ДОМАШНА ЗАДАЧА 1 | **АГЕНТНО-БАЗИРАНИ СИСТЕМИ** | *МАРКОВИ ПРОЦЕСИ НА ОДЛУЧУВАЊЕ*

1. Нова верзија на Phone или не?

Замислете дека користите Маркови процеси на одлучување (MDPs) при решавање на проблемот на компанијата Apple, која секои 6 месеци треба да донесе одлука дали да се појави на пазарот со нова верзија на iPhone или да причека. Проблемот може да се претстави со 3 состојби **P, SS** и **N**, кои одговараат на позитивно, така-така (неутрално) или негативно расположение/сентимент кон компанијата. Дозволени акции се **Нова верзија** и **Причекај.** Наградата, која се добива за било која акција која го води системот во состојбата на позитивен сентимент кон компанијата **P** е +2 (наградата се добива и за премин од **P** во **P**). Наградата за премин во неутралната состојбата **SS** е 0, додека состојбата **N** се смета за непожелна и за премин во неа се добива негативна награда од -1. Факторот на намалување е γ = 1.

Веројатностите за премин од една во друга состојба се прикажани во следната табела.

s	***a***	***T(s, a, s')***		
		s' = P	***s' = SS***	***s'= N***
P	**Нова верзија**	0.1	0.9	0
	Причекај	0.2	0.8	0
SS	**Нова верзија**	0.1	0.9	0
	Причекај	0	0.3	0.7
N	**Нова верзија**	0.9	0	0.1
	Причекај	0	0.5	0.5

За дадениот проблем чие решавање треба да го дефинирате како MDP, потребно е да одговорите на следните прашања/задачи:

(a) Да се пресметаат вредностите на состојбите, *V(s, a, s') & Q (s, a, s')* за првите 3 чекори од постепеното рекурзивно пресметување на состојбите (првите 3 итерации) и пополни табелата. Образложете ги пресметките преку формулите кои сте ги користеле при пополнување на табелата.

	P	***SS***	***N***
$V^*_0(s)$	0	0	0
$Q^*_1(s,$ Нова верзија$)$	0.2	0.2	1.7
$Q^*_1(s,$ Причекај$)$	0.4	-0.7	-0.5
$V^*_1(s)$	0.4	0.2	1.7
$Q^*_2(s,$ Нова верзија$)$	0.42	0.42	2.23
$Q^*_2(s,$ Причекај$)$	0.64	0.55	0.45
$V^*_2(s)$	0.64	0.55	2.23

За $V^*_0(s)$ каде s = {P,SS,N} под претпоставка ги иницијализираме сита да бидат еднакви на 0.

За s = P следува:

$Q^*_1(s,$ Нова верзија$)$ = T(s,Нова верзија, P)[R(s,Нова верзија,P)+ $\gamma V_0(P)$] + T(s,Нова верзија, SS)[R(s,Нова верзија,SS)+ $\gamma V_0(SS)$] + T(s,Нова верзија, N)[R(s,Нова верзија,N)+ $\gamma V_0(N)$] = 0.1[2+1*0]+0.9[0+1*0]+0[-1+1*0] = 0.1*2 = 0.2

$Q^*_1(s,$ Причекај$)$ = T(s,Причекај, P)[R(s,Причекај,P)+ $\gamma V_0(P)$] + T(s,Причекај, SS)[R(s,Причекај,SS)+ $\gamma V_0(SS)$] + T(s,Причекај, N)[R(s,Причекај,N)+ $\gamma V_0(N)$] = 0.2[2+1*0] + 0.8[0+1*0]+0[-1+1*0] = 0.4

$V^*_1(s) = max(0.2,0.4)=0.4$

$Q^*_2(s,$ Нова верзија$)$ = T(s, Нова верзија, P)[R(s, Нова верзија,P)+ $\gamma V_1(P)$] + T(s, Нова верзија, SS)[R(s, Нова верзија,SS)+ $\gamma V_1(SS)$] + T(s,Нова верзија,N)[R(s,Нова Верзија,N)+ $\gamma V_1(N)$] = 0.1[2+1*0.4] + 0.9[0+1*0.2]+0[-1+1*1.7] = 0.24 + 0.18 + 0 = 0.42

$Q^*_2(s,$ Причекај$)$ = T(s, Причекај, P)[R(s, Причекај,P)+ $\gamma V_1(P)$] + T(s, Причекај, SS)[R(s, Причекај,SS)+ $\gamma V_1(SS)$] + T(s, Причекај, N)[R(s, Причекај,N)+ $\gamma V_1(N)$] = 0.2[2+1*0.4] + 0.8[0+1*0.2]+0[-1+1*1.7] = 0.48 + 0.16 = 0.64

$V^*_2(s) = max(0.42,0.64)=0.64$

За s = N следува:

$Q^*_1(s,$ Нова верзија$)$ = T(s, Нова верзија, P)[R(s, Нова верзија,P)+ $\gamma V_0(P)$] + T(s, Нова верзија, SS)[R(s, Нова верзија,SS)+ $\gamma V_0(SS)$] + T(s,Нова верзија, N)[R(s, Нова верзија,N)+ $\gamma V_0(N)$]= 0.9[2+1*0]+0[0+1*0]+0.1[-1+1*0] = 0.9*2 – 0.1 = 1.7

$Q^*_1(s,$ Причекај$)$ = T(s, Причекај, P)[R(s, Причекај,P)+ $\gamma V_0(P)$] + T(s, Причекај, SS)[R(s, Причекај,SS)+ $\gamma V_0(SS)$] + T(s, Причекај, N)[R(s, Причекај,N)+ $\gamma V_0(N)$] = 0[2+1*0] + 0.5[0+1*0]+0.5[-1+1*0] = -0.5

$V^*_1(s) = max(-0.5,1.7)=1.7$

$Q^*_2(s,$ Нова верзија$)$ = T(s, Нова верзија, P)[R(s, Нова верзија,P)+ $\gamma V_1(P)$] + T(s, Нова верзија, SS)[R(s, Нова верзија,SS) +$\gamma V_1(SS)$]+T(s,Нова верзија,N)[R(s,Нова верзија,N)+$\gamma V_1(N)$]=0.9[2+1*0.4] +0[0+1*0.2]+0.1[-1+1*1.7] = 2.16 +0+0.07 = 2.23

$Q^*_2(s,$ Причекај) = T(s, Причекај, P)[R(s, Причекај,P)+ $\gamma V_1(P)$] + T(s, Причекај, SS)[R(s, Причекај,SS)+ $\gamma V_1(SS)$] + T(s, Причекај, N)[R(s, Причекај,N)+ $\gamma V_1(N)$] = 0[2+1*0.4] + 0.5[0+1*0.2]+0.5[-1+1*1.7] = 0+0.1+0.35 = 0.45

$V^*_2(s) = max(2.23,0.45)=2.23$

За s = SS следува:

$Q^*_1(s,$ Нова верзија) = T(s, Нова верзија, P)[R(s, Нова верзија,P)+ $\gamma V_0(P)$] + T(s, Нова верзија, SS)[R(s, Нова верзија,SS)+ $\gamma V_0(SS)$] + T(s, Нова верзија, N)[R(s, Нова верзија,N)+ $\gamma V_0(N)$]= 0.1[2+1*0] + 0.9[0+1*0]+0[-1+1*0] = 0.1*2 = 0.2

$Q^*_1(s,$ Причекај$)$ = T(s, Причекај, P)[R(s, Причекај,P)+ $\gamma V_0(P)$] + T(s, Причекај, SS)[R(s, Причекај,SS)+ $\gamma V_0(SS)$] + T(s, Причекај, N)[R(s, Причекај,N)+ $\gamma V_0(N)$] = 0[2+1*0] + 0.3[0+1*0]+0.7[-1+1*0] = -0.7

$V^*_1(s) = max(0.2,-0.7)=0.2$

$Q^*_2(s,$ Нова верзија$)$ = T(s, Нова верзија, P)[R(s, Нова верзија,P)+ $\gamma V_1(P)$] + T(s, Нова верзија, SS)[R(s, Нова верзија,SS)+ $\gamma V_1(SS)$] + T(s,Нова верзија,N)[R(s,Нова верзија,N)+$\gamma V_1(N)$] = 0.1[2+1*0.4] +0.9[0+1*0.2]+0[-1+1*1.7] = 0.24+0.18 = 0.42

$Q^*_2(s,$ Причекај) = T(s, Причекај, P)[R(s, Причекај,P)+ $\gamma V_1(P)$] + T(s, Причекај, SS)[R(s, Причекај,SS)+ $\gamma V_1(SS)$] + T(s, Причекај, N)[R(s, Причекај,N)+ $\gamma V_1(N)$] = 0[2+1*0.4] + 0.3[0+1*0.2]+0.7[-1+1*1.7] = 0+0.06+0.49 = 0.55

$V^*_2(s) = max(0.42,0.55)=0.55$

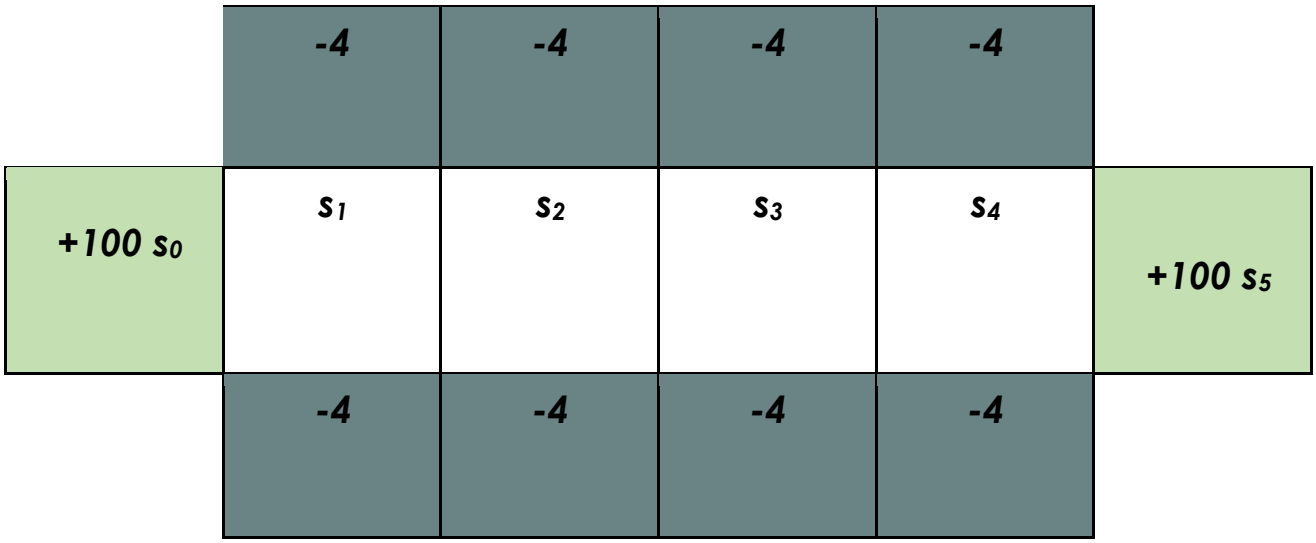
(б) Која е оптималната политика која Агентот ќе ја преземе доколку се наоѓа во состојбата **SS** и има уште два чекори до крајот на играта? Образложете го вашето решение.

- Доколку агентот се наоѓа ви состојбата **SS** и има уште два чекора за до крајот на играта тогаш агентот како следна акција ќе ја земе 'Причекај' бидејќи според последните резултати добиени од **value iterations** $Q^*_2(SS,$ Нова верзија$) < Q^*_2(SS,$ Причекај$)$. Потоа откако ќе ја направи оваа акција тогаш агентот може да се пронајде во состојба **SS** со веројатност 0.3 или па во состојбата **N** со веројатност 0.7. Ако агентот се пронајде во состојба **SS** тогаш како следна акција со најголем исход ќе ја земе повторно 'Причекај'. Но доколку

агентот се најде во состојбата **N** тогаш како следна ќе ја земе 'Нова верзија' бидејќи таа акција го има најголемиот исход од сите возможни акции што може агентот да ги направи.

2. Во потрага по богатство

Агентот се движи во лавиринт каде е скриено богатство. Акциите кои може да ги преземе агентот се помести се Лево (**L)** или Десно (**R)**, но поради стохастичноста може да заврши во поле различно од очекуваното. Ако агентот избере акција Десно со веројатност 0.5 ќе се придвижи во посакуваната насока, но со иста со веројатност 0.5 може да се лизне и падне во бездна каде играта завршува со казна од -4. Ако агентот избере акција Лево се поместува во полето лево со веројатност 1. Воедно, постои и акцијата **Exit** за излез од терминалните полиња, кои се означена со награда. Вредноста на g ***= 1***



За дадениот проблем чие решавање треба да го дефинирате како MDP, потребно е да одговорите на следните прашања/задачи:

(a) Да се пресметаат вредностите на состојбите, *V(s, a, s')* за првите 3 чекори од постепеното рекурзивно пресметување на состојбите (i=0, i=1, i=2) и пополни табелата. Образложете ги пресметките преку формулите кои сте ги користеле при пополнување на табелата. (F= бездна)

	S1	S2	S3	S4
$V^*_0(s)$	0	0	0	0
$Q^*_1(s, Лево)$	100	0	0	0
$Q^*_1(s, Десно)$	-2	-2	-2	48
$V^*_1(s)$	100	0	0	48
$Q^*_2(s, Лево)$	100	100	0	0
$Q^*_2(s, Десно)$	-2	-2	22	48
$V^*_2(s)$	100	100	22	48

За $V^*_0(s)$ каде s = {S1,S2,S3,S4} под претпоставка ги иницијализираме сита да бидат еднакви на 0.

За s=S1:

$Q^*_1(s, Лево) = T(s,Лево,S0)[R(s,Лево,S0)+ \gamma V_0(S0)] = 1*[100+1*0]=100$

$Q^*_1(s, Десно) = T(s,Десно,S2)[R(s,Десно,S2)+ \gamma V_0(S2)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_0(F)] = 0.5[0+1*0] +0.5*[-4+1*0] = 0-2 = -2$

$V^*_1(s) = max(-2,100)=100$

$Q^*_2(s, Лево) = T(s,Лево,S0)[R(s,Лево,S0)+ \gamma V_1(S0)] = 1*[100+1*0]=100$

$Q^*_2(s, Десно) = T(s,Десно,S2)[R(s,Десно,S2)+ \gamma V_1(S2)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_1(F)] = 0.5[0+1*0] +0.5[-4+1*0] = -2$

$V^*_2(s) =max(100,-2)=100$

За s=S2:

$Q^*_1(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1)+ \gamma V_0(S1)] = 1*[0+1*0]=0$

$Q^*_1(s, Десно) = T(s,Десно,S3)[R(s,Десно,S3)+ \gamma V_0(S3)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_0(F)] = 0.5[0+1*0] +0.5*[-4+1*0] = 0-2 = -2$

$V^*_1(s) = max(0,-2)=0$

$Q^*_2(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1)+ \gamma V_1(S1)] = 1*[0+1*100]=100$

$Q^*_2(s,$ Десно$) = T(s,$Десно$,S3)[R(s,$Десно$,S3)+ \gamma V_1(S3)] + T(s,$Десно$,F)[R(s,$Десно$,F)+ \gamma V_1(F)] = 0.5[0+1*0] +0.5[-4+1*0] = -2$

$V^*_2(s) = max(100,-2)=100$

За s=S3:

$Q^*_1(s,$ Лево$) = T(s,$Лево$,S2)[R(s,$Лево$,S2)+ \gamma V_0(S2)] = 1*[0+1*0]=0$

$Q^*_1(s,$ Десно$) = T(s,$Десно$,S4)[R(s,$Десно$,S4)+ \gamma V_0(S4)] + T(s,$Десно$,F)[R(s,$Десно$,F)+ \gamma V_0(F)] = 0.5[0+1*0] +0.5*[-4+1*0] = 0-2 = -2$

$V^*_1(s) = max(0,-2)=0$

$Q^*_2(s,$ Лево$) = T(s,$Лево$,S2)[R(s,$Лево$,S2)+ \gamma V_1(S2)] = 1*[0+1*0]=0$

$Q^*_2(s,$ Десно$) = T(s,$Десно$,S4)[R(s,$Десно$,S4)+ \gamma V_1(S4)] + T(s,$Десно$,F)[R(s,$Десно$,F)+ \gamma V_1(F)] = 0.5[0+1*48] +0.5[-4+1*0] = 24-2=22$

$V^*_2(s) = max(0,22)=22$

За s=S4:

$Q^*_1(s,$ Лево$) = T(s,$Лево$,S3)[R(s,$Лево$,S3)+ \gamma V_0(S3)] = 1*[0+1*0]=0$

$Q^*_1(s,$ Десно$) = T(s,$Десно$,S5)[R(s,$Десно$,S5)+ \gamma V_0(S5)] + T(s,$Десно$,F)[R(s,$Десно$,F)+ \gamma V_0(F)] = 0.5[100+1*0] +0.5*[-4+1*0] = 50-2 = 48$

$V^*_1(s) = max(0,48)=48$

$Q^*_2(s,$ Лево$) = T(s,$Лево$,S3)[R(s,$Лево$,S3)+ \gamma V_1(S3)] = 1*[0+1*0]=0$

$Q^*_2(s,$ Десно$) = T(s,$Десно$,S5)[R(s,$Десно$,S5)+ \gamma V_1(S5)] + T(s,$Десно$,F)[R(s,$Десно$,F)+ \gamma V_1(F)] = 0.5[100+1*0] +0.5[-4+1*0] = 50-2=48$

$V^*_2(s) = max(0,48)=48$

Оптимална политика: $\pi(S_1) = L,\ \pi(S_2) = L\ ,\ \pi(S_3) = R,\ \pi(S_4) = R$

Наоѓате мапа на скриено богатство од претходниот трагач со запис за неговото искуство, т.е политиката која ја преземал и вредноста на состојбата:

$$\boldsymbol{\pi(S_1) = R\ \pi(S_2) = L\ \ \pi(S_3) = L\ \pi(S_4) = R}$$

$$\boldsymbol{V(S_1) = -4 \quad V(S_2) = -4 \quad V(S_3) = 8 \quad V(S_4) = 10}$$

(б) Следејќи ја политиката на претходниот трагач, направете евалуација на политиката за првите 3 чекор (i=0, i=1, i=2) и пресметајте ги вредностите на сите 4 состојби, кои не се терминални.

Иницијализација: $V^*_0(S_1) = -4,\ V^*_0(S_2) = -4,\ V^*_0(S_3) = 8,\ V^*_0(S_4) = 10$

Пресметка на $V^*_1(s)$:

За $s=S1: V^*_1(s) = Q^*_1(s,Десно) = T(s,Десно,S2)[R(s,Десно,S2) + \gamma V_0(S2)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V0(F)] = 0.5[0+1*-4] + 0.5*[-4+1*0] = -2-2 = -4$

За $s=S2:\ V^*_1(s) = Q^*_1(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1) + \gamma V_0(S1)] = 1*[0+1*(-4)] = -4$

За $s=S3:\ V^*_1(s) = Q^*_1(s, Лево) = T(s,Лево,S2)[R(s,Лево,S2) + \gamma V_0(S2)] = 1*[0+1*-4] = -4$

За $s=S4:\ V^*_1(s) = Q^*_1(s,Десно) = T(s,Десно,S5)[R(s,Десно,S5) + \gamma V_0(S5)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V0(F)] = 0.5[100+1*0] + 0.5*[-4+1*0] = 50-2 = 48$

Пресметка на $V^*_2(s)$:

За $s=S1: V^*_2(s) = Q^*_2(s,Десно) = T(s,Десно,S2)[R(s,Десно,S2) + \gamma V_1(S2)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V_1(F)] = 0.5[0+1*-4] + 0.5*[-4+1*0] = -2-2 = -4$

За $s=S2:\ V^*_2(s) = Q^*_2(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1) + \gamma V_1(S1)] = 1*[0+1*(-4)] = -4$

За $s=S3:\ V^*_2(s) = Q^*_2(s, Лево) = T(s,Лево,S2)[R(s,Лево,S2) + \gamma V_1(S2)] = 1*[0+1*-4] = -4$

За $s=S4:\ V^*_2(s) = Q^*_2(s,Десно) = T(s,Десно,S5)[R(s,Десно,S5) + \gamma V_1(S5)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V_1(F)] = 0.5[100+1*0] + 0.5*[-4+1*0] = 50-2 = 48$

Крајна состојба: $V^*(S_1) = -4,\ V^*(S_2) = -4,\ V^*(S_3) = -4,\ V^*_0(S_4) = 48$

(в) Направете подобрување на политиката, врз основа на вредностите на состојбите добиени во претходното барање за чекор i=2!

За s=S1:

$Q^*_1(s, Десно) = T(s,Десно,S2)[R(s,Десно,S2) + \gamma V_0(S2)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V_0(F)] = 0.5[0+1*-4] + 0.5*[-4+1*0] = -2-2 = -4$

$V^*_1(s) = -4$

$Q^*_2(s, Лево) = T(s,Лево,S0)[R(s,Лево,S0) + \gamma V_1(S0)] = 1*[100+1*0] = 100$

$Q^*_2(s, Десно) = T(s,Десно,S2)[R(s,Десно,S2) + \gamma V_1(S2)] + T(s,Десно,F)[R(s,Десно,F) + \gamma V_1(F)] = 0.5[0+1*(-4)] + 0.5[-4+1*0] = -2-2 = -4$ $\quad V^*_2(s) = max(100,-4) = 100$

	S1	S2	S3	S4
$V^*_0(s)$	-4	-4	8	10
$Q^*_1(s, Лево)$	/	-4	-4	/
$Q^*_1(s, Десно)$	-4	/	/	48
$V^*_1(s)$	-4	-4	-4	48
$Q^*_2(s, Лево)$	100	-4	-4	-4
$Q^*_2(s, Десно)$	-4	-4	22	48
$V^*_2(s)$	100	-4	22	48

За s=S2:

$Q^*_1(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1)+ \gamma V_0(S1)] = 1*[0+1*(-4)]=-4$

$V^*_1(s) =-4$

$Q^*_2(s, Лево) = T(s,Лево,S1)[R(s,Лево,S1)+ \gamma V_1(S1)] = 1*[0+1*-4]=-4$

$Q^*_2(s, Десно) = T(s,Десно,S3)[R(s,Десно,S3)+ \gamma V_1(S3)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_1(F)] = 0.5[0+1*(-4)] +0.5[-4+1*0] =-2-2 = -4$

$V^*_2(s) =-4$ *Нема никаква промена*

За s=S3:

$Q^*_1(s, Лево) = T(s,Лево,S2)[R(s,Лево,S2)+ \gamma V_0(S2)] = 1*[0+1*-4]=-4$

$V^*_1(s) = -4$

$Q^*_2(s, Лево) = T(s,Лево,S2)[R(s,Лево,S2)+ \gamma V_1(S2)] = 1*[0+1*-4]=-4$

$Q^*_2(s, Десно) = T(s,Десно,S4)[R(s,Десно,S4)+ \gamma V_1(S4)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_1(F)] = 0.5[0+1*48] +0.5[-4+1*0] = 24-2=22$

$V^*_2(s) =max(-4,22)=22$

За s=S4:

$Q^*_1(s, Десно) =T(s,Десно,S5)[R(s,Десно,S5)+ \gamma V_0(S5)] + T(s,Десно,F)[R(s,Десно,F)+ \gamma V_0(F)] = 0.5[100+1*0] +0.5*[-4+1*0] = 50-2 = 48$

$V^*_1(s) = 48$

$Q^*_2(s, \text{Лево}) = T(s,\text{Лево},S3)[R(s,\text{Лево},S3)+ \gamma V_1(S3)] = 1*[0+1*-4]=-4$

$Q^*_2(s, \text{Десно}) = T(s,\text{Десно},S5)[R(s,\text{Десно},S5)+ \gamma V_1(S5)] + T(s,\text{Десно},F)[R(s,\text{Десно},F)+ \gamma V_1(F)] = 0.5[100+1*0] +0.5[-4+1*0] = 50-2=48$

$V^*_2(s) =max(-4,48)=48$

За да се направи подобрување на политиката, врз основа на вредностите на состојбите добиени во претходното барање за чекор i=2, потребно е наместо $\pi(S_1) = R$ треба $\pi(S_1) =L$ и $\pi(S_3) = L$ да се замени со $\pi(S_3) = R$.