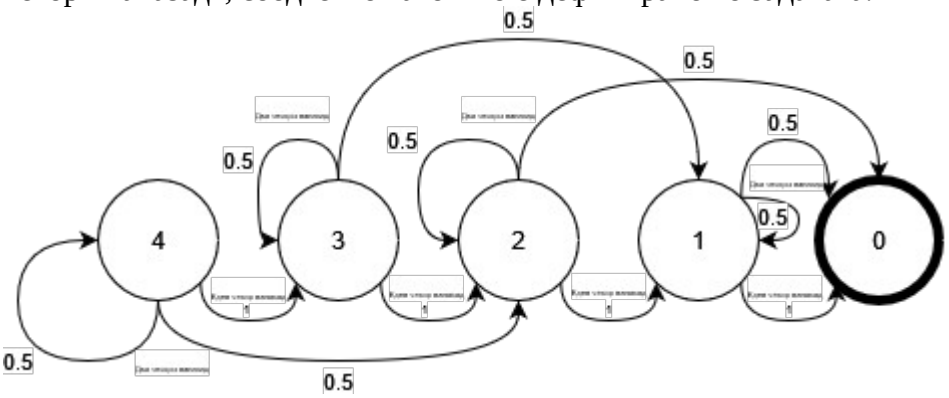
# Домашна задача 1

Индекс: 211017

## Задача 1: Чекор или два наназад

a) Не се чита најдобро на фотографијата, но текстовите се „Еден чекор наназад“ и „Два чекори наназад“, соодветно како што е дефинирано во задачата.




b) $V^*(2)$. a={„Еден чекор наназад“, „Два чекори наназад“}

| States | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| $V_0$ | 0 | 0 | 0 | 0 | 0 |
| $V_1$ | 0 | 0 | 2 | 1 | 0 |
| $V_2$ | 2.25 | 2.25 | 2.5 | 1 | 0 |

a) $V_0 = 0$ За сите состојби бидејќи при k=0 агентот нема повеќе чекори, крај на играта

b) k=1:
Во состојбите 4 и 3 пак вредноста ќе биде 0 бидејќи од тие состојби невозможно е да се стигне до терминална состојба ако имаме само уште еден timestep до крај на играта.

За состојба 2:

$$V_1(2) = max \sum T(2,a,s')[R(2,a,s') + \gamma V_0(s')] =$$
$$max[1 + \gamma * V_0(1), 0.5 * (4 + \gamma * V_0(0)) + 0.5 * (0 + \gamma * V_0(2))] =$$
$$max[1 + 0.5 * 0, 0.5 * (4 + 0.5 * 0) + 0] = max[1, 2 + 0] = 2$$

За состојба 1:

$$V_1(1) = max \sum T(1,a,s')[R(1,a,s') + \gamma V_0(s')] =$$
$$max[1 + \gamma * V_0(0), 0.5 * (1 + \gamma * V_0(0)) + 0.5 * (0 + \gamma * V_0(1))] =$$
$$max[1 + 0.5 * 0, 0.5 * (1 + 0.5 * 0) + 0] = max[1, 0.5 + 0] = 1$$

Состојба 0 има вредност 0 бидејќи таа е терминална и од неа нема акции кои агентот може да ги преземе и притоа да добие некаква награда за нив. (пример немаме награда за завршување со играта)

c)

$$V_2(4)=max\sum T(4,a,s')[R(4,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(3),0.5*(4+\gamma*V_1(2))+0.5*(0+\gamma*V_1(4))]=$$
$$max[1+0.5*0,0.5*(4+0.5*1)+0.5*(0+0.5*0)]=max[1,2.25+0]=2.25$$

$$V_2(3)=max\sum T(3,a,s')[R(3,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(2),0.5*(4+\gamma*V_1(1))+0.5*(0+\gamma*V_1(3))]=$$
$$max[1+0.5*1,0.5*(4+0.5*1)+0.5*(0+0.5*0)]=max[1.5,2.25+0]=2.25$$

$$V_2(2)=max\sum T(2,a,s')[R(2,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(1),0.5*(4+\gamma*V_1(0))+0.5*(0+\gamma*V_1(2))]=$$
$$max[1+0.5*1,0.5*(4+0.5*0)+0.5*(0+0.5*2)]=max[1.5,2+0.5]=2.5$$

$$V_2(1)=max\sum T(1,a,s')[R(1,a,s')+\gamma V_2(s')]=$$
$$max[1+\gamma*V_0(0),0.5*(1+\gamma*V_0(0))+0.5*(0+\gamma*V_1(1))]=$$
$$max[1+0.5*0,0.5*(1+0.5*0)+0.5*(0+0.5*1)]=max[1,0.5+0.25]=1$$

