

Alkalmazott mesterséges intelligencia (AMI)

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

11. ea. (2019 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



BME I.E. 414, 463-26-79

pataki@mit.bme.hu,

<http://www.mit.bme.hu/general/staff/pataki>

Eddigiekben egy-egy mintához tartozott egy-egy jó válasz, kívánt eredmény. 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.
- A pénzünket elköltsük-e (mozijegyre, sörre, új nadrágra, biciklire, autóra, házra) vagy fektessük be (X részvénybe, Y részvénybe, államkötvénybe, aranyba...)

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

$S_0 \Rightarrow S_{17} \Rightarrow \bullet \bullet \bullet \Rightarrow S_{103} \Rightarrow S_{53}$

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 vagy 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), ha például a fenti állapotsorozatnál:

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

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

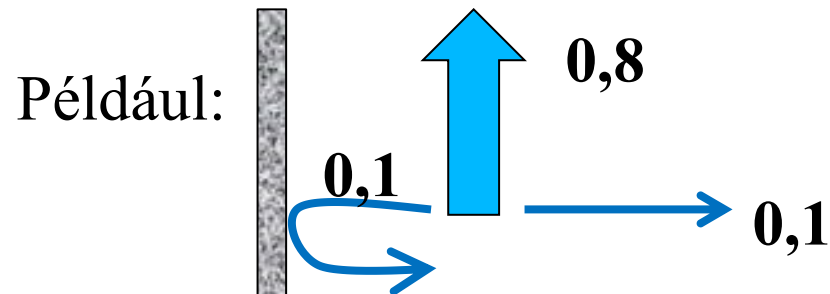
Példa (Russel-Norvig könyv példája):

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 80% valószínűséggel a kívánt irányba megy, 10-10% valószínűséggel valamelyik merőleges irányba. Ha falnak ütközik helyben marad.



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: s_0

KVÍZ következik

Állapotátmenet-modell: $T(s, a, s')$: $s \rightarrow s'$ átmenet
valószínűsége, amikor az a cselekvést választottuk s -ben

Markov tulajdonság: elég az s állapotot ismernem, a megelőző
állapotok nem érdekesek az $s \rightarrow s'$ átmenet szempontjából

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

KVÍZ 11.1.

A JOBBRA cselekvés esetén milyen valószínűség található a kérdőjellel jelölt $(4,1) \rightarrow (4,1)$ cellában?

$P(s \rightarrow s' \mid \text{Cselekvés})$ vagy másképp jelölve $T((4,1), \text{JOBBRA}, (4,1))$

Jobbra

s↓ s'→		(3,3)	(4,1)	(4,2)	
....				...	...
(3,3)				...	...
(4,1)			?	...	...
(4,2)				...	...

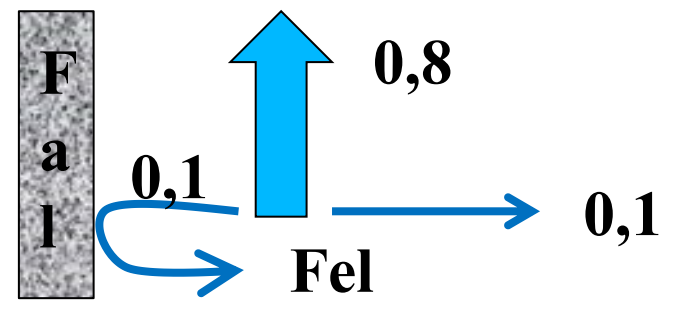
- A. 0,1

B. 0,2

C. 0,8

D. 0,9

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



KVÍZ 11.1.

újabb KVÍZ következik

A JOBBRA cselekvés esetén milyen valószínűség található a kérdőjellel jelölt $(4,1) \rightarrow (4,1)$ cellában?

$P(s \rightarrow s' \mid \text{Cselekvés})$ vagy másképp jelölve $T((4,1), \text{JOBBRA}, (4,1))$

Jobbra

s↓ s'→		(3,3)	(4,1)	(4,2)	
....				...	...
(3,3)				...	...
(4,1)			0,9	...	...
(4,2)				...	...

