.8\,[0.6 \cdot (-150) + 0.2 \cdot 160 + 0.2 \cdot V_2(B)]$$
$$= -46.4 + 0.8 \cdot V_2(B) \cdot 0.2$$
$$= -60.93104$$

$$V_3(B) = 0.8 \cdot [0.6\,(-150) + 0.2\,(V_0(A) + V_2(B))]$$
$$= -72 + 0.8 \cdot 0.2 \cdot (V_2(A) + V_2(B))$$
$$= -95.81824$$

$$V_4(A) = -46.4 + 0.16\,(V_3(B))$$
$$= -61.7309$$

$$V_4(B) = -72 + 0.16 \cdot (V_3(A) + V_3(B))$$
$$= -97.08319$$

$V_5(A) = -61.9333$

$V_5(B) = -97.4162$

$V_6(A) = -61.9856$

$V_6(B) = -97.4950$

$V_7(A) = -61.9952$

$V_7(B) = -97.5169$

So the values converge to about

$V(A) = -62.0$

$V(B) = -97.5$

if we always choose west.