**Согласно барањата од задачата, за состојба 2, оптималната вредност $V^*(2)=2$. Оптималната политика се добива така што со argmax на местото од max во претходните пресметки за $V_2$ ја добиваме акцијата која треба да се превземе во секоја состојба:**

| State | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| $\pi^*$ | Два чекори назад | Два чекори назад | Два чекори назад | Еден черкор назад | Крај? |

**Оптималната политика која агентот ќе ја преземе во $V^*(2)$ е Два чекори наназад**

c) Овој дел ќе го пресметам во однос на табелата со итерации добиена на претходното барање, бидејќи проверуваме дали сме ги добиле оптималните вредности рачно, би било многу долг процес(многу итерации). Односно, ќе сметам дека алгоритамот конвергирал по пресметаните две итерации.

$$Q^*(4,чекор\ наназад)=\sum T(4,чекор\ наназад,s')[R(4,чекор\ наназад,s')+\gamma V_2(s')]=$$
$$1+\gamma*V_2(3)=1+0.5*2.25=2.125$$

d)

| States | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| $V_0$ | 0 | 0 | 0 | 0 | 0 |
| $V_1$ | 2 | 2 | 2 | 4.5 | 16 |
| $V_2$ | 3 | 3.625 | 7 | 9 | 17 |

e) **Забелешка: При премин од состојба 1 во состојба 4, го сметам тоа како квадратна разлика од 9, односно $(1-4)^2=3^2=9$, со што наградата е 9, и аналогно за 0 при премин во состојби 4 и 3.**

Образложение:

$$V_1(4)=max\sum T(4,a,s')[R(4,a,s')+\gamma V_0(s')]=$$
$$max[1+\gamma*V_0(3),0.5*(4+\gamma*V_0(2))+0.5*(0+\gamma*V_0(4))]=$$
$$max[1+0.5*0,0.5*(4+0.5*0)+0]=max[1,2+0]=2$$
$$V_1(3)=max\sum T(3,a,s')[R(3,a,s')+\gamma V_0(s')]=$$
$$max[1+\gamma*V_0(3),0.5*(4+\gamma*V_0(1))+0.5*(0+\gamma*V_0(3))]=$$
$$max[1+0.5*0,0.5*(4+0.5*0)+0]=max[1,2+0]=2$$

$V_1(1)=2$, бидејќи како што е наведено во алтернативната форма на задачата, два чекори наназад од состојба 1 ќе го врати агентот во состојба 4.

Во оваа алтернативна форма на задачата, во k=1:

$$V_1(1)=max\sum T(1,a,s')[R(1,a,s')+\gamma V_0(s')]=$$
$$max[1+\gamma*V_0(0),0.5*(9+\gamma*V_0(4))+0.5*(0+\gamma*V_0(1))]=$$
$$max[1+0.5*0,0.5*(9+0.5*0)+0]=max[1,4.5+0]=4.5$$

$$V_1(0)=max\sum T(0,a,s')[R(0,a,s')+\gamma V_0(s')]=$$
$$max[16+\gamma*V_0(4),0.5*(9+\gamma*V_0(3))+0.5*(0+\gamma*V_0(0))]=$$
$$max[16+0.5*0,0.5*(9+0.5*0)+0]=max[16,9+0]=16$$
$$V_2(4)=max\sum T(4,a,s')[R(4,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(3),0.5*(4+\gamma*V_1(2))+0.5*(0+\gamma*V_1(4))]=$$
$$max[1+0.5*2,0.5*(4+0.5*2)+0.5*(0+0.5*2)]=max[2,2.5+0.5]=3$$
$$V_2(3)=max\sum T(3,a,s')[R(3,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(2),0.5*(4+\gamma*V_1(1))+0.5*(0+\gamma*V_1(3))]=$$
$$max[1+0.5*2,0.5*(4+0.5*4.5)+0.5*(0+0.5*2)]=max[2,3.125+0.5]=3.625$$

$$V_2(2)=max\sum T(2,a,s')[R(2,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(1),0.5*(4+\gamma*V_1(0))+0.5*(0+\gamma*V_1(2))]=$$
$$max[1+0.5*4.5,0.5*(4+0.5*16)+0.5*(0+0.5*2)]=max[3.25,6+1]=7$$

$$V_2(1)=max\sum T(1,a,s')[R(1,a,s')+\gamma V_1(s')]=$$
$$max[1+\gamma*V_1(0),0.5*(4+\gamma*V_1(4))+0.5*(0+\gamma*V_1(1))]=$$
$$max[1+0.5*16,0.5*(9+0.5*2)+0.5*(0+0.5*4.5)]=max[9,5+1.125]=9$$

$$V_2(0)=max\sum T(0,a,s')[R(0,a,s')+\gamma V_2(s')]=$$
$$max[1+\gamma*V_1(4),0.5*(4+\gamma*V_1(3))+0.5*(0+\gamma*V_1(0))]=$$
$$max[16+0.5*2,0.5*(9+0.5*2)+0.5*(16+0.5*2)]=max[17,5+8.5]=17$$

Оптималната политика е:

| State | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| $\pi^*$ | Два чекори назад | Два чекори назад | Два чекори назад | Еден чекор назад | Еден чекор назад |

**Забелешка: Оптималната политика може да се промени доколку на алгоритамот му се дозволат повеќе итерации од ова. Во моменталната состојба, оваа погоре е оптималната политика.**

Оптималната политика е добиена со тоа што за секоја состојба се заменети соодветните вредности во формулата за пресметување на политика:
$\pi^*=argmax\sum T(s,n,s')*(R(s,n,s')+\gamma V^*(s'))$. За да скратам на препишување на истата формула како што правев во претходните пресметки, тоа е всушност истата

пресметка од погоре, но наместо користење на функција max, се користи argmax функцијата која ќе го врати аргументот кој ја дава најголемата вредност, наместо самата најголема вредност. На пример:

$$\pi^*(4) = argmax \sum T(4,a,s')[R(4,a,s') + \gamma V_1(s')] =$$

$$argmax[1 + \gamma * V_1(3), 0.5 * (4 + \gamma * V_1(2)) + 0.5 * (0 + \gamma * V_1(4))] =$$

$$argmax[1 + 0.5 * 2, 0.5 * (4 + 0.5 * 2) + 0.5 * (0 + 0.5 * 2)] = argmax[2, 2.5 + 0.5] = \textit{Два чекори назад}$$

# Задача 2: Save the Earth!

a) actions={do nothing(n), environmental policy(p), environmental disaster(d), irresponsible behavior(r)}
$\gamma = 1$

| State | Status Quo (SQ) | Global Warming Alarm (GW) | Still a Chance (SC) | Doomsday (DD) |
|---|---|---|---|---|
| $V_0$ | 0 | 0 | 0 | 0 |
| $V_1$ | 0.36 | 0.2 | 1.92 | -10 |
| $V_2$ | 7.2304 | 3.232 | 2.592 | -20 |
| $V_3$ | 9.31504 | 4.848 | 8.84064 | -30 |

За овој дел од проблемот повторно како и во првата задача ќе ја користам формулата за Value фукцијата (односно Белмановата равенка):
$V_{k+1}(s) = max \sum T(s,a,s')[R(s,a,s') + V_k(s')]$

Во k=1

$$V_1(SQ) = max[\sum T(SQ,n,s') * (R(SQ,n,s') + \gamma V_0(s')), \sum T(SQ,p,s') * (R(SQ,p,s') + \gamma V_0(s')), \sum T(SQ,r,s') * (R(SQ,r,s') + \gamma V_0(s'))]$$
$$= max[1*(-0.1+1*0), 0.6*(-0.1+1*0)+0.4*(10+1*0), 0.1*(-0.1+1*0)+0.9*(-10+1*0)] =$$
$$= max[-0.1, -0.06+0.4, -0.01-0.9] = max[-0.1, 0.36, -0.91] = 0.36$$

(За сите понатамошни пресметки постапката е слична, па ќе ги пишувам само резултатите.)
$V_1(GW) = 0.2$

$V_1(SC) = 1.92$

$V_1(DD) = -10$

Во k=2:
$V_2(SQ) = 7.2304$

$V_2(GW) = 3.232$

$V_2(SC) = 2.592$

$V_2(DD) = -20$

Во k=3:
$V_3(SQ) = 9.31504$

$V_3(GW) = 4.848$

$V_3(SC) = 8.84064$

$V_3(DD) = -30$

b) Доколку пресметките ми беа добри(пишував рачно, на тетратка), изгледа дека вредностите на состојбите **НЕ** конвергираат после првите 3 чекори, туку сè уште значително се зголемуваат.

c) Користејќи ја формулата:
$\pi^* = argmax \sum T(s,n,s')*(R(s,n,s')+\gamma V^*(s'))$
Доколку добиените вредности за последната итерација на вредноста за Status Quo состојбата ги замениме во формулата погоре, (се добива $max[4.104, 7.3352]$), акцијата која доведува до поголема вредност е **Environmental policy.**

d)

| State | Status Quo\| (SQ) | Still a Chance (SC) |
|---|---|---|
| $\pi_i$ | Do nothing | Do Nothing |
| $V^{\pi i}$ | -0.1 | 1.92 |
| $\pi_{i+1}$ | Environmental policy | Do Nothing |

Ажурираната оптимална стратегија ја добив на ист начин како што добив за претходното барање.