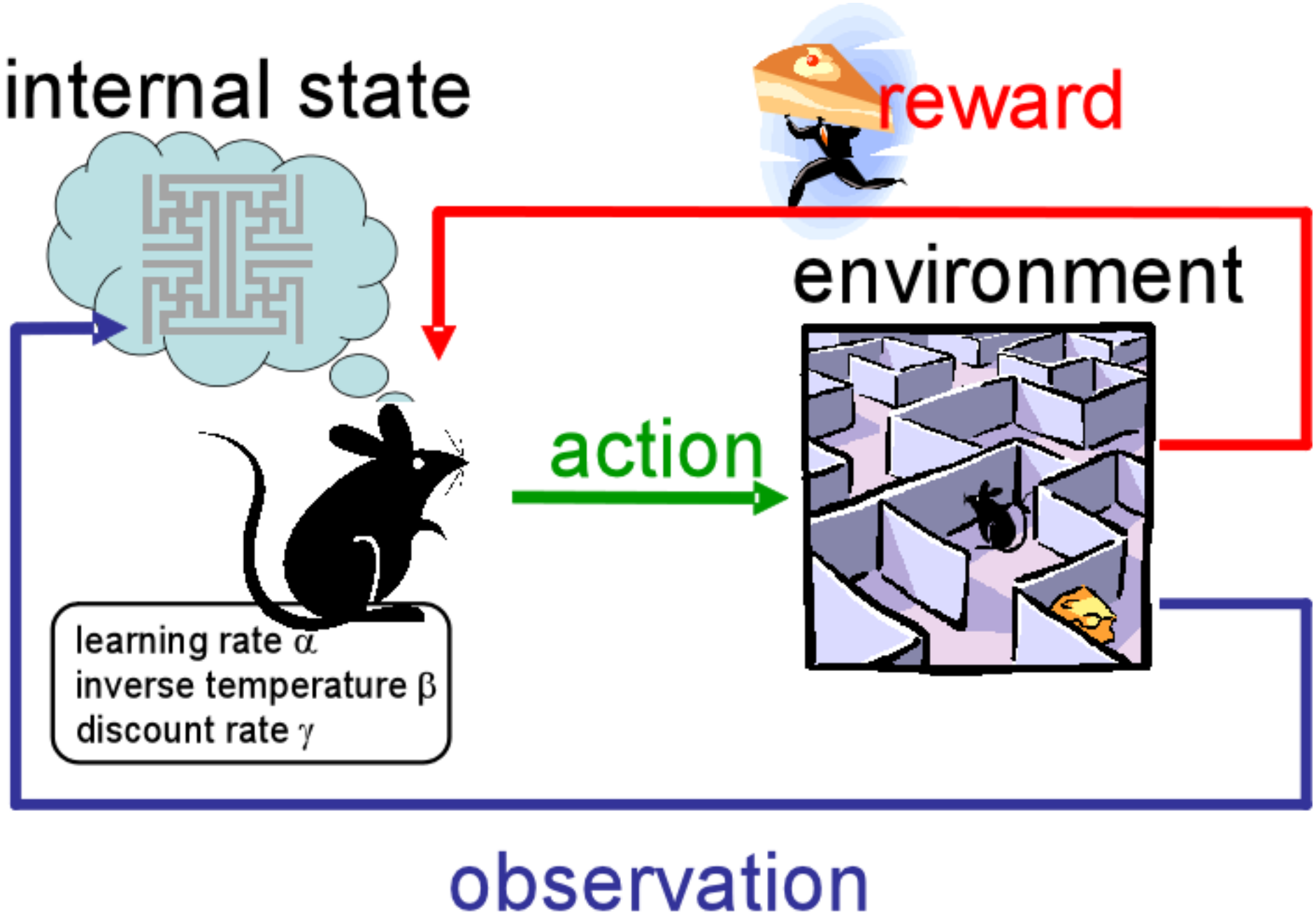
Mesterséges Intelligencia MI

Megerősítéses tanulás

Pataki Béla

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


internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

Ágens tudása:

Induláskor: vagy ismeri már a **környezetet** és a **cselekvéseinek hatását**, vagy pedig még ezt is **meg kell tanulnia**. Ez utóbbi a megerősítéses tanulás területe.

Megerősítés: lehet csak a **végállapotban**, de lehet menet közben **bármelyikben** is.

Ágens:

passzív tanuló: figyeli a világ alakulását és tanul, *előre rögzített eljárásmód* (ettől passzív), nem mérlegel, előre kódoltan cselekszik

aktív tanuló: a megtanult információ birtokában az eljárásmódot is alakítja, optimálisan cselekednie is kell (**felfedező** …)

Ágens tanítása: megerősítés → hasznosság leképezés

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

γ - leszámítolási tényező

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

Két lehetőség:

U(s) **hasznosság**függvény tanulása, ennek alapján a cselekvések eldöntése úgy, hogy az elérhető hasznosság várható értéke maximális legyen (**környezet**/ágens modellje **szükséges,** környezet: **T(*s*,*a*,*s*')**)

Q(a, s) **cselekvésérték** függvény (állapot-cselekvés párok) tanulása, valamilyen várható hasznot tulajdonítva egy adott helyzetben egy adott cselekvésnek (**környezet**/ágens modell **nem szükséges**) – **Q tanulás (model-free)**

Passzív megerősítéses tanulás

Markov döntési folyamat (MDF), szekvenciális döntés

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

Passzív tanulás esetén az ágens **π(s) stratégiája rögzített**, az s állapotban mindig az $a = \pi(s)$ cselekvést hajtja végre.
A cél egyszerűen a stratégia jóságának – tehát az $U^{\pi}(s)$ hasznosságfüggvénynek – a megtanulása.
Fontos különbség a Markov döntési folyamathoz, szekvenciális döntéshez képest, hogy a passzív ágens **nem ismeri** az **állapotátmenet-modellt** (**transition model**), a $T(s, a, s')$-t, amely annak a valószínűséget adja meg, hogy az a cselekvés hatására az s állapotból az s' állapotba jutunk, továbbá nem ismeri a **jutalomfüggvény**t (**reward function**), $R(s)$-et, amely minden állapothoz megadja az ott elnyerhető jutalmat.

Egyszerűbb jelölés:

$$T(s_k, \pi(s_k), s_m) \quad \text{helyett } T_{km} = P(s_k \to s_m \,|\, a = \pi(s_k))$$

Passzív tanulás - Közvetlen hasznosságbecslés

Kvíz

Az állapotok hasznosságát a definíció (hátralevő jutalom várhatóértéke, praktikusan átlaga) alapján tanuljuk.

Sok kísérletet végzünk, és feljegyezzük, hogy amikor az *s* állapotban voltunk, akkor mennyi volt a végállapotig gyűjtött jutalom, amit kaptunk. Ezen jutalmak átlaga – ha nagyon sok kísérletet végzünk – az U(s)-hez konvergál.

Viszont nem használja ki a Bellman egyenletben megragadott összefüggést az állapotok közt, ezért lassan konvergál, nehezen tanul!

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

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

Kvíz

Az egyszerű közvetlen tanulásba hogy kerül bele az állapotátmenet-valószínűség?

A. A gyakori (valószínű) átmenetek gyakrabban fognak előfordulni a sorozatokban

B. Nem kerül bele, ezért lassú a tanulás

C. A tanulás során figyeljük, és a végén súlyozunk a valószínűségekkel

D. A leszámítolási tényezőn keresztül

Passzív tanulás - Adaptív Dinamikus Programozás

Az állapotátmeneti valószínűségek $T(s_k,s_m)= T_{km}$ a megfigyelt gyakoriságokkal becsülhetők.
Amint az ágens megfigyelte az összes állapothoz tartozó jutalom értéket, és elég tapasztalatot gyűjtött az állapotátmenetek gyakoriságáról, a hasznosság-értékek a következő **lineáris egyenletrendszer megoldásával** kaphatók meg:

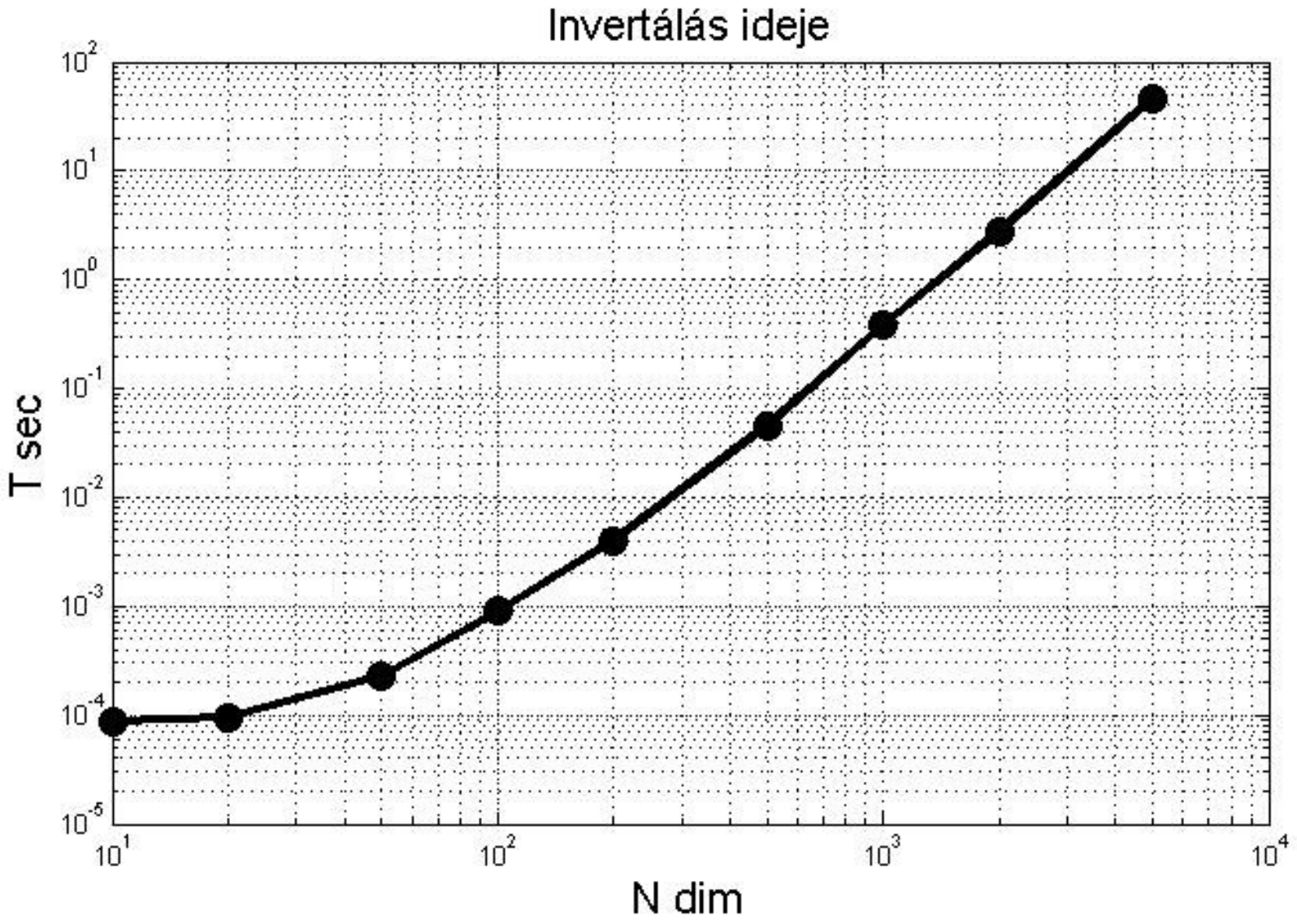
$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s,s') \cdot U^{\pi}(s')$$

megtanult — megfigyelt

$$\begin{pmatrix} U(1) \\ U(2) \\ U(3) \\ \dots \\ U(N) \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(N) \end{pmatrix} + \gamma \begin{pmatrix} T_{11} & T_{12} & T_{13} & \dots & T_{1N} \\ T_{21} & T_{22} & T_{23} & \dots & T_{2N} \\ T_{31} & T_{32} & T_{33} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ T_{N1} & T_{N2} & \dots & \dots & T_{NN} \end{pmatrix} \begin{pmatrix} U(1) \\ U(2) \\ U(3) \\ \dots \\ U(N) \end{pmatrix}$$

$$\mathbf{U} = \mathbf{R} + \gamma\, \mathbf{T}\mathbf{U} \qquad \mathbf{U} = (\mathbf{I} - \gamma \mathbf{T})^{-1}\mathbf{R}$$

….és ha nem megy?



A fontosak nem a konkrét értékek, hanem a jelleg!
(logaritmikus skála!)

Passzív tanulás - Időbeli különbség (IK) tanulás (TD – Time Difference)

Ha csak az aktuális jutalmat néznénk: $U(s) \leftarrow U(s) + \alpha(R^{\circ}, \quad U(s))$

Ha a teljes megfigyelt hátralévő jutalomösszeget használjuk – ki kell várni az epizód végét.

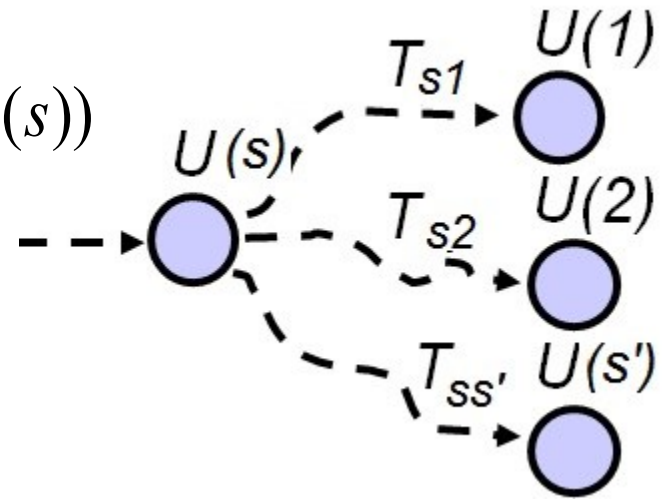
Alapötlet: a hátralevő összjutalom helyett használjuk fel a megfigyelt állapotátmeneteknél a hasznosság aktuálisan becsült értékét:

A jutalom becslése: $\hat{U}(s) = R(s) + \gamma U(s')$

Az előző becsléstől való eltérés: $\delta(s) = \hat{U}(s) - U(s) = R(s) + \gamma U(s') - U(s)$

Frissítés: $U(s) \leftarrow U(s) + \alpha\,\delta(s) =$

$$U(s) + \alpha(R(s) + \gamma U(s') - U(s))$$



α - **bátorsági faktor,** tanulási tényező

γ - **leszámítolási** tényező (ld. MDF)

Az egymást követő állapotok hasznosság-különbsége **időbeli különbség (IK)** .

Az összes időbeli különbség eljárásmód alapötlete

1. korrekt hasznosság-értékek esetén **lokálisan fennálló feltételek rendszere**: $U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s,s')U^{\pi}(s')$

   Ha egy adott s→s' átmenetet tapasztalunk a jelenlegi sorozatban, akkor úgy vesszük, mintha T(*s*,*s*')=1 lenne, tehát fenn kéne álljon. $U^{\pi}(s) = R(s) + \gamma\, U^{\pi}(s')$

2. Ebből a **módosító egyenlet**, amely a becsléseinket ezen **„egyensúlyi" egyenlet irányába** módosítja:

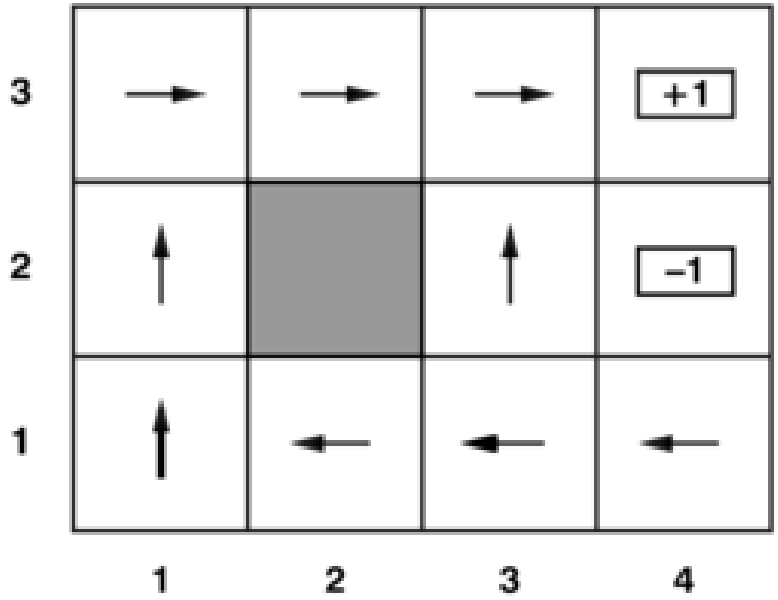
$$U(s) \leftarrow U(s) + \alpha(R(s) + \gamma U(s') - U(s))$$

Az **ADP** módszer felhasználta a **modell teljes** ismeretét.
Az **IK** a szomszédos (egy lépésben elérhető) állapotok közt fennálló kapcsolatokra vonatkozó információt használja, de csak azt, amely az **aktuális tanulási sorozatból származik**.
Az **IK** változatlanul **működik ismeretlen környezet esetén is**. (Nem használtuk ki a T(s,s') ismeretét!)

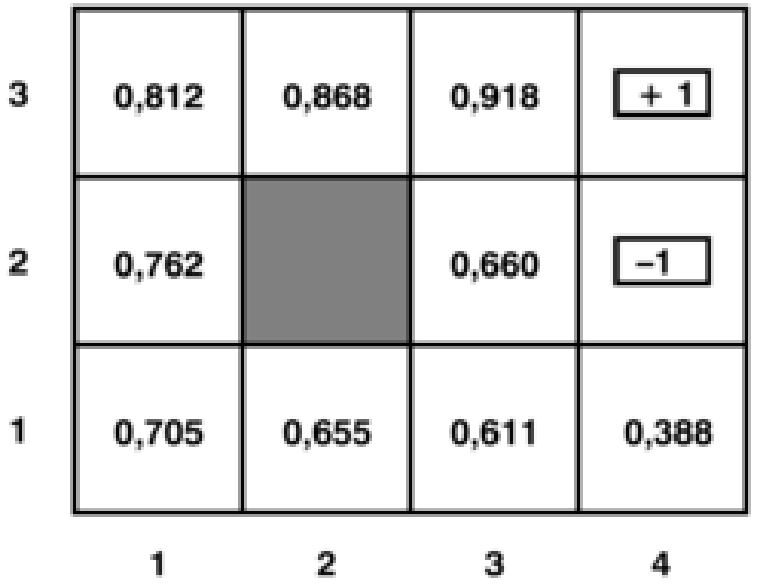
Egy egyszerű demóprobléma

- végállapotok: (4,2) és (4,3)
- γ=0
- minden állapotban, amelyik nem végállapot R(s)= - 0,04
  (pl. így lehet modellezni a lépésköltséget)

0,8 valószínűséggel a szándékolt irányba lép, 0,2 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

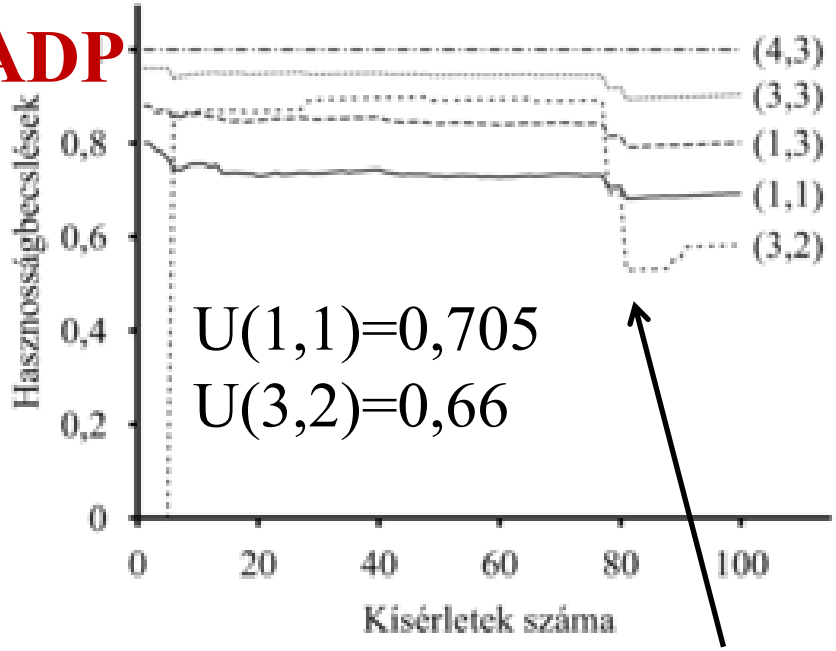


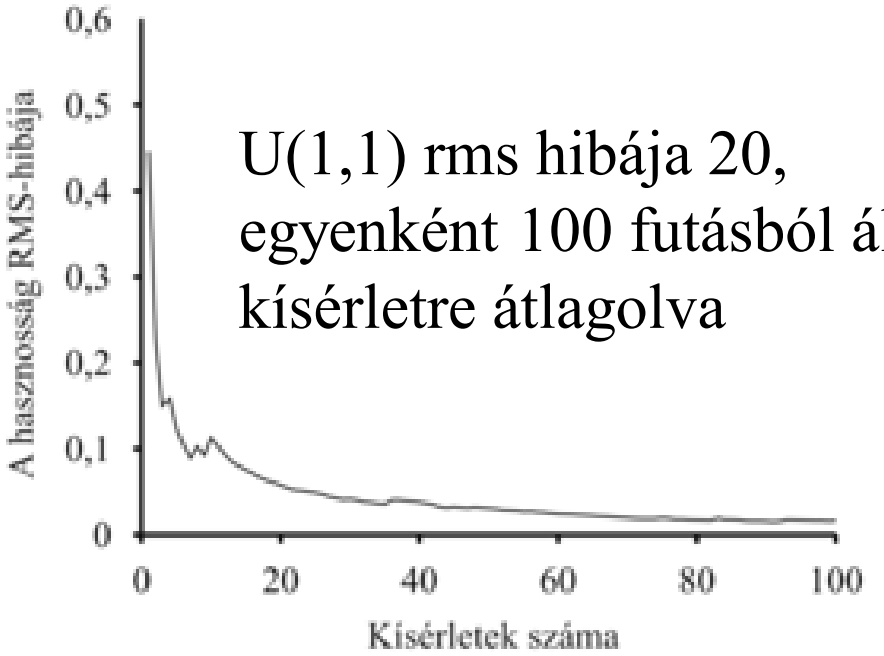
Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok

ADP

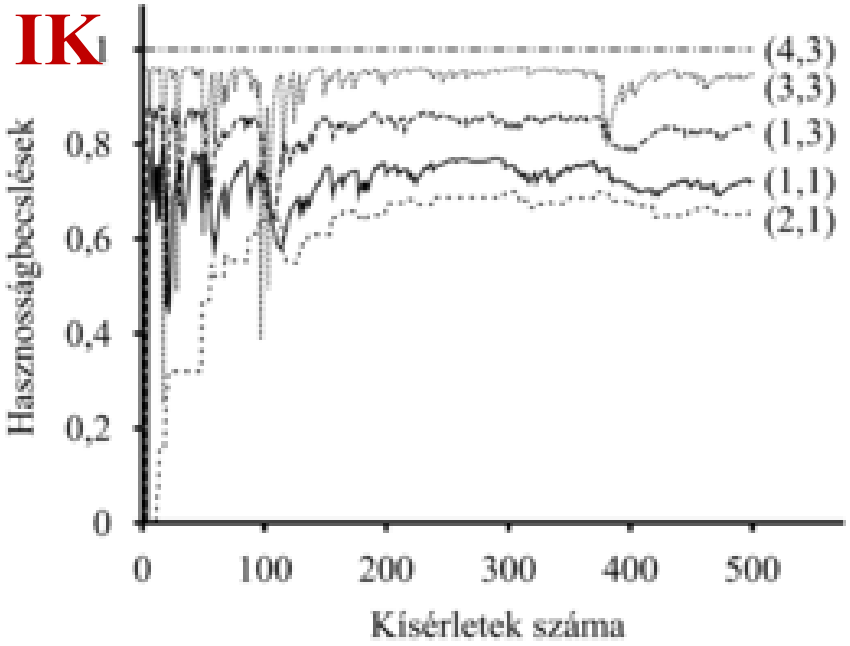


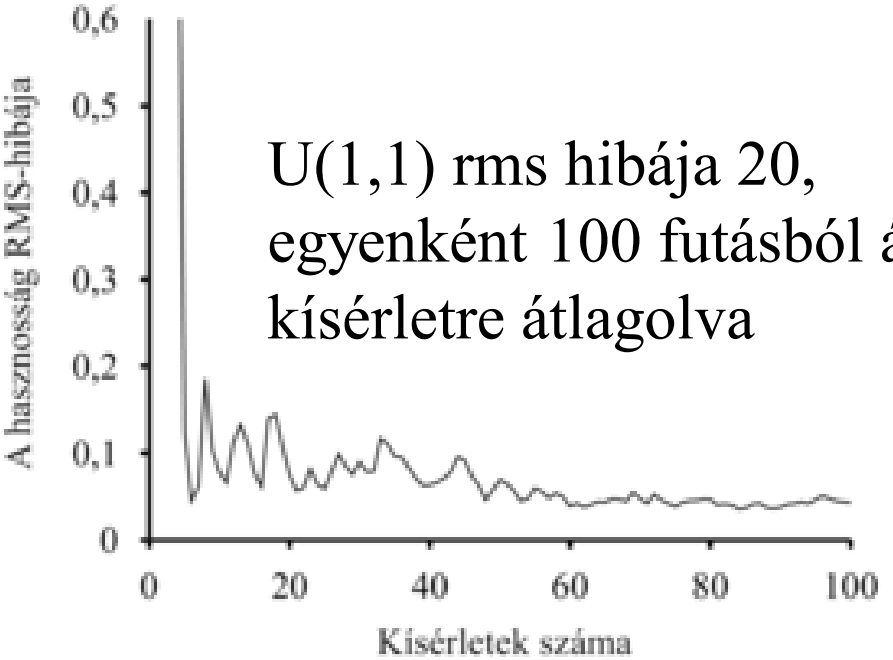


U(1,1) rms hibája 20, egyenként 100 futásból álló kísérletre átlagolva

Egyetlen 100-as sorozat, a 78-diknál jut először a -1-be

IKI





U(1,1) rms hibája 20, egyenként 100 futásból álló kísérletre átlagolva

Ismeretlen környezetben végzett **aktív megerősítéses tanulás**

Döntés: melyik cselekvés? a cselekvésnek mik a kimeneteleik?
hogyan hatnak az elért jutalomra?

A környezeti modell: a többi állapotba való **átmenet valószínűsége egy adott cselekvés esetén**.

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

A feladat kettős: az állapottér felderítése és a jutalmak begyűjtése

- milyen cselekvést válasszunk?
- a két cél gyakran ellentmondó cselekvést igényel

Nehéz probléma (exploration-exploitation, felfedezés-kiaknázás)

Helyes-e azt a cselekvést választani, amelynek a jelenlegi hasznosságbecslés alapján legnagyobb a közvetlen várható hozama?

Ez figyelmen kívül hagyja a **cselekvésnek a tanulásra gyakorolt hatását**!

Az állapottér felderítése

Kvíz

A döntésnek kétféle hatása van:

1. **Jutalma(ka)t** eredményez a **jelenlegi** szekvenciában.
2. Befolyásolja az észleléseket, és ezáltal az ágens **tanulási képességét –** így jobb jutalmakat eredményezhet a **jövőbeni szekvenciákban.**

Két szélsőséges megközelítés a cselekvés kiválasztásában:
„**Hóbortos, Felfedező**”: véletlen módon cselekszik, annak reményében, hogy végül is felfedezi az egész környezetet
„**Mohó**”: a jelenlegi becslésre alapozva maximalizálja a hasznot.

Kvíz

Mi a probléma a mohó megközelítéssel?

A. Könnyen ragadhat egy szuboptimális megoldásba

B. Nem jól hasznosítja a megszerzett ismereteit

C. Lassan (de biztosan) konvergál a helyes eljárásmódhoz

D. Lassan (de biztosan) konvergál a pontos hasznosságértékekhez

Az állapottér felderítése

A döntésnek kétféle hatása van:

1. **Jutalma(ka)t** eredményez a **jelenlegi** szekvenciában.
2. Befolyásolja az észleléseket, és ezáltal az ágens **tanulási képességét** – így jobb jutalmakat eredményezhet a **jövőbeni szekvenciákban.**

Kompromisszumot keresünk: a **jelenlegi jutalom**, amit a pillanatnyi hasznosság-becslés tükröz, és a **hosszú távú előnyök** közt.

Két szélsőséges (kompromisszumképtelen) megközelítés a cselekvés kiválasztásában:

„**Hóbortos, Felfedező**”: véletlen módon cselekszik, annak reményében, hogy végül is felfedezi az egész környezetet

„**Mohó**”: a jelenlegi becslésre alapozva maximalizálja a hasznot.

Hóbortos: képes jó hasznosság-becsléseket megtanulni az összes állapotra. Sohasem sikerül fejlődnie az optimális hozam elérésében.

Mohó: gyakran talál egy *viszonylag jó* utat. Utána ragaszkodik hozzá, és soha nem tanulja meg a többi állapot hasznosságát, a legjobb utat.

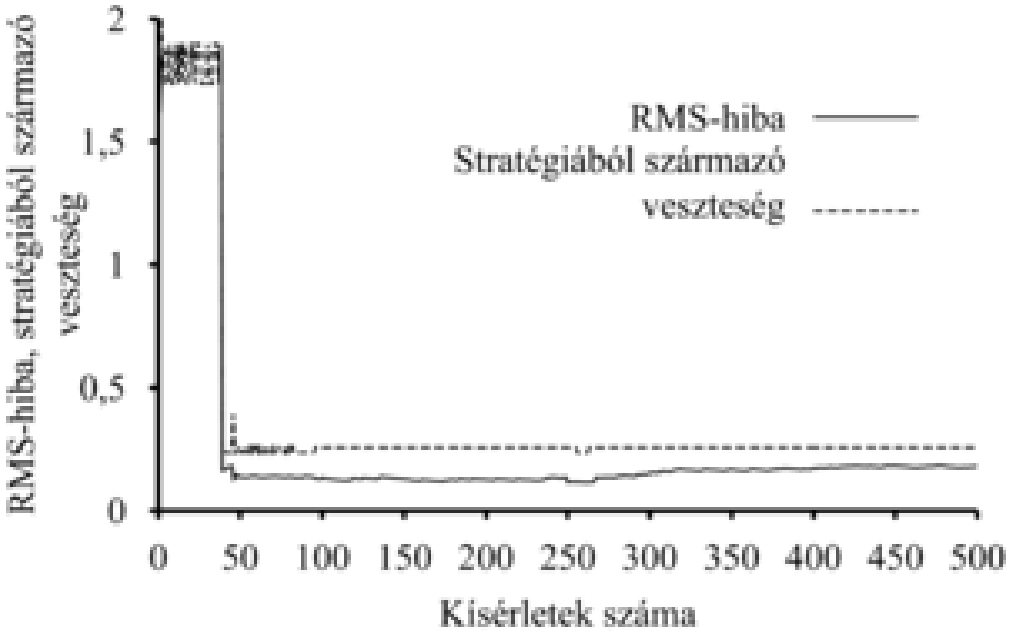
Hóbortos: képes jó hasznosság-becsléseket megtanulni az összes állapotra. Sohasem sikerül fejlődnie az optimális jutalom elérésében.

Mohó: gyakran talál egy viszonylag jó utat. Utána ragaszkodik hozzá, és soha nem tanulja meg a többi állapot hasznosságát, a legjobb utat.

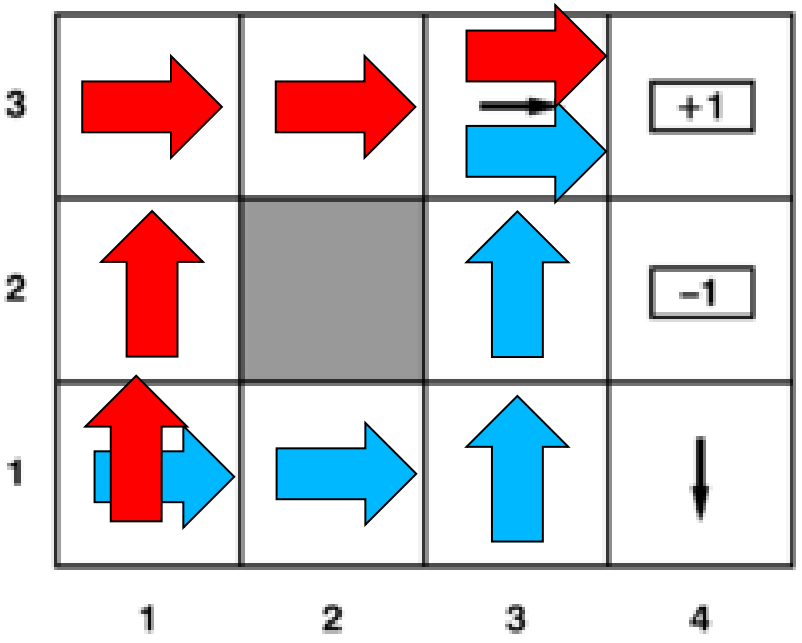
A **mohó ágens** által talált stratégia ⟶

A mohó ágens által talált optimális út

A valódi optimális út



(a)



(b)

A **mohó** és **hóbortos** eljárásmód tipikus eredményei

Felfedezési stratégia

Az ágens addig legyen hóbortos, amíg kevés fogalma van a környezetről, és legyen mohó, amikor a valósághoz közeli modellel rendelkezik.

Létezik-e optimális felfedezési stratégia?

Az ágens

- előnybe kell részesítse azokat a cselekvéseket, amelyeket még nem nagyon gyakran próbált,
- a már elég sokszor kipróbált és kis hasznosságúnak gondolt cselekvéseket pedig kerülje el, ha lehet.

1. Felfedezési stratégia - <u>felfedezési függvény</u>

$$f(u,n) = \begin{cases} R^+ & \text{ha} \quad n < N_e \\ u & \text{különben} \end{cases}$$

$$U^+(s) \leftarrow R(s) + \gamma \max_a f(\sum_{s'} T(s,a,s') \cdot U^+(s'), N(a,s))$$

$U^+(i)$: az i állapothoz rendelt hasznosság ***optimista*** becslése
$N(a,i)$: az i állapotban hányszor próbálkoztunk az a cselekvéssel

R^+ a tetszőleges állapotban elérhető legnagyobb jutalom *optimista* becslése

Az a cselekvés, amely felderítetlen területek *felé* vezet, nagyobb súlyt kap.

mohóság ↔ kíváncsiság

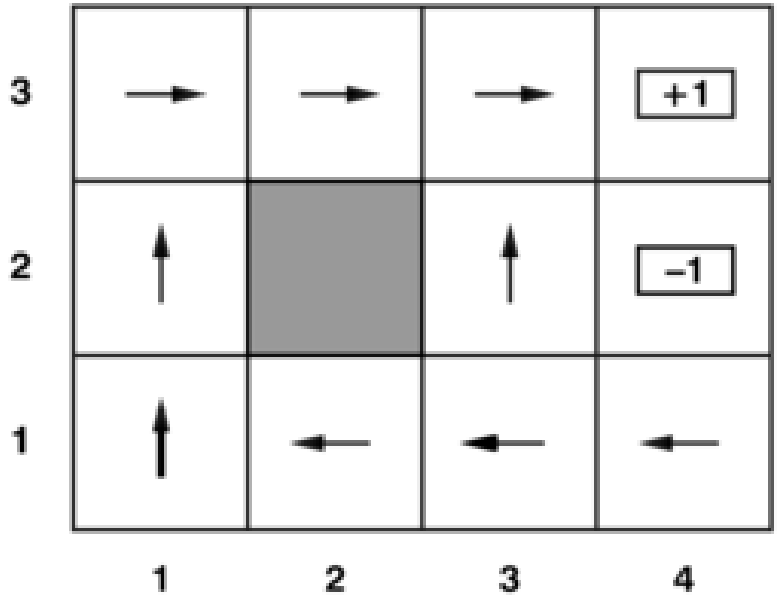
$f(u,n)$: u-ban monoton növekvő (mohóság),
n-ben monoton csökkenő (a próbálkozások számával csökkenő kíváncsiság)

Ismételjük meg a régi fóliát

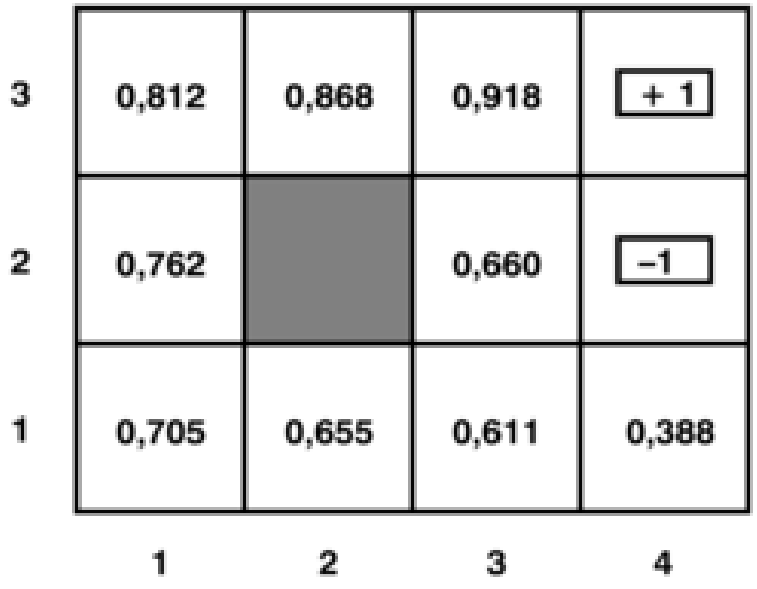
Egy egyszerű demóprobléma:

- végállapotok: (4,2) és (4,3)
- $\gamma=0$
- minden állapotban, amelyik nem végállapot R(s)= - 0,04
  (pl. így lehet modellezni a lépésköltséget)

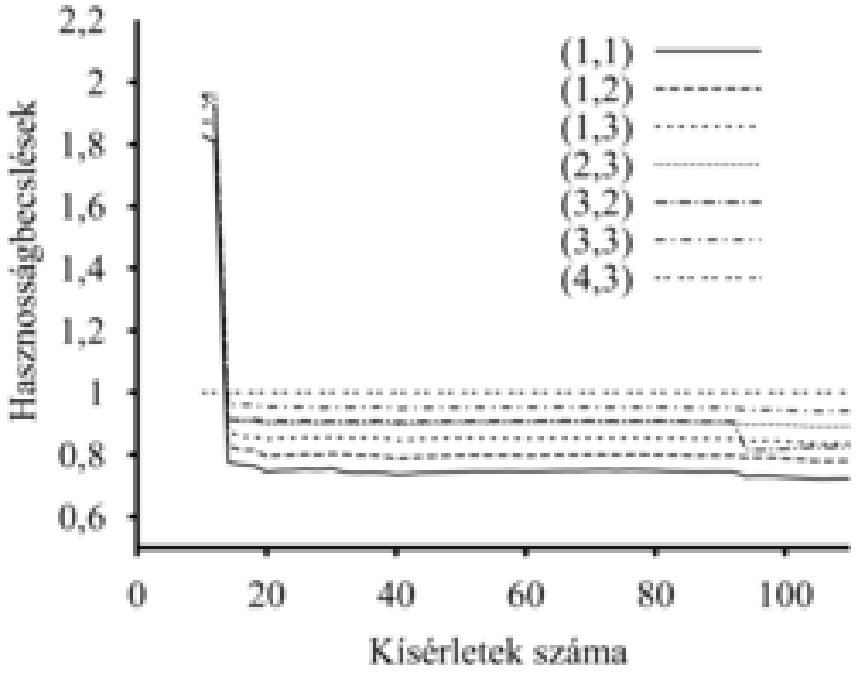
0,8 valószínűséggel a szándékolt irányba lép, 0,2 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

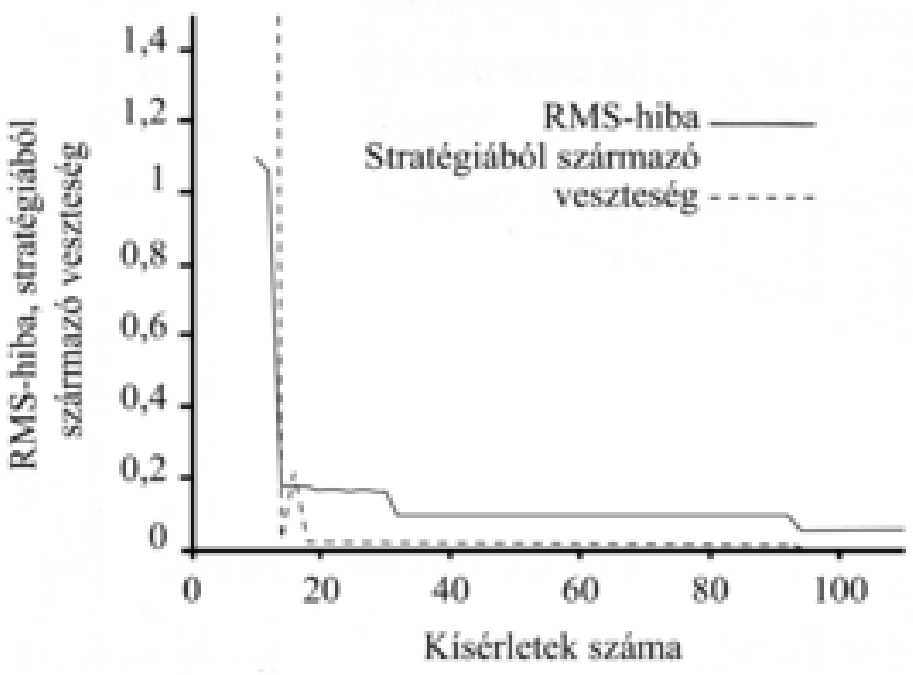


Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok





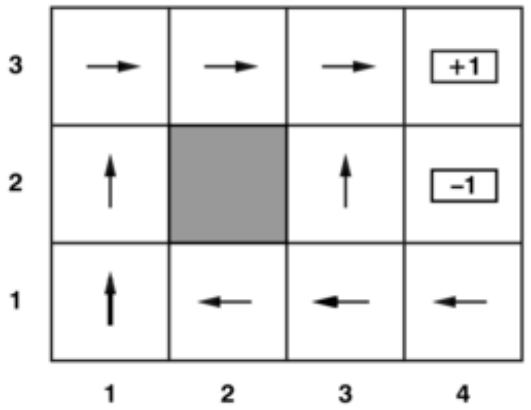
(b)

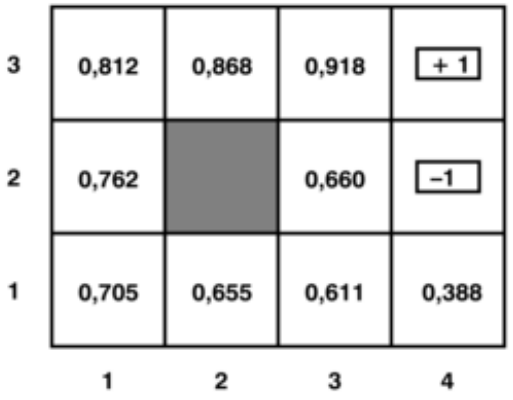
Felfedezési függvény: $R^+=2$, $N_e=5$

$$U^+(s) \leftarrow R(s) + \gamma \max_a f(\sum_{s'} T(s,a,s') \cdot U^+(s'), N(a,s))$$

$$f(u,n) = \begin{cases} R^+ & \text{ha} \quad n < N_e \\ u & \text{különben} \end{cases}$$

Valós hasznosságok, opt. stratégia





2. Felfedezési stratégia - <u>ε-mohóság</u> (*N* cselekvési lehetősége van)

- 1-ε valószínűséggel mohó, tehát a legjobbnak gondolt cselekvést választja ilyen valószínűséggel
- az ágens ε valószínűséggel véletlen cselekvést választ egyenletes eloszlással, tehát ekkor 1/(*N*-1) valószínűséggel a nem a legjobbnak gondoltak közül

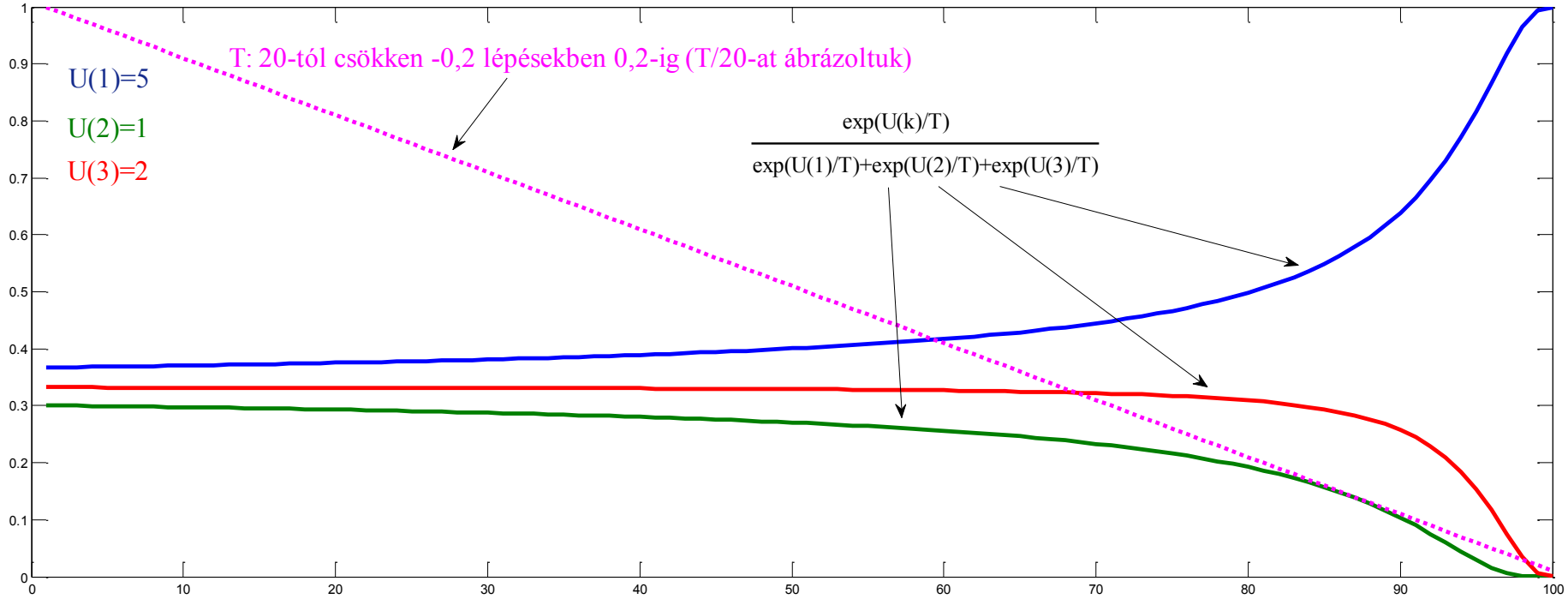
3. Felfedezési stratégia - <u>Boltzmann-felfedezési modell</u>

egy *a* cselekvés megválasztásának valószínűsége egy *s* állapotban:

$$P(a,s) = \frac{e^{hasznosság(a,s)/T}}{\sum_{a'} e^{hasznosság(a',s)/T}}$$

← Az összes, *s*-ben lehetséges *N* db *a* cselekvésre összegezve

a T „hőmérséklet” a két véglet között szabályoz.
Ha $T \rightarrow \infty$, akkor a választás tisztán (egyenletesen) véletlen,
ha $T \rightarrow 0$, akkor a választás mohó.



Boltzmann felfedezési modell, 3 lehetséges cselekvés, 3 követő állapotba visz. A követő állapotok hasznosságából becsülhető értékük U(a1,1)=5, U(a2,2)=1, U(a3,3)=2

$$P(a,s) = \frac{e^{hasznosság(a,s)/T}}{\Sigma_{a'} e^{hasznosság(a',s)/T}}$$

A cselekvésérték függvény tanulása

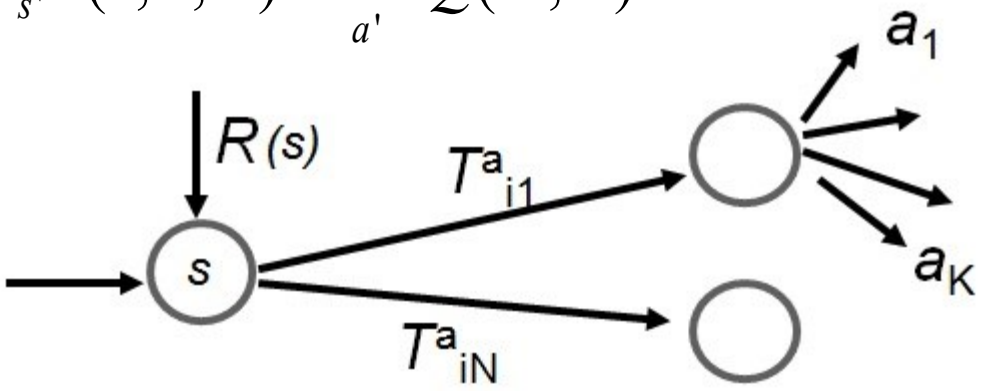
A cselekvés-érték függvény egy adott állapotban választott adott cselekvéshez egy várható hasznosságot rendel: **Q-érték**

$$U(s) = \max_a Q(a,s)$$

A Q-értékek fontossága:
a feltétel-cselekvés szabályokhoz hasonlóan lehetővé teszik a döntés tanulását, de **modell használata** nélkül, **közvetlenül** a jutalom visszacsatolásával.

Mint a hasznosság-értékeknél, itt is felírhatunk egy kényszeregyenletet, amely **egyensúlyi állapotban**, amikor a **Q-értékek korrektek**, fenn kell álljon:

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



A cselekvés-érték függvény tanulása

Ezt az egyenletet közvetlenül felhasználhatjuk egy olyan iterációs folyamat módosítási egyenleteként, amely egy adott modell esetén a pontos Q-értékek számítását végzi. De ez nem lenne modell-mentes.

Az **időbeli különbség** viszont nem igényli a modell ismeretét. Az **IK** módszer **Q-tanulásának** módosítási egyenlete:

$$Q(a,s) \leftarrow Q(a,s) + \alpha \left[R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s) \right]$$

SARSA Q-tanulás (State-Action-Reward-State-Action) megspóroljuk a max() műveletet, aktuális a'-t használjuk

$$Q(a,s) \leftarrow Q(a,s) + \alpha [R(s) + \gamma Q(a',s') - Q(a,s)]$$

(a' megválasztása pl. Boltzmann felfedezési modellből, felfedezési függvényből stb.)

Itt is használható ugyanaz a **felfedezési-függvény**

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
$\mathbf{N}_{sa}$, 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
$\mathbf{Q}[a,s] \leftarrow \mathbf{Q}[a,s] + \alpha(r + \gamma \max_{a'} \mathbf{Q}[a',s'] - \mathbf{Q}[a,s])$
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.)

A megerősítéses tanulás általánosító képessége

Eddig: ágens által tanult hasznosság: **táblázatos** (mátrix) formában reprezentált = **explicit reprezentáció**
Kis állapotterek: elfogadható, de a konvergencia idő és (ADP esetén) az iterációnkénti idő gyorsan nő a tér méretével. Ráadásul hogyan általánosít egy táblázat? A nem ismert sorokban milyen választ adjunk?

Gondosan vezérelt közelítő ADP: 10.000 állapot, vagy ennél is több.
De a való világhoz közelebb álló környezetek szóba se jöhetnek.

(Sakk – állapottere 10^{50}-10^{120} nagyságrendű. Az összes állapotot látni kell, hogy megtanuljunk játszani?!)

A megerősítéses tanulás általánosító képessége

Egyetlen lehetőség: a függvény **implicit reprezentációja**:
Pl. egy játékban a táblaállás tulajdonságainak valamilyen halmaza $f_1,\ldots f_n$.
Az állás becsült hasznosság-függvénye:

$$\hat{U}_\theta(s) = \theta_1 f_1(s) + \theta_2 f_2(s) + \ldots \theta_n f_n(s)$$

A hasznosság-függvény n értékkel jellemezhető, pl. 10^{120} érték helyett.
Egy átlagos sakk kiértékelő függvény kb. 10-20 súllyal épül fel,
tehát *hatalmas* tömörítést értünk el.

Az implicit reprezentáció által elért tömörítés teszi lehetővé, hogy a tanuló ágens általánosítani tudjon a már látott állapotokról az eddigiekben nem látottakra.

Az implicit reprezentáció legfontosabb aspektusa nem az, hogy kevesebb helyet foglal, hanem az, hogy lehetővé teszi a **bemeneti állapotok induktív általánosítását**. Az olyan módszerekről, amelyek ilyen reprezentációt tanulnak, azt mondjuk, hogy **bemeneti általánosítást** végeznek.

4 x 3 világ

Hasznosság implicit reprezentációja

$$\hat{U}_\theta(x,y) = \theta_0 + \theta_1 x + \theta_2 y$$

Hibafüggvény: a jósolt és a kísérleti ténylegesen tapasztalt hasznosságok különbségