Alkalmazott mesterséges intelligencia (AMI)

http://www.mit.bme.hu/oktatas/targyak/vimibb01

8. – 9. ea. (2023 ősz)

Szekvenciális döntési problémák

http://http://mialmanach.mit.bme.hu/aima/ch17 17. fejezet

Előadó: Pataki Béla

a fóliák
Dobrowiecki Tadeusz és
Hullám Gábor anyagainak
felhasználásával készültek

https://www.esrcheck.com/2023/06/05/artificial-intelligence-ai-experts-sign-statement-on-ai-risk/

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

Tanulás alapvető fajtái:

felügyelt tanulás egy komponensnek mind a bemenetét, mind a kimenetét észlelni tudjuk (bemeneti minta + kívánt válasz)

megerősítéses tanulás az ágens az általa végrehajtott tevékenység csak bizonyos értékelését kapja meg, esetleg nem is minden lépésben (jutalom, büntetés, **megerősítés**)

felügyelet nélküli tanulás semmilyen információ sem áll rendelkezésünkre a helyes kimenetről (az észlelések közötti összefüggések tanulása)

féligellenőrzött tanulás a tanításra használt esetek egy részénél mind a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz), a másik – tipikusan nagyobb – részénél csak a bemeneti leírás ismert

Eddigiekben *felügyelt tanulás* esetén egy-egy mintához tartozott egy-egy jó válasz, kívánt eredmény. *Nemfelügyelt tanulásnál* pedigegyáltalán nem tudtuk a kívánt váaszt.
Fontos problémacsoport, amikor egy ***lépéssorozat*** mentén ***csak időnként kapunk visszajelzést***, hogy jó vagy rossz, amit csinálunk.

- Triviális példa a sakk, csak a végén derül ki pontosan, hogy jó volt-e a lépéssorozatunk.
- Amikor az egyetemről hazamegyünk, egy sor döntést hozunk, hogy milyen járművel, merre menjünk – a végén derül ki, hogy gyorsan vagy lassan értünk-e haza.
- Több napon át döntéseket hozunk, hogy tanuljunk, zh-ra készüljünk, vagy barátainkkal bulizzunk vagy dolgozzunk vagy látogassuk meg a nagymamát. Később derül ki (ha egyáltalán), hogy jól döntöttünk-e.
- A pénzünket elköltsük-e (mozijegyre, sörre, új nadrágra, biciklire, autóra, házra) vagy fektessük be (X részvénybe, Y részvénybe, államkötvénybe, aranyba, szotyolába…).

A ***lépéssorozatot*** $\Rightarrow$ ***állapotsorozat***ként modellezzük, pl:

S0 $\Rightarrow$ S17 $\Rightarrow$ • • • $\Rightarrow$ S103 $\Rightarrow$ S53

Pl. a hazafele utazós példában:

S0: az egyetemen vagyok

S1: a Petőfi-hídnál lévő villamosmegállóban vagyok

S2: a 4-esen vagy 6-oson vagyok Pest fele

S3: a 4-esen vagyok Újbuda központ fele

S4: a 6-oson vagyok a Móricz Zs. körtér fele

S5: a Móricz Zs. körtéren leszálltam a villamosról

……..

Bizonyos lépéseknél kapunk **jutalmat** (***reward***, ez lehet negatív jutalom, büntetés vagy költség is, pl. hosszú ideig megy a villamos), ha például a fenti (S0$\Rightarrow$S17$\Rightarrow$ • • •) állapotsorozatnál:

R(17)= +2, R(103)= -17, R(53)=+100 a többinél nincs jutalom,

akkor a lépéssorozat hasznossága (**additív jutalom**) $\Rightarrow$+2-17+100=+85

Példa (a Russel-Norvig könyv példája alapján):

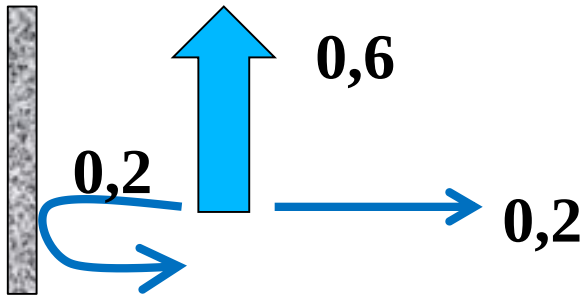
Kezdőállapot: bal alsó sarok
Végállapotok: *s*(4,3) és *s*(4,2) – csak itt van jutalom, másutt 0.

Választható cselekvések: ’**Fel**’, ’**Le**’, ’**Jobbra**’, ’**Balra**’.

A cselekvés hatására 60% valószínűséggel a kívánt irányba megy,

20-20% valószínűséggel valamelyik merőleges irányba. Ha falnak ütközik helyben marad.

Például „Fel” az (1,1)-en :



3	(1,3)	(2,3)	(3,3)	R(4,3)=+1
2	(1,2)		(3,2)	R(4,2)=-1
1	(1,1)	(2,1)	(3,1)	(4,1)
	1	2	3	4

Szekvenciális döntési probléma

Kezdőállapot: s0

Állapotátmenet-modell: **T(*s*, a, *s'*):** ***s*→*s'*** átmeneti valószínűség valószínűsége, amikor az ***a*** cselekvést választottuk s-ben

Jelölések: T(*s* , a , *s'*) vagy másképp jelölve *P*(*s*→*s'* **|** *a*)

Markov tulajdonság: elég az *s* állapotot ismernem, a megelőző állapotok nem érdekesek az *s*→*s'* átmenet szempontjából

Jutalom: **R(*s*)** (Nálunk most: additív!)

Jelölés: R(s), v. R(s, a, s') – az R(s)-nél az állapothoz kötjük a jutalmat, az R(s, a, s')-nél a jutalmat az *s* állapotban az *a* cselekvés végrehajtásához kötjük akkor, ha a cselekvés hatására s'-be jutottunk.

Kvíz 08.3 megoldás

A LE cselekvés esetén milyen valószínűség található a kérdőjellel jelölt (2,1) → (2,1) cellában?

