Gyakorló feladatok a 2. pótzh. előtt
Megerősítéses tanulás

Pataki Béla,

BME I.E. 414, 463-26-79
pataki@mit.bme.hu,
http://www.mit.bme.hu/general/staff/pataki

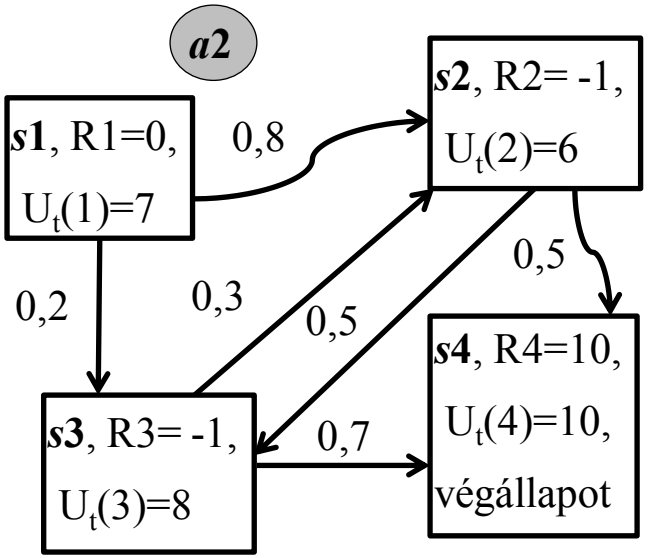
A 2. zh megerősítéses tanulásos feladata

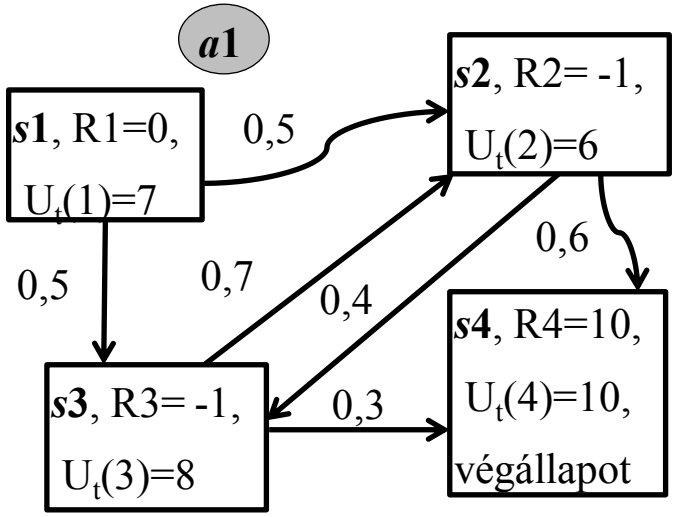
Aktív megerősítéses tanulást végzünk a 4 állapottal rendelkező térben. Minden állapotban két cselekvés közt választhatunk: *a*1, illetve *a*2. A végállapot *s*4, a kiinduló állapot mindig *s*1. Az általunk használt felfedezési függvény (a szokásos jelölésekkel):

$$f(u,n) = \begin{cases} 20 & \text{ha} \quad n < 4 \\ u & \text{különben} \end{cases}$$

Az eddigi futások során nyert információinkat az alábbi két ábrán, illetve az ábrák alatti táblázatban foglaltuk össze. Az egyes szövegdobozokban az állapot, az állapotban kapható jutalom, illetve az állapotra vonatkozó jelenlegi (*t* iterációs lépés utáni) optimista hasznosságbecslésünk van. A baloldali ábrán az *a*1, a jobboldalin az *a*2 cselekvés esetén tapasztalt állapotátmenetvalószínűségek láthatók a dobozok közt az átmenetet mutató nyilak mellett. A táblázatban az látható,hogy az egyes állapotokban hányszor választottuk az egyik, illetve a másik cselekvést.