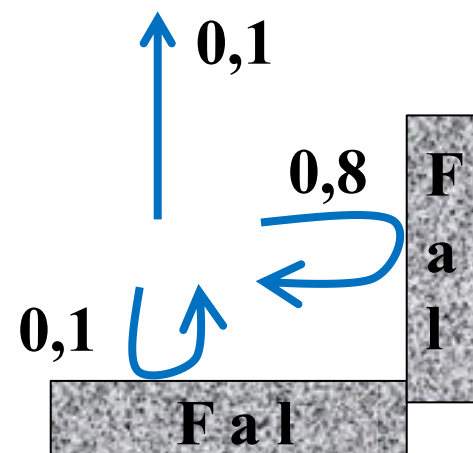
A. 0,1

B. 0,2

C. 0,8

D. 0,9

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



$$T(s, A, s') = P(s \rightarrow s' \mid \text{Cselekvés}=A)$$

$s \downarrow s' \rightarrow$	s1	s2	s3	s4	
s1				...	...
s2				...	...
s3				...	...
....				...	...

Kvíz 11.2.

A $T(s, a, s')$ adott $a=A$ mellett egy kétdimenziós mátrix (táblázat). Erre **a $T(s, A, s')$ mátrixra a következő igaz:**

- A. Az egész mátrixban található összes szám összege mindig 1.**
- B. A mátrix egy-egy oszlopában található számok összege mindig 1.**
- C. A mátrix egy-egy sorában található számok összege mindig 1.**
- D. A fentiek közül egyik sem igaz.**

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.

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

(De általában nem tudjuk biztosan, hogy s_0 -ból milyen lépéssorozattal jutunk el $s_{k_{VÉG}}$ -be!)

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

$$s_0 \rightarrow s_1 \rightarrow \dots \rightarrow s_{k_{VÉG}} \qquad U(s) = \sum_{k=0}^{k_{VÉG}} \gamma^k \cdot R(s_k) \quad ; \quad 0 \leq \gamma \leq 1$$

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

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

Azt például nem tudjuk megmondani, hogy mennyi időbe telik ma , ha 4-es villamossal megyünk haza, de *átlagosan* meg lehet mondani, hogy várhatóan mekkora az átlagos időszükséglet. (Persze itt az eltelt idő 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 várható 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 a cselekvés az adott állapotban, ami maximalizálja a hátralévő jutalmat.

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

Példa:

KVÍZ következik

Kezdőállapot: s11

Végállapotok: s6 és s8 – csak itt van jutalom, másutt 0.

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

Az s1, s2, s3, s4, s5, s12, s13, s14, s15 állapotokban csak a 'Jobbra' cselekvést választhatjuk.

Az s11 kezdőállapotban mindkettőt választhatjuk, de csak bizonyos valószínűséggel teljesül a választott cselekvés, a többi állapotban 100%-ban az történik, amit „akarunk” ('Fel', illetve 'Jobbra').

$$T(s11, a='Fel', s9)=0,8$$

$$T(s11, a='Fel', s12)=0,2$$

$$T(s11, a='Jobbra', s9)=0,1$$

$$T(s11, a='Jobbra', s12)=0,9$$

s1	s2	s3	s4	s5	s6, R(6)=+10
s7	x	x	x	x	s8, R(8)=-4
s9	x	x	x	x	s10
s11	s12	s13	s14	s15	s16

Kvíz 11.3. Mekkora lesz s11 hasznossága – U(11)?

- A. 10
- B. 6
- C. 7,2
- D. 4,4

Kezdőállapot: s11

Végállapotok: s6 és s8 – csak itt van jutalom, másutt 0.

$T(s11,a='Fel',s9)=0,8$

$T(s11,a='Fel',s12)=0,2$

$T(s11,a='Jobbra',s9)=0,1$

$T(s11,a='Jobbra',s12)=0,9$

s1	s2	s3	s4	s5	s6, R(6)=+10
s7	x	x	x	x	s8, R(8)=-4
s9	x	x	x	x	s10
s11	s12	s13	s14	s15	s16

**Mekkora lesz
s11 hasznossága
– U(11)?**

s1	s2	s3	s4	s5	s6, R(6)=+10
s7	X	X	X	X	s8, R(8)=-4
s9	X	X	X	X	s10
s11	s12	s13	s14	s15	s16

Tegyük fel, hogy a '**Fel**' cselekvést választjuk s11-ben,

- ha jó irányba indultunk, utána s6-ba navigálhatjuk magunkat,
- ha s12-be jutottunk, akkor nem tudunk mit csinálni, előbb-utóbb s8-ba jutunk.

A várható nyereség ez esetben: $0,8*10-0.2*4=+7,2$

Ha a '**Jobbra**' cselekvést választjuk s11-ben,

- ha mégis felfele indultunk, utána s6-ba navigálhatjuk magunkat,
- ha s12-be jutottunk, akkor nem tudunk mit csinálni, előbb-utóbb s8-ba jutunk.

A várható nyereség ez esetben: $0,1*10-0.9*4= -2,6$

Rajtunk áll, hogy milyen cselekvést választunk s11-ben: a várható
hátralévő összjutalom (az s11 hasznossága) **U(11)=+7,2**

Kvíz 11.4.

Egy vetélkedőben a következő választás elé állítanak: 4 fiók van előttük, mindegyikben egy adott összeg. Az elsőben 8.000 Ft van, ezt megmondják. A másik háromban ismeretlen sorrendben 6.000 Ft, 9.000 Ft, 21.000 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. 12.000 Ft**

Kvíz 11.4. Egy vetélkedőben a következő választás elé állítanak: 4 fiók van előttük, mindegyikben egy adott összeg. Az elsőben 8.000 Ft van, ezt megmondják. A másik háromban ismeretlen sorrendben 6.000 Ft, 9.000 Ft, 21.000 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. 12.000 Ft

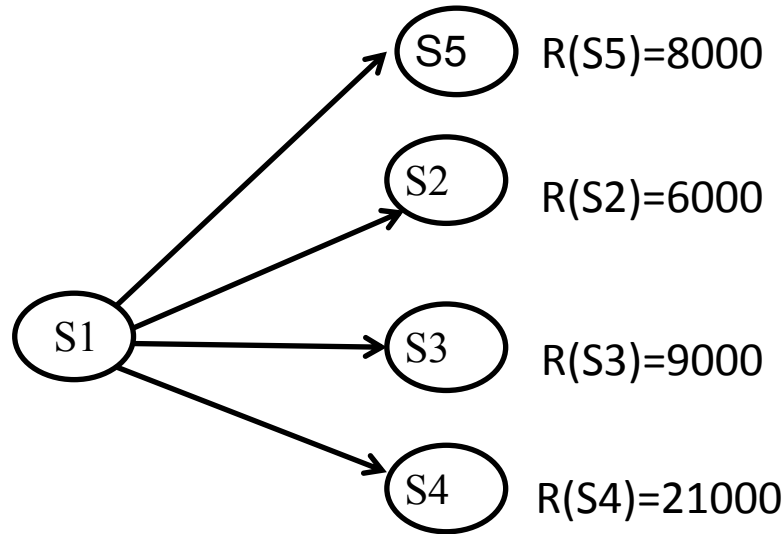
a) Ha az 1-es fiókot választom, mindig **8.000 Ft**-t kapok.

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

$$\begin{aligned} \text{Várható Jutalom} &= \sum_k p(k) \cdot \text{Jutalom}(k) = \\ &= 6000 \cdot \frac{1}{3} + 9000 \cdot \frac{1}{3} + 21000 \cdot \frac{1}{3} = 2000 + 3000 + 7000 = \mathbf{12000} \end{aligned}$$

Tehát a jó döntés itt a kockáztatás, mert átlagban (várható érték) többet nyerhetek!

Az előző (vetélkedős) példa:



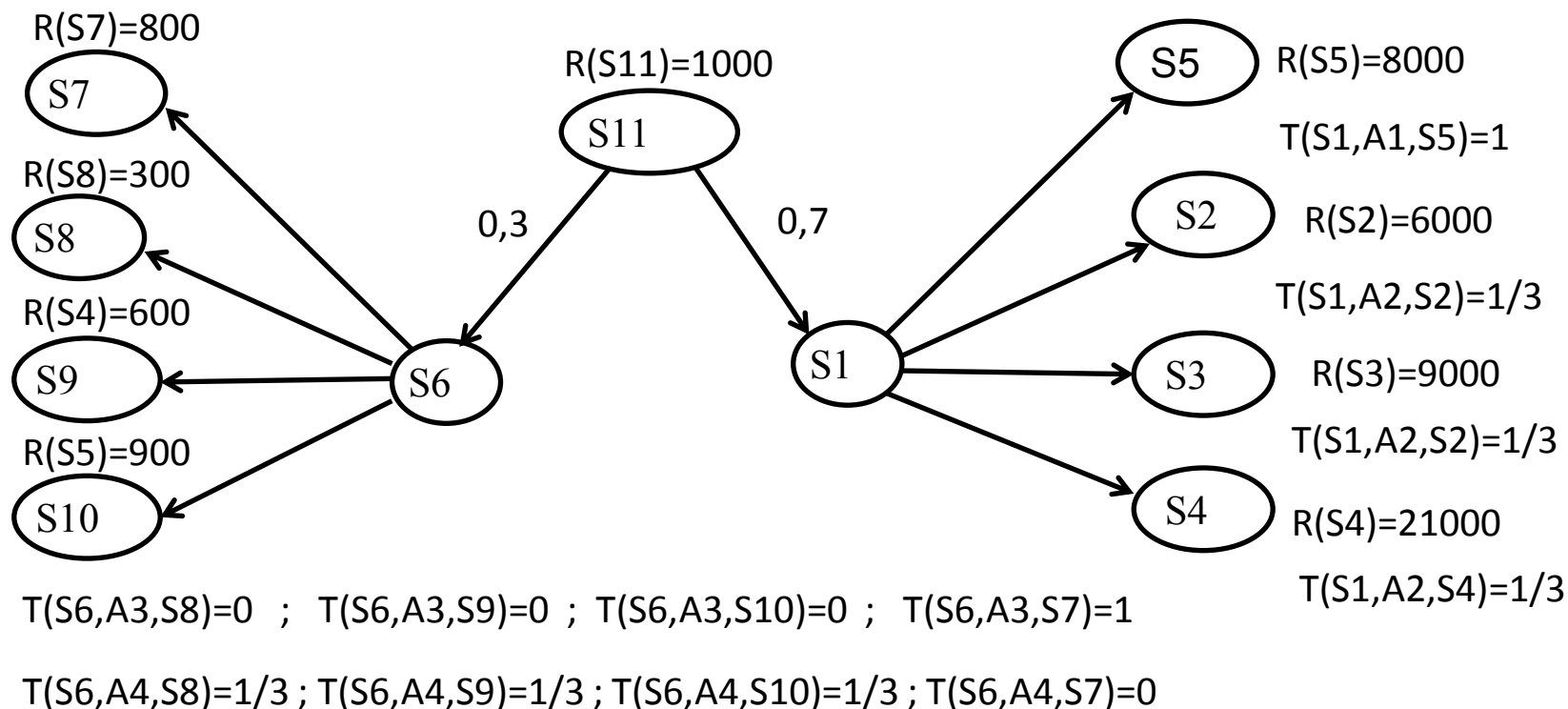
A1: $T(S1, A1, S2)=0$; $T(S1, A1, S3)=0$; $T(S1, A1, S4)=0$; $T(S1, A1, S5)=1$

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

$$\pi^*(S1) = A2$$

$$U(S1) = 12000$$

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 (A3 és A4 cselekvés, valószínűségeket lásd lent):
 S7→800 Ft, S8→300Ft, S9→600 Ft, S10→900 Ft. **Kvíz következik!**