P((2,1)→2,1) | *LE*) vagy másképp jelölve T((2,1) , *LE* , (2,1))

LE

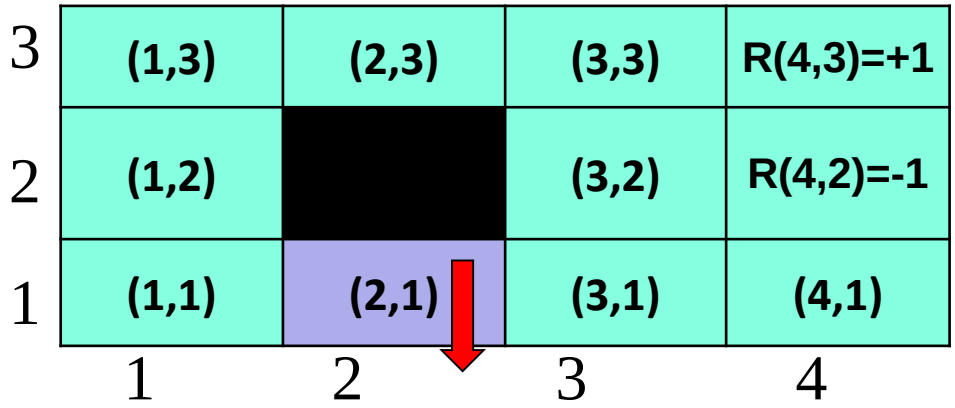
s↓ s'→	….	(1,3)	(2,1)	(2,3)	….
….	….	….	….	…	…
(1,3)	….	….	….	…	…
(2,1)	….	….	?	…	…
(2,3)	….	….	….	…	…

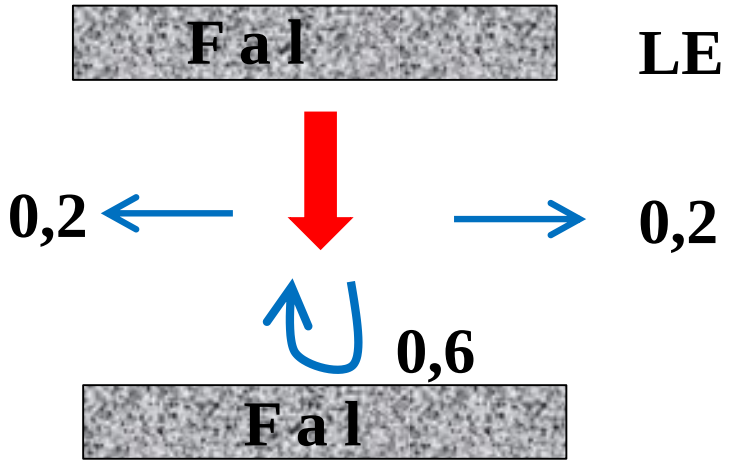
A. 0,3

B. 0,4

C. 0,6

D. 0,8





s↓ s'→	s'=s1	s'=s2	s'=s3	s'=s4	….
s=s1	….	….	….	…	…
s=s2	….	….	….	…	…
s=s3	….	….	….	…	…
….	….	….	….	…	…

$$T(s, A, s') = P(s \rightarrow s' \mid a=A)$$

A T(s,a,s') adott *a*=A mellett egy kétdimenziós mátrix (táblázat). (itt most az sorok tartoznak egy adott kiindulóállapothoz, és az oszlopok egy-egy követőállapothoz)

Az ábrán látható T(s,A,s') mátrixra az alábbiak közül melyik állítás igaz?

A mátrix egyetlen eleme se lehet 0.

A mátrix egy-egy oszlopában található számok összege mindig 1.

Az egész mátrixban található összes szám összege mindig 1.

A mátrix egy-egy sorában található számok összege mindig 1.

Szekvenciális döntési probléma

Valahogy el kell döntenünk, hogy melyik állapotsorozat jobb nekünk.

Kell egy ***skalár mérőszám*** (teljesítménymérték), hogy sorba tudjuk rendezni a lehetőségeinket!

1. Az állapot **hátralevő-jutalma** (**reward-to-go**) az adott lépéssorozatban azon jutalmak összege, amelyet akkor kapunk, ha az adott állapotból valamelyik végállapotig eljutunk.

$$s_0 \to s_1 \to .... \to s_{k_{VÉG}} \qquad U(s) = \sum_{k=0}^{k_{VÉG}} R(s_k)$$

(*De általában nem tudjuk biztosan, hogy s_0-ból milyen lépéssorozattal jutunk el $S_{kVÉG}$-be!*)

2. Leszámított jutalom („jobb ma egy veréb, mint holnap egy túzok”)

$$s_0 \to s_1 \to .... \to s_{k_{VÉG}} \qquad U(s) = \sum_{k=0}^{k_{VÉG}} \gamma^k \cdot R(s_k) \;\; ; \;\; 0 \le \gamma \le 1$$

Mivel az s állapotból bizonyos valószínűséggel jutunk különböző lépéssorozatokon át a végállapotba, ezért nem tudjuk előre megmondani, hogy mennyi lesz akár csak a mostani lépéssorozatban a pontos összjutalom a végállapotig. (Lehet, hogy azt se tudjuk megmondani, hogy melyik végállapotba jutunk, ha több van.)

Csak a ***várható értékét*** (***várható jutalom átlagosan*** sok esetre, próbálkozásra) tudjuk kiszámítani.

Azt például nem tudjuk megmondani, hogy mennyi időbe telik *ma*, hogy 4-es villamossal eljussunk az Oktogonig, de *átlagosan* meg lehet mondani, hogy várhatóan mekkora az *átlagos időszükséglet*. (Persze itt az eltelt idő legtöbbünknek negatív jutalom – költség vagy büntetés.)

A várható érték nagyszámú kísérlet esetén az egyes kísérletekben kapott összjutalom átlaga lesz!

Szekvenciális döntési probléma

Az s állapot hasznossága: ***U*(*s*) a legnagyobb <u>várható</u> hátralévő jutalom**,
ami az innen kiinduló lépéssorozatokban átlagosan *elérhető,* az egyes sorozatok valószínűségével súlyozva. (Nem az adott állapotban, hanem a végállapotig tartó lépéssorozat során összegezve!)

Az elérhető jutalom függ attól, hogy milyen cselekvéseket választunk!

Az s állapot hasznossága ***U*(*s*)** *–* ***optimális cselekvések*** *választása esetén!*

Optimális eljárásmód (stratégia, policy): megadja minden állapotra, hogy melyik az az ***a*** cselekvés az adott állapotban, ami maximalizálja a hátralévő jutalmat (optimális cselekvés).

$$a(s) = \pi^*(s)$$

Példa

Kezdőállapot: s11
Végállapotok: **s6** és **s8** – csak itt van jutalom, mindenhol máshol 0.

Az s7, s9, s10, s16 állapotokban csak a ’**Fel**’ cselekvés választható.

Az s1, s2, s3, s4, s5, s11, s13, s14, s15 állapotokban csak a ’**Jobbra**’ cselekvést választhatjuk.

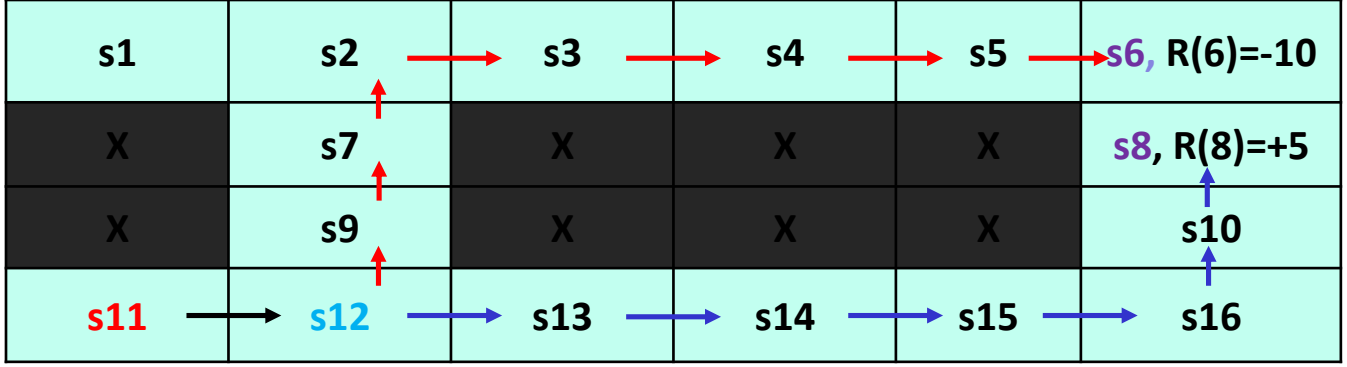
Az **s12** állapotban mindkettőt választhatjuk, de csak bizonyos valószínűséggel teljesül a választott cselekvés, a többi állapotban 100%-ban az történik, amit „akarunk” (’Fel’, illetve ’Jobbra’).

T(**s12**,a=’Fel’,s9)=0,7 T(**s12**,a=’Fel’,s13)=0,3

T(**s12**,a=’Jobbra’,s9)=0,3 T(**s12**,a=’Jobbra’,s13)=0,7

s1	s2	s3	s4	s5	s6, R(6)=-10
X	s7	X	X	X	s8, R(8)=+5
X	s9	X	X	X	s10
s11	s12	s13	s14	s15	s16

Mekkora lesz s11 hasznossága – U(11)?



s11-ből mindig s12-be jutunk, ott van egyedül választásunk.
Tegyük fel, hogy a '**Fel**' cselekvést választjuk s12-ben,

- ha jó irányba (s13) indul, utána s8-ba navigálhatjuk magunkat,
- ha s9-be jutottunk, akkor nem tudunk mit csinálni, előbb-utóbb s6-ba jutunk.

A várható nyereség ez esetben: 0,7*(-10)+0,3*(+5)= **-7+1,5 = -5,5**

Ha a '**Jobbra**' cselekvést választjuk s12-ben,

- ha mégis felfele indultunk, akkor nem tudunk mit csinálni, előbb-utóbb s6-ba jutunk,
- ha s13-ba jutottunk, akkor előbb-utóbb s8-ba jutunk.

A várható nyereség ez esetben: 0,3*(-10)+0,7*(+5)=)= **-3+3,5 = +0,5**
Rajtunk áll, hogy milyen cselekvést választunk s12-ben: a várható hátralévő összjutalom (az s11 hasznossága) **U(11)=+0,5**

Egy vetélkedőben a következő választás elé állítanak: 3 fiók van előttem, mindegyikben egy adott összeg. Az elsőben 11.000 Ft van, ezt megmondják. A másik kettőben ismeretlen sorrendben 2.000 Ft, 20.000 Ft. Bármelyiket választhatom, és akkor az enyém lesz a benne található pénz.

Várható értékben (=sok kísérlet átlagában) mi a maximális elérhető összeg?

A. 6.000 Ft
B. 8.000 Ft
C. 11.000 Ft
D. 20.000 Ft

a) Ha az 1-es fiókot választom, mindig **11.000** Ft-t kapok.

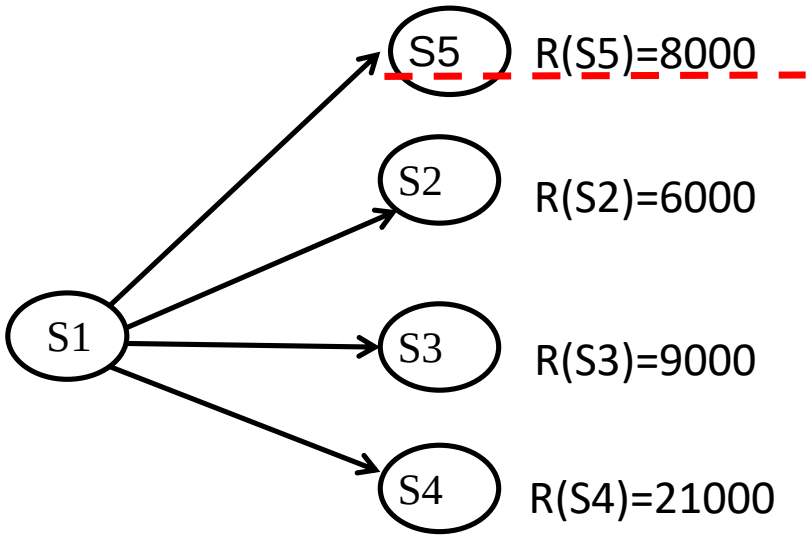
b) Ha a másik kettőből bármelyiket, akkor a várható jutalom:

$$VárhatóJutalom = \sum_k p(k) \cdot Jutalom(k) =$$

$$= 2000 \cdot \frac{1}{2} + 20000 \cdot \frac{1}{2} = 1000 + 10000 = 11000$$

Tehát itt mindegy, hogy melyik stratégiát választjuk, átlagban ugyanannyit nyerhetek! (Innen kezdve pszichológiai kérdés.)

Egy másik (ugyanilyen ostoba vetélkedős) példa (itt is 8000 Ft a fix az egyik fiókban, de 3 másik fiók is van az ábrán látható jutalmakkal)
A1: a fix összegű fiókot választom, A2: kockáztatok



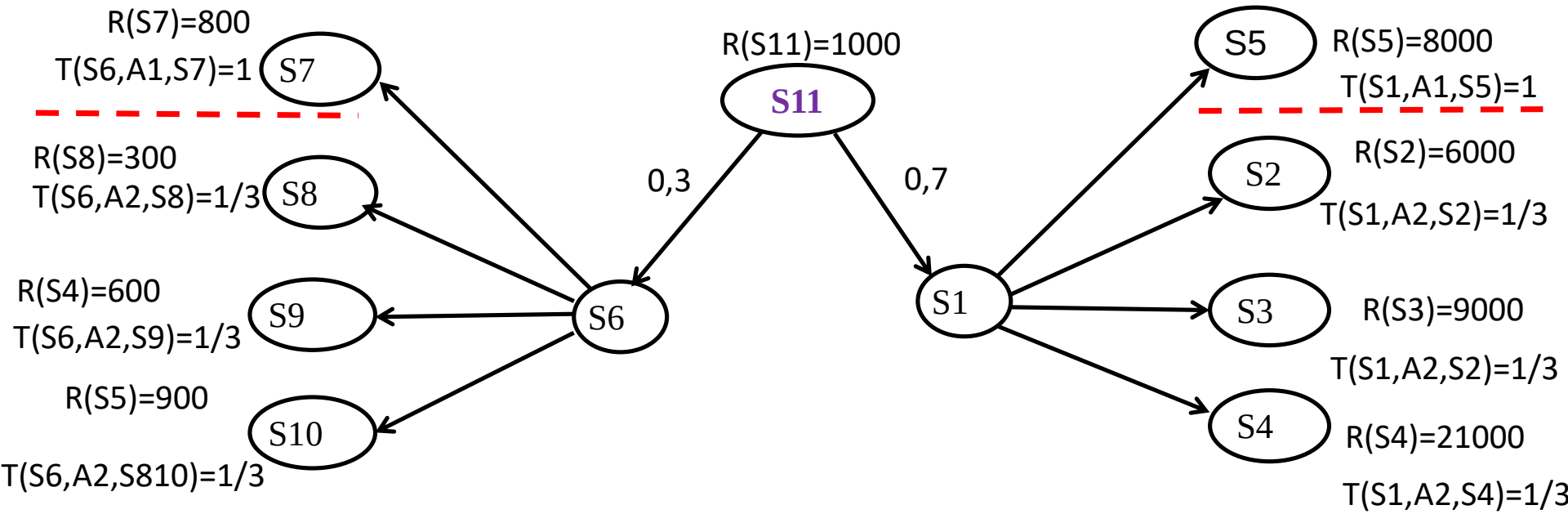
A1: T(S1,A1,S2)=0 ; T(S1,A1,S3)=0 ; T(S1,A1,S4)=0 ; T(S1,A1,S5)=1

A2: T(S1,A2,S2)=1/3 ; T(S1,A2,S3)=1/3 ; T(S1,A2,S4)=1/3 ; T(S1,A1,S5)=0

$\pi(S1) = A1$ $\quad U^{\pi}(S1)=8.000$

$\pi^{*}(S1) = A2$ $\quad U^{*}(S1)=12.000$

Bonyolítsuk tovább a példát! Tegyük fel, hogy amikor jelentkezem a vetélkedőre, akkor egyből kapok 1000 Ft-t. Viszont csak 70% eséllyel jutok a fent vázolt szituációba (4 fiók – S5, S2, S3, S4 – a fent részletezett valószínűségekkel és összegekkel), de 30% eséllyel csak egy kisebb költségvetésű hasonló helyzetbe kerülök, ahol a 4 fiókban a következő összegek vannak (**A1** = biztos jutalom és **A2** = hazárdírozás cselekvés, valószínűségeket lásd lent): S7→800 Ft, S8→ 300Ft, S9→ 600 Ft, S10→ 900 Ft.



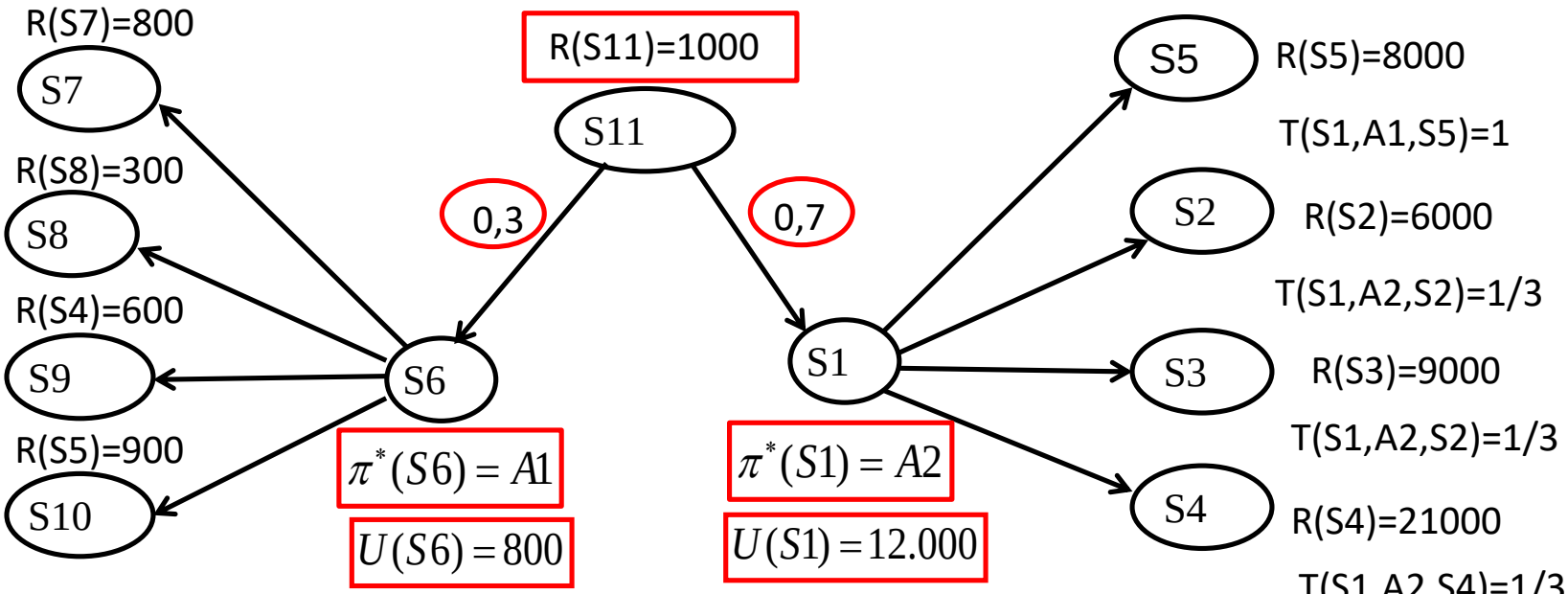
T(S11,A1,S1)=0,7 ; T(S11,A1,S6)=0,3 ; T(S11,A2,S1)=0,7 ; T(S11,A2,S6)=0,3