A 2. zh megerősítéses tanulásos feladata





$$f(u,n) = \begin{cases} 20 & \text{ha} \quad n < 4 \\ u & \text{különben} \end{cases}$$

N(sk,am)	s1	s2	s3	s4
a1	9	2	3	
a2	7	4	6	

Jelenleg az *s*2 állapotban vagyunk, milyen cselekvést fog választani az aktív tanuló ágens? (Természetesen röviden indokolja a választ!)

A 2. zh Q-tanulásos feladata

Aktív megerősítéses Q-tanulást végzünk időbeli különbség SARSA módszerrel. Az alábbi lépéssorozat (állapot-cselekvés párok) mentén módosítunk, a tanulás bátorsági faktora 0,1; a leszámítolási tényező 0,9. Jutalmat csak az *s*4 végállapotban kapunk, R(s4)=1.

*s*1, *a*1 $\Rightarrow$ *s*2,*a*2 $\Rightarrow$ *s*2,*a*1 $\Rightarrow$ *s*1,*a*1 $\Rightarrow$ ***s*3,*a*1** $\Rightarrow$ *s*1,*a*2 $\Rightarrow$ *s*4, *végállapot*

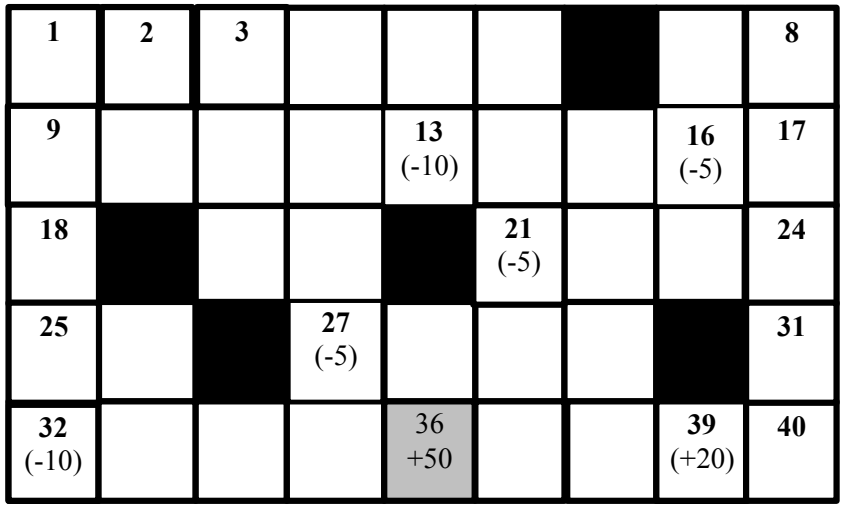
Mi lesz az *s*3-as állapot Q-értékeinek (Q(***a*1,*s*3**) és Q(***a*2,*s*3**)) új becslése a lépéssorozat után, ha a lépéssorozatot megelőzően az egyes állapot-cselekvés párok Q értékeinek becslése a következő volt.

Q(am,sk)	s1	s2	s3	s4
a1	0,3	0,22	0,3	
a2	0,8	0,37	0,6	

Passzív megerősítéses tanulás

Az alábbi útvesztőben passzív megerősítéses tanulást végzünk. Az egyes mezőkben a felső szám az állapot sorszáma, az alsó, zárójelben lévő szám – ha van – az állapothoz rendelt jutalom értéke. Ahol nincs feltüntetve a jutalom értéke, ott nulla. Nem minden mezőben tüntettük fel az állapot sorszámát, értelemszerűen sorban növekednek a számok. A végállapot a szürke színnel is megjelölt 36-os. A tanulást rekurzív átlagolással végezzük, a leszámítolási tényező 1.

U1 = -4	U2 = -3	U3 = -2	U4 = -1	U5 = 0
U6 = 1	U7 = 1	U8 = 0	U9 = -6	U10 =-5
U11 = 4	U12 = -8	U13 =-9	U14 = -4	U15 = -2
U16 = -8	U17 = -1	U18 = 0	U19 = 2	U20 = 3
U21 = -8	U22 = -7	U23 = -6	U24 = 15	U25 = 20
U26 = 25	U27 = 32	U28 = 36	U29 = 30	U30 = 25
U31 = 55	U32 = 30	U33 = 32	U34 = 36	U35 = 47
U36 = 50	U37 = 48	U38 = 47	U39 = 66	U40 = 65



Közvetlen (naiv) módosítás eljárással tanítunk az alábbi lépéssorozat mentén:

$$1 \Rightarrow 9 \Rightarrow 10 \Rightarrow 11 \Rightarrow 19 \Rightarrow 20 \Rightarrow 27 \Rightarrow 28 \Rightarrow 36$$

Mi lesz a 11-es állapot hasznosságértékének új becslése, ha a tanulást megelőzően a hasznosságbecslések a táblázatban láthatók voltak, és előtte már 7-szer jártunk a 11-ben?

Egy **szekvenciális döntési problémában** 4 állapot alkotja az állapotteret, *S*3 a végállapot. Minden állapotban kétféle cselekvés (a_1 és a_2) közt választhatunk. Az alábbi baloldali táblázatban láthatók az a_1-hez tartozó állapotátmenet-valószínűségek, a jobboldali táblázatban az a_2 cselekvéshez tartozók. A leszámítolási tényező γ=0,5.

a_1 esetén		**s'**			
T(s → s')		S1	S2	S3	S4
s	S1	0,5	0	0,5	0
	S2	0	0,5	0	0,5
	S3	0	0	0	0
	S4	0	0	0,5	0,5

a_2 esetén		**s'**			
T(s → s')		S1	S2	S3	S4
s	S1	0	0,5	0	0,5
	S2	0,2	0	0,8	0
	S3	0	0	0	0
	S4	1	0	0	0

Eljárásmóditerációs algoritmust alkalmazunk, az eddigi iterációk eredményeképp a *t*. lépésben a becsült eljárásmód: $\pi_t(S_1)$ = a1; $\pi_t(S_2)$ = a2; $\pi_t(S_4)$=a1. Ez alapján a a *t*. lépésben a becsült hasznosságok $U_t(S_1) = 0$; $U_t(S_2) = 1$; $U_t(S_3)$= $R(S_3)$= 1,2 ; $U_t(S_4) = 1$.
Adja meg a *t*+1-dik iterációs lépésre előálló eljárásmódot!

Szekvenciális döntési probléma

$$U(s) = R(s) + \gamma \cdot \max_a \sum_{s'} T(s,a,s') \cdot U(s')$$

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U(s')$$

Eljárásmód-iteráció:

1. t =0 - valamilyen kiinduló eljárásmód π_0(s)

2. $U_t(s)$ meghatározása az $U_t(s) = R(s) + \gamma \sum_{s'} T(s,\pi_t(s),s') \cdot U_t(s')$ lineáris egyenletrendszerből (*n* állapot, *n* egyenlet)

3. Amelyik *s*-re $\max_a \sum_{s'} T(s,a,s') \cdot U_t(s') > \sum_{s'} T(s,\pi_t(s),s') \cdot U_t(s')$

   arra $\pi_{t+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U_t(s')$ a többire $\pi_{t+1}(s) = \pi_t(s)$

4. $t \leftarrow t+1$

5. Ha már egyik *s*-nél sincs változás - KÉSZ , ha volt változás, akkor folytassuk újra a 2.-nél

Egy **szekvenciális döntési problémában** 3 állapot alkotja az állapotteret, *S*2 a végállapot. Minden állapotban kétféle cselekvés (a_1 és a_2) közt választhatunk. Az alábbi baloldali táblázatban láthatók az a_1-hez tartozó állapotátmenet-valószínűségek, a jobboldali táblázatban az a_2 cselekvéshez tartozók. A leszámítolási tényező γ=0,5.

P(*s* → *s'*|a1)

s \ s'	1	2	3
1	0	0,3	0,7
2	0	0	0
3	0,4	0,6	0

P(*s* → *s'*|a2)

s \ s'	1	2	3
1	0	1	0
2	0	0	0
3	0,6	0,4	0

Az állapotokban elnyerhető jutalmak, és a kiinduló hasznosságbecslések:

s	1	2	3
R(s)	+1	+20	-4
U_0(s)	0	+20	0

Értékiterációs algoritmussal adja meg a következő lépésben kapott hasznosságbecsléseket!

Szekvenciális döntési probléma

Bellman egyenlet: $U(s) = R(s) + \gamma \cdot \max_a \sum_{s'} T(s,a,s') \cdot U(s')$

Értékiteráció (t változó az iterációs lépések számlálója):

1. $t = 0$ - valamilyen kiinduló hasznosságfüggvény $U_0(s)$
2. $U_{t+1}(s)$ meghatározása (nem oldjuk meg az egyenletrendszert, csak kiszámítunk mindegyikből egy-egy új $U(s)$ értéket, a max-ot persze használjuk!)

$$U_{t+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s,a,s') \cdot U_t(s')$$

3. $t \leftarrow t+1$
4. Ha már egyik $U(s)$-nél sincs változás (vagy egy adott értéknél kisebb) - KÉSZ , ha volt változás, folytassuk újra a 2.-nél

Aktív megerősítéses Q-tanulást végzünk felfedezési függvényt használó IK módszerrel. A tanulás bátorsági faktora 0,1; a leszámítolási tényező 0,4. Jutalmat csak az *s*4 végállapotban kapunk, R(s4)=1.
Mi lesz az *s*3-as állapot Q-értékeinek (Q(***a*1,*s*3**) és Q(***a*2,*s*3**)) új becslése a lépéssorozat után, ha a lépéssorozatot megelőzően az egyes állapot-cselekvés párok Q értékeinek becslése a következő volt.

N(sk,am)	*s*1	*s*2	*s*3	*s*4
***a*1**	4	6	7	
***a*2**	7	5	6	

Q(am,sk)	*s*1	*s*2	*s*3	*s*4
***a*1**	0,3	0,2	0,3	
***a*2**	0,8	0,3	0,6	

$$f(u,n)=\begin{cases}10 & \text{ha } n<5\\ u & \text{egyébként}\end{cases}$$

A következő átmenet után mi lesz a Q(*s*3,*a*1) új becslése? Milyen cselekvést javasol most a állapotban az eljárás?

$$\boldsymbol{s3,a1 \Rightarrow s1}$$

function Q-Tanuló-Ágens(*észlelés*) **returns** egy cselekvés
inputs: *észlelés* , egy észlelés, amely a pillanatnyi s' állapotot és az r' jutalmat tartalmazza
static: **Q** , egy cselekvésérték-tábla; állapottal és cselekvéssel indexelünk

N$_{sa}$, az állapot-cselekvés párok gyakorisági táblája
s, a, r az előző állapot, előző cselekvés, jutalom, kezdeti értékük nulla

if s nem nulla **then do**
inkrementáljuk $N_{sa}[s,a]$-t
Q$[a,s] \leftarrow$ **Q**$[a,s]+\alpha(r+\gamma \max_{a'}$**Q**$[a',s']-$**Q**$[a,s])$
if Végállapot?[s'] **then** s,a,r ←nulla
else $s,a,r \leftarrow s'$, $\text{argmax}_{a'} f(Q[a',s'], N_{sa}[a',s'])$, r'
return *a*

(aztán a meghívó függvény a következő lépésben végrehajtja *s'*-ben és a visszakapott *a* cselekvést, eljut *s"*-be és frissíti Q[*a*,*s'*]-t stb.)