11.5. Kvíz

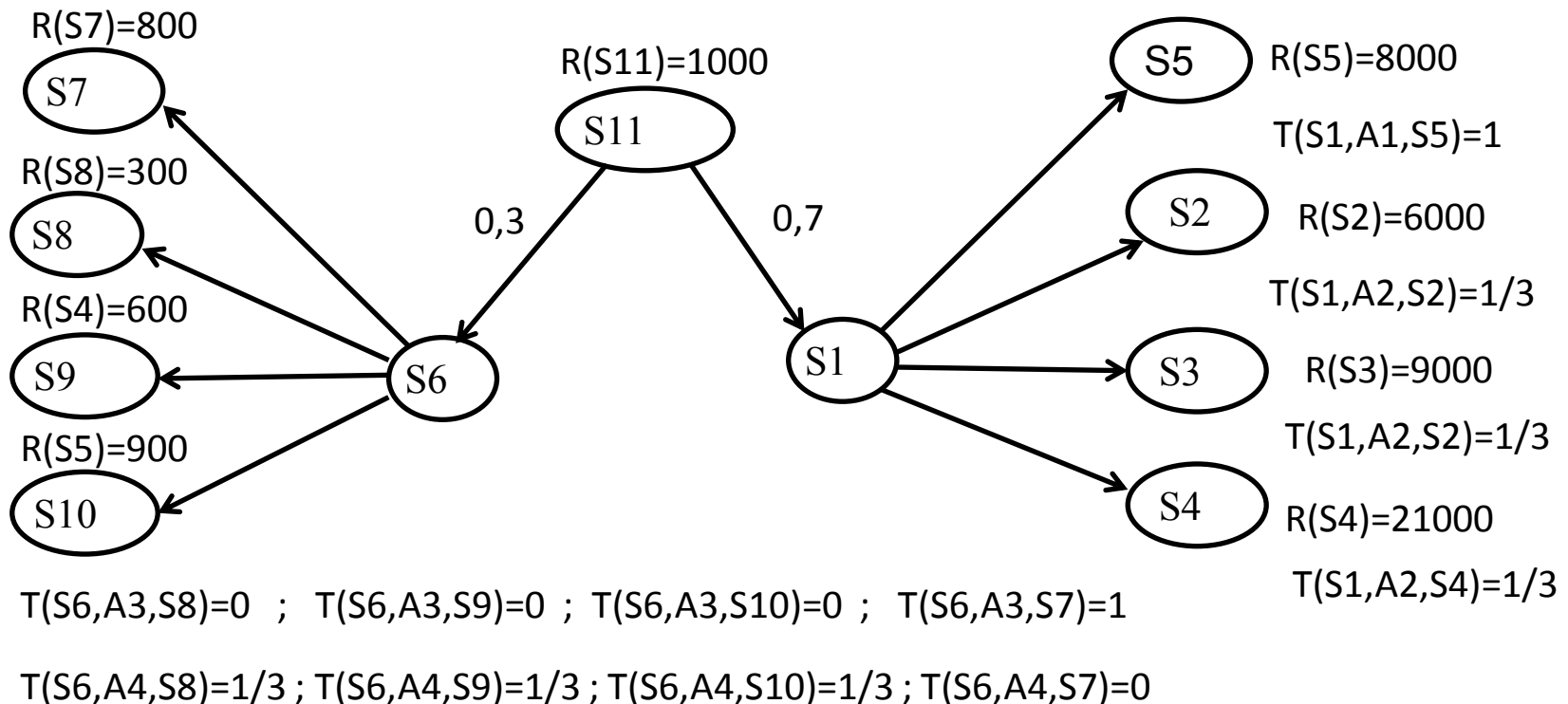
Mekkora az S11 állapot hasznossága?