T(S6,A1,S8)=0 ; T(S6,A1,S9)=0 ; T(S6,A1,S10)=0 ; T(S6,A1,S7)=1

T(S6,A2,S8)=1/3 ; T(S6,A2,S9)=1/3 ; T(S6,A2,S10)=1/3 ; T(S6,A2,S7)=0

Mekkora lesz most U(s11)?

Bonyolítsuk tovább a példát! Tegyük fel, hogy amikor jelentkezem a vetélkedőre, akkor egyből kapok 1000 Ft-t. Viszont csak 70% eséllyel jutok a fent vázolt szituációba (4 fiók – S5, S2, S3, S4 –fent részletezett valószínűségekkel és összegekkel), de 30% eséllyel csak egy kisebb költségvetésű hasonló helyzetbe kerülök, ahol a 4 fiókban a következő összegek vannak (**A1** és **A2** cselekvés, valószínűségeket lásd lent):
S7→800 Ft, S8→ 300Ft, S9→ 600 Ft, S10→ 900 Ft.



T(S6,A1,S8)=0 ; T(S6,A1,S9)=0 ; T(S6,A1,S10)=0 ; T(S6,A1,S7)=1
T(S6,A2,S8)=1/3 ; T(S6,A2,S9)=1/3 ; T(S6,A2,S10)=1/3 ; T(S6,A2,S7)=0

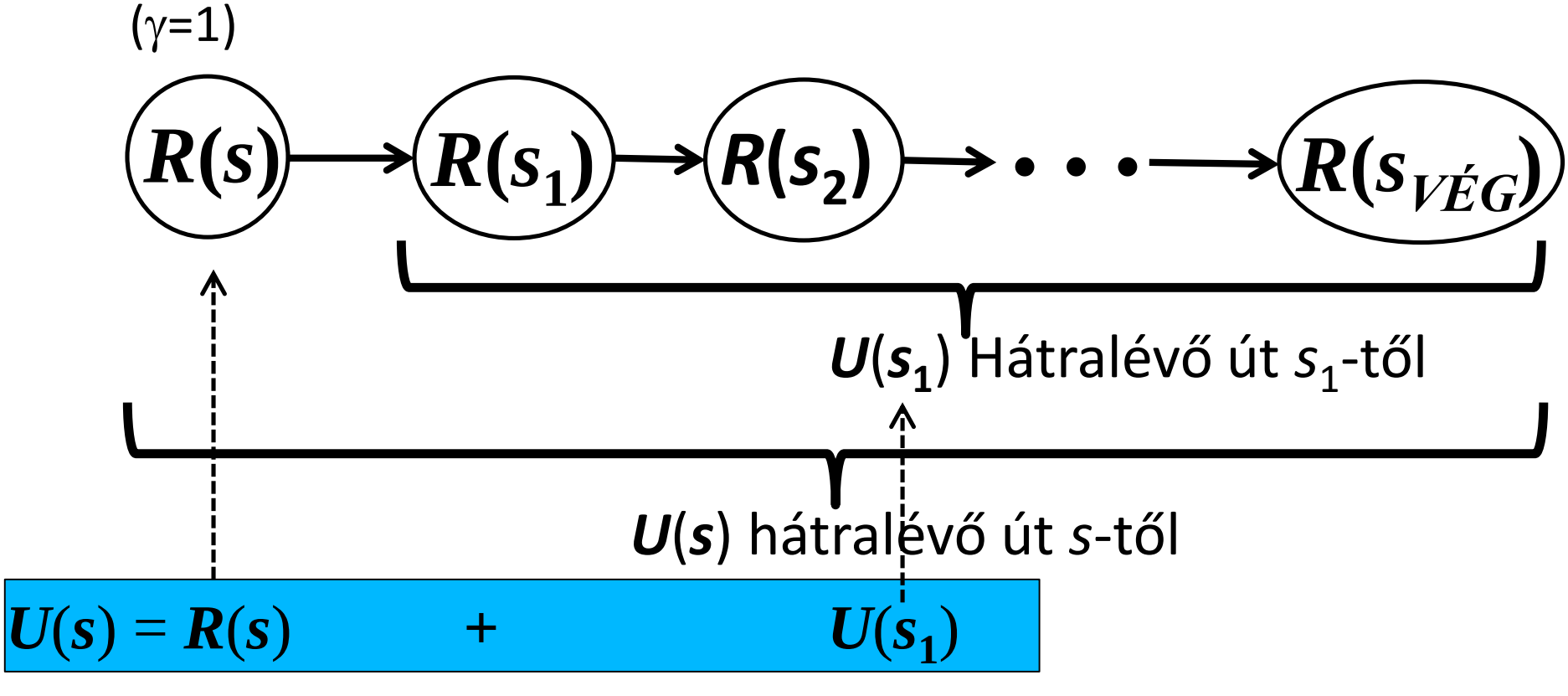
U(s11)=R(s11)+0,3*U(s6)+0,7*U(s1)=1000+240+8400=9640

Fontos egyszerűsítés volt: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság additív)

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

$$U^{\pi}(s) = E\left\{\sum_k \gamma^k \cdot R(s_k) \,|\, \pi, s_0 = s\right\}$$

Ha s-ből determinisztikusan egyetlen úton jutnánk el a végállapotba:

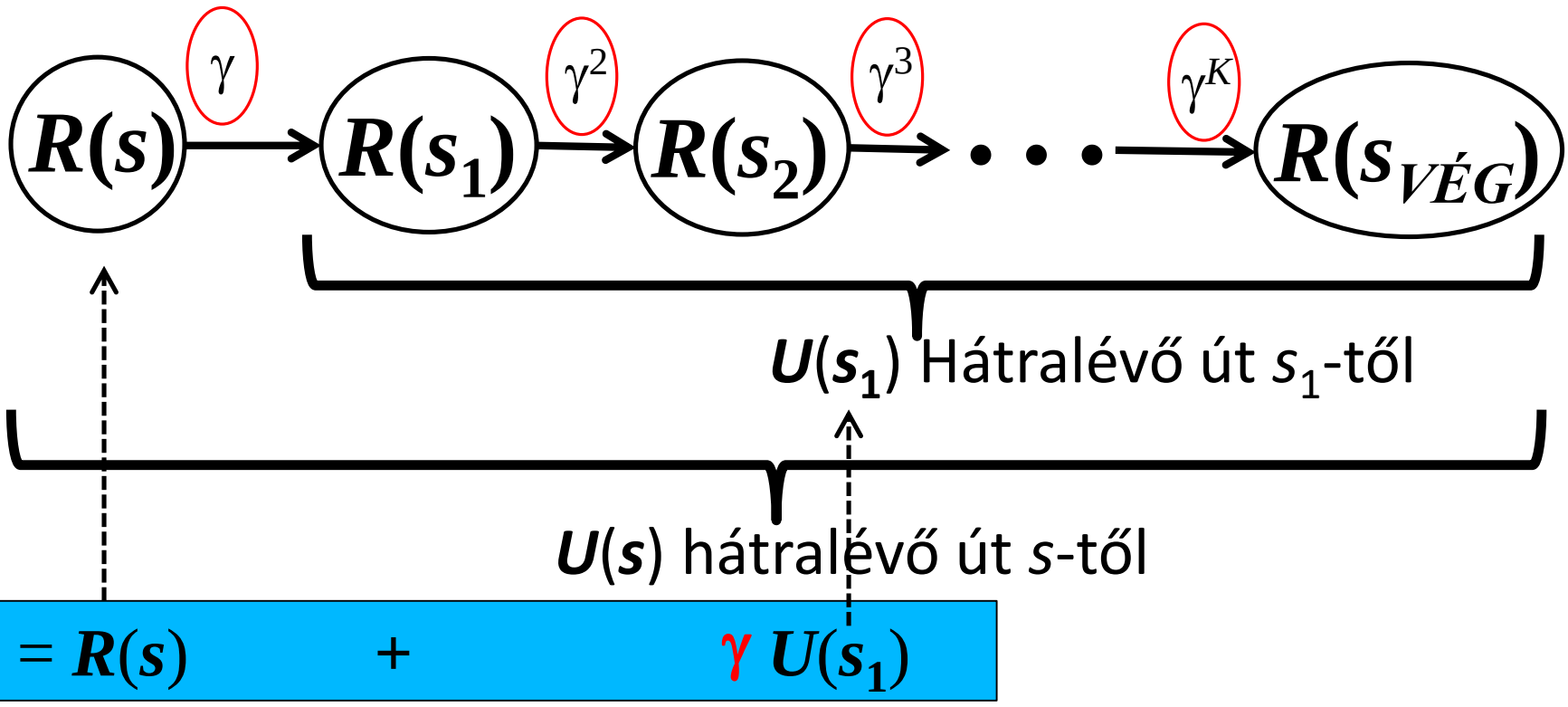


Fontos egyszerűsítés volt: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság additív)

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

$$U^{\pi}(s) = E\left\{\sum_k \gamma^k \cdot R(s_k) \,|\, \pi, s_0 = s\right\}$$

Ha s-ből determinisztikusan egyetlen úton jutnánk el a végállapotba:

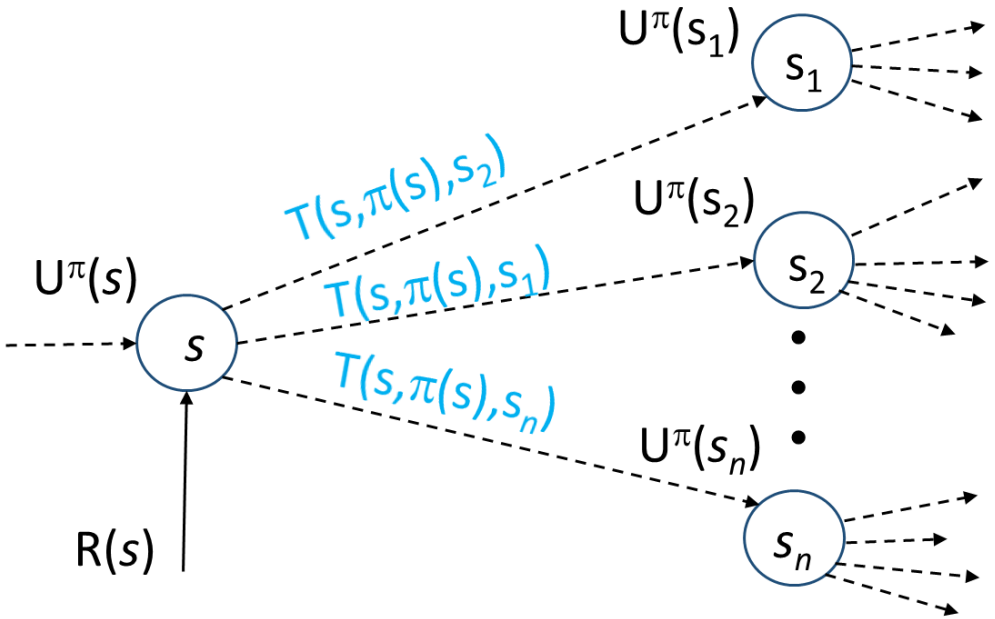


Fontos egyszerűsítés volt: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege (a hasznosság additív)

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke, a legjobb cselekvés választása esetén

$$U^{\pi^*}(s) = E\left\{\sum_k \gamma^k \cdot R(s_k) \,\middle|\, \pi^*, s_0 = s\right\}$$

(közönségesen: várható érték=az előfordulási gyakorisággal súlyozott átlag)



Jelölések:

$$T(s,a,s_k) = T(s,\pi(s),s_k) = T(s,s_k) = T_{sk}$$

$$U^{\pi}(s) = R(s) + \gamma T(s,s_1)\cdot U^{\pi}(s_1) + \ldots + \gamma T(s,s_n)\cdot U^{\pi}(s_n)$$

Szekvenciális döntési probléma

Optimális eljárásmód (stratégia, policy): megadja minden állapotra, hogy melyik az a cselekvés az adott állapotban, ami maximalizálja a hátralévő jutalmat.

$$a(s) = \pi^*(s)$$

Véges horizontú probléma: N lépés múlva véget ér, akár eljutottunk egy végállapotba, akár nem!

Az eljárásmód időfüggő (lépésszámfüggő)! $\Rightarrow$ nehezebb jó eljárásmódot kialakítani!!!

Pl. *N*=1000 esetén más kell legyen a stratégiánk a második lépésben, mint a 990-dikben! (A 20 éves és a 98 éves emberek stratégiája tipikusan kicsit eltérő. A focimeccs elején más a stratégia, mint a 88. percben. Merhogy véges a horizont.)

Végtelen horizontú: **nincs ilyen leketyegő számláló!** Lehet, hogy véges lépésszám után célba érünk, de ha nem, akkor sem fújják le soha a meccset, mindig ugyanaz az optimum.

Véges horizontú probléma: Egy adott, rögzített *N* lépés múlva véget ér a sorozat, akár eljutottunk egy végállapotba, akár nem!

Végtelen horizontú: Lehet, hogy véges lépésszám után célba érünk, de ha nem, soha nem fújják le a meccset.

Általában melyik probléma esetén nehezebb jó eljárásmódot (stratégiát) találni, a véges vagy a végtelen horizontú esetén?

Szekvenciális döntési probléma

Kapcsolat van az aktuális és az aktuálist követő állapotok hasznossága közt! Hiszen a hátralévő utak – ahol a jutalmakat gyűjtjük – részei a követő állapotok !

Markov döntési folyamat (MDF)

$a \in A$ – lehetséges cselekvések

s0- kiinduló (start) állapot, $s \in S$ – állapottér elemei

T(s, a, s') –állapotátmenet-valószínűségek

R(s) – jutalomfüggvény

Bellman egyenlet:

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

γ - leszámítolási tényező

Optimális eljárásmód:

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

Szekvenciális döntési probléma

Szeretnénk az állapotok hasznosságát kiszámítani.

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

Ha *K* állapottal írható le a probléma, akkor ez egy K ismeretlenes egyenletrendszer:

$$U(s_1) = R(s_1) + \gamma \cdot \max_a \sum_{k=1}^{K} T(s_1,a,s_k) \cdot U(s_k)$$

$$U(s_2) = R(s_2) + \gamma \cdot \max_a \sum_{k=1}^{K} T(s_2,a,s_k) \cdot U(s_k)$$

.....

$$U(s_K) = R(s_K) + \gamma \cdot \max_a \sum_{k=1}^{K} T(s_K,a,s_k) \cdot U(s_k)$$

A valós U(s_k)-kat akkor tudjuk kiszámítani, ha az optimális eljárásmódot ismerjük, DE az optimális eljárásmódot csak akkor tudjuk meghatározni, ha U(s_k)-kat ismerjük! ***Nemlineáris egyenletrendszert kéne megoldani*** **! (mindegyik egyenletben van „max”, ami nemlineáris függvény)**

Szekvenciális döntési probléma

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

Értékiteráció: ***Konvergens iteratív eljárás!***

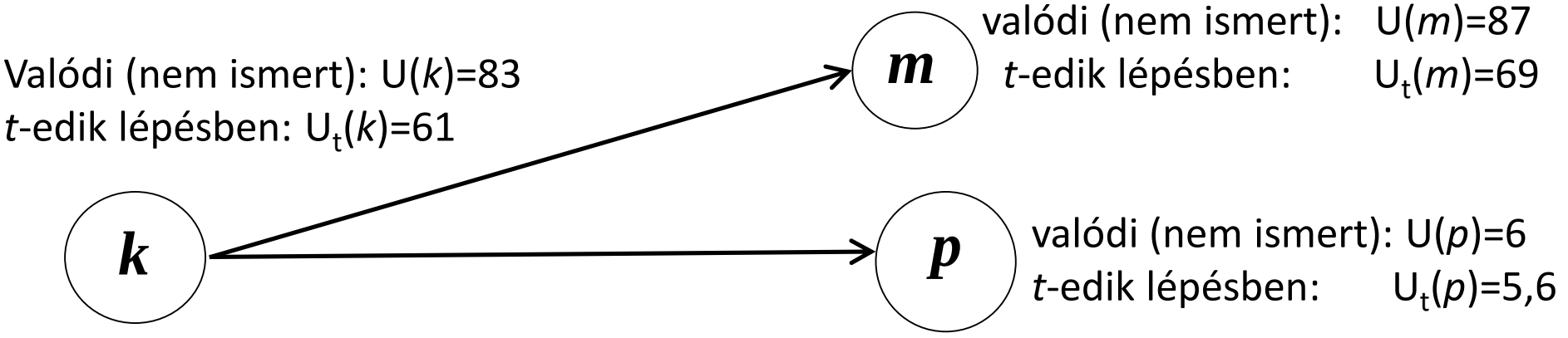
1. t =0 - valamilyen kiinduló hasznosságfüggvény $U_0(s_k)$, k=1,…,K
2. $U_{t+1}(s_k)$-k meghatározása (*nem oldjuk meg az egyenletrendszert*, csak kiszámítunk mindegyikből egy-egy új U(s) értéket, a max-ot persze használjuk!)

$$U_{t+1}(s_k) \leftarrow R(s_k) + \gamma \max_a \sum_{s'} T(s_k,a,s') \cdot U_t(s'), \quad k = 1,2,\ldots,K$$

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

Eljárásmód-iteráció:

A hasznosságértékek lassan konvergálnak az értékiterációnál, de a ***legjobb cselekvés*** rendszerint már jóval a pontos konvergencia előtt eldönthető! Például (iteráció=1,2,3…,*t*,…. lépésekben):



Már látható a *t*-edik lépésben is, hogy a k→m átmenetet eredményező cselekvések a jobbak, pedig az $U_t(k)$, $U_t(m)$, $U_t(p)$ értékek még messze vannak a valóstól!

Eljárásmód-iteráció ötlete:

Tapasztalat: már látható a *t*-dik lépésben is, hogy pl. a $k \to m$ átmenetet eredményező cselekvések a jobbak, pedig az $U_t(\cdot)$ értékek még messze vannak a valóstól!

Ötlet: ha a *t*-edik iterációs lépésben **az eljárásmódunkat, $\pi_t(s)$-t rögzítjük** (a jelenleg ismert hasznosságok alapján), akkor az $U_t(s)$-hez *nem kell a nemlineáris max!*
Ugyanis rögzítettük az eljárásmóddal, hogy jelenleg (*t*-dik iteráció) milyen cselekvést választunk. Tehát *n* állapot esetén *egy n egyenletből álló* ***lineáris*** *egyenletrendszer*t kell csak megoldanunk a *t*-edik lépésben!

$$U_{t+1}(s_k) \leftarrow R(s_k) + \gamma \max_a \sum_{s'} T(s_k, a, s') \cdot U_t(s'), \quad k = 1, 2, \ldots, K$$

Ha *a*-t rögzítjük, az éppen használt eljárásmód megadja a választandó cselekvést: $a = \pi_t(s)$

$$U_t(s) = R(s) + \gamma \sum_{s'} T(s, \pi_t(s), s') \cdot U_t(s')$$

Szekvenciális döntési probléma

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

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

Eljárásmód-iteráció:

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

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

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 kezdjük a 2.-nél újra

Egy régebbi vizsgafeladat: Egy szekvenciális döntési probléma mindegyik állapotában két cselekvést választhatunk: A1-et vagy A2-őt. A rendszer végállapota *s*4, a leszámítolási tényező 0,8. A választott cselekvéstől függően az alábbi állapotátmenet-valószínűségek jellemzik a rendszert.
(Jelölésmagyarázó példa: a baloldali táblázatban narancssárgával megjelölt cella a T(*s*3,A1,*s*1) valószínűséget tartalmazza, tehát A1 választása esetén a P(*s*3→*s*1) valószínűséget.)

A1 cselekvés esetén P(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,8	0,0	0,0
s2	0,1	0,1	0,1	0,7
s3	0,4	0,4	0,2	0,0
s4	0,0	0,0	0,0	0,0

A2 cselekvés esetén P(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,0	0,8	0,0
s2	0,1	0,2	0,1	0,6
s3	0,2	0,2	0,2	0,4
s4	0,0	0,0	0,0	0,0

Az egyes állapotokban kapható jutalmak, illetve az állapothasznosságok kiinduló becslése:

s	s1	s2	s3	s4
R(s)	-1	-1	-1	5
U_0(s)	0	-1	2	5

Értékiterációt végzünk. Adja meg az első iterációs lépés után az s1 állapot hasznosságának új becslését : U_1(s1)-et!

A1 cselekvés esetén T(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,8	0,0	0,0
s2	0,1	0,1	0,1	0,7
s3	0,4	0,4	0,2	0,0
s4	0,0	0,0	0,0	0,0

A2 cselekvés esetén T(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,0	0,8	0,0
s2	0,1	0,2	0,1	0,6
s3	0,2	0,2	0,2	0,4
s4	0,0	0,0	0,0	0,0

s	s1	s2	s3	s4
R(s)	-1	-1	-1	5
$U_0(s)$	0	-1	2	5

Adja meg az első iterációs lépés után az s1 állapot hasznosságának új becslését: $U_1(s1)$-et!

Ha az **A1** döntést választjuk, akkor a várható hátralévő jutalom a jelenlegi becsléseink alapján:

$U^{(A1)}(s1)=R(s1)+\gamma[T(s1,A1,s1)*U_0(s1)+T(s1,A1,s2)*U_0(s2)+T(s1,A1,s3)*U_0(s3)+T(s1,A1,s4)*U_0(s4)]$

$U^{(A1)}(s1)=-1+0,8*[0,2*0 + 0,8*(-1)+ 0*2+ 0*5]= -1,64$

Ha az **A2** döntést választjuk, akkor a várható hátralévő jutalom a jelenlegi becsléseink alapján:

$U^{(A2)}(s1)=R(s1)+\gamma[T(s1,A2,s1)*U_0(s1) + T(s1,A2,s2)*U_0(s2) + T(s1,A2,s3)*U_0(s3) + T(s1,A2,s4)*U_0(s4)]$

$U^{(A2)}(s1)=-1+0,8*[0,2*0 + 0*(-1)+ 0,8*2+ 0*5]= +0,28$

Tehát ***jelenleg*** az A2 cselekvést fogjuk választani, és az új becsült hasznosság $U1(s1) = U1^{\pi 1(s1)=A2} = +0,28$ lesz.

Alakítsuk át eljárásmód-iterációra: Egy szekvenciális döntési probléma mindegyik állapotában két cselekvést választhatunk: A1-et vagy A2-őt. A rendszer végállapota *s*4, a leszámítolási tényező 0,8. A választott cselekvéstől függően az alábbi állapotátmenet-valószínűségek jellemzik a rendszert.
(Jelölésmagyarázó példa: a baloldali táblázatban narancssárgával megjelölt cella a T(*s*3,A1,*s*1) valószínűséget tartalmazza, tehát A1 választása esetén a P(*s*3→*s*1) valószínűséget.)

A1 cselekvés esetén P(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,8	0,0	0,0
s2	0,1	0,1	0,1	0,7
s3	0,4	0,4	0,2	0,0
s4	0,0	0,0	0,0	0,0

A2 cselekvés esetén P(s→s')				
s\s'	s1	s2	s3	s4
s1	0,2	0,0	0,8	0,0
s2	0,1	0,2	0,3	0,4
s3	0,2	0,2	0,2	0,4
s4	0,0	0,0	0,0	0,0

Az egyes állapotokban kapható jutalmak, illetve az eljárásmód és azállapothasznosságok becslése:

s	s1	s2	s3	s4
R(s)	-1	-1	-4	5
$U_t(s)$	2	3	4	5
$\pi_t(s)$	A1	A1	A2	-

Eljárásmód-iterációt végzünk. Adja meg az következő iterációs lépés után a négy állapot hasznosságának, illetve az eljárásmód új becslését : $U_{t+1}(sk)$ és $\pi_{t+1}(sk)$-t, k=1,2,3,4!

Nem oldottam meg, maradhatott benne rossz adat!