A. 8400

B. 8640

C. 9640

D. 12000

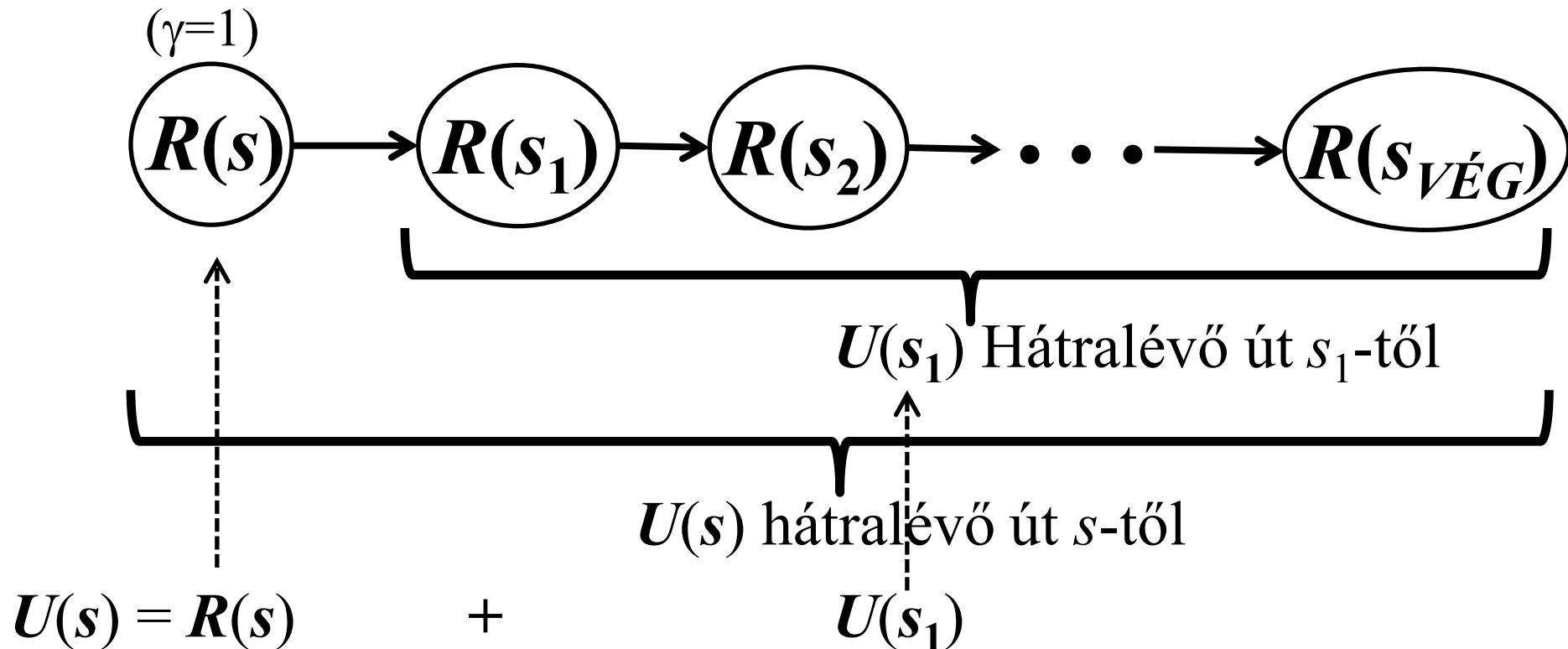


Fontos egyszerűsítés volt: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság additív)

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

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

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

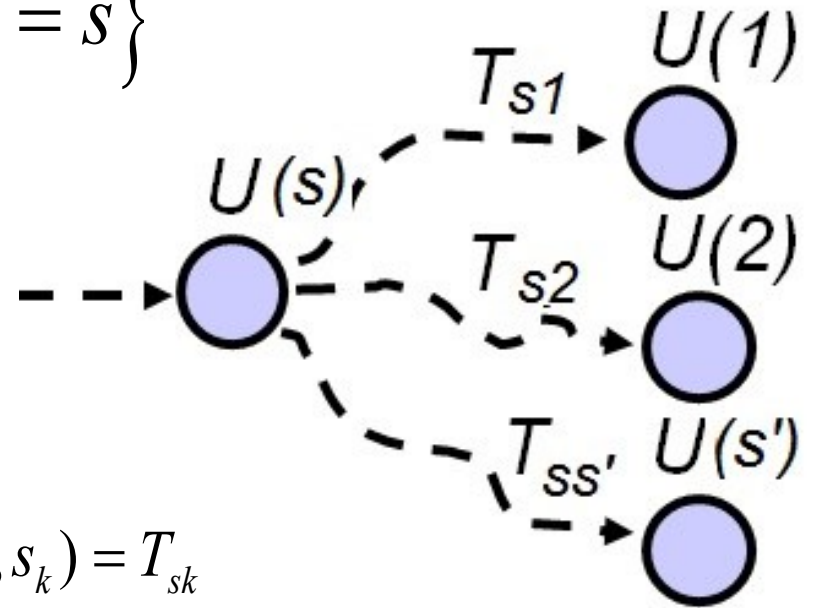


Fontos egyszerűsítés volt: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság additív)

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

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

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



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

$$U^{\pi}(s) = R(s) + \gamma T(s, s_1) \cdot U^{\pi}(s_1) + \dots + \gamma T(s, s') \cdot U^{\pi}(s')$$

Véges horizontú probléma: Egy adott 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.

Melyik probléma esetén könnyebb jó eljárásmodot találni, a véges vagy a végtelen horizontú esetén?

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. 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 .



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

s_0 - kiinduló (start) állapot, $s \in S$ – állapottér elemei

$T(s, a, s')$ –áll. átmenet 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 eljárásmodot ismerjük, DE az optimális eljárásmodot csak akkor tudjuk meghatározni, ha $U(s_k)$ -kat ismerjük! ***Nemlineáris (mindegyik egyenletben van max) egyenletrendszert kéne megoldani!***

Szekvenciális döntési probléma

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

Értékiteráció: *Konvergens eljárás!*

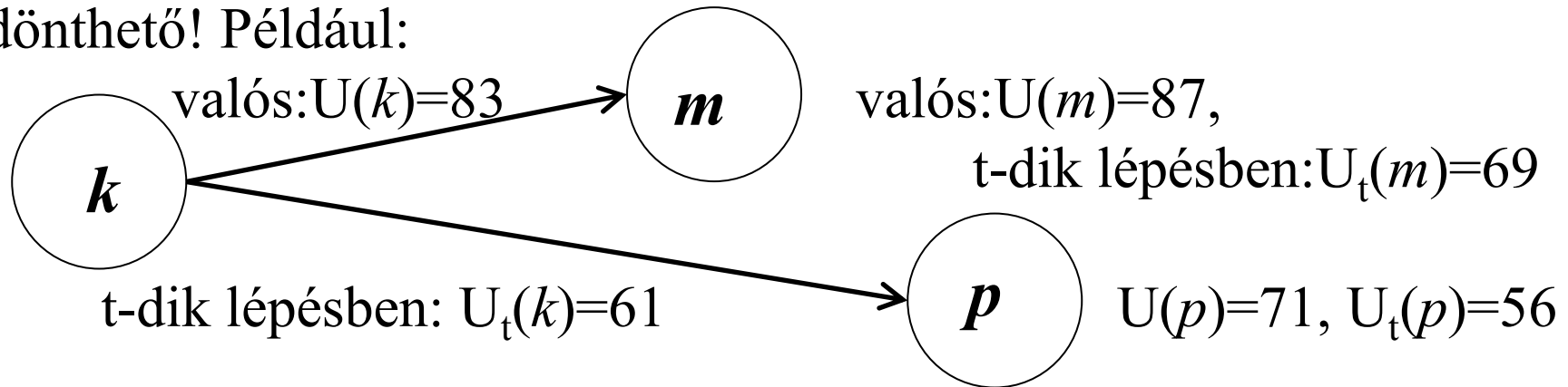
1. $t=0$ - valamilyen kiinduló hasznosságfüggvény $U_0(s_k)$,
 $k=1, \dots, K$
2. $U_{t+1}(s_k)$ -k meghatározása (nem oldjuk meg az egyenletrendszert, csak kiszámítunk mindegyikből egy-egy új $U(s)$ értéket, a max-ot persze használjuk!)

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

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:



Már látható a t -dik lépésben is, hogy a $k \rightarrow m$ átmenetet eredményező cselekvések a jobbak, pedig az $U_t(\cdot)$ értékek még messze vannak a valóstól!

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

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

Szekvenciális döntési probléma

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

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

Eljárásmód-iteráció:

1. $t=0$ - valamilyen kiinduló eljárás mód $\pi_0(s)$ (minden s -re)
2. $U_t(s)$ 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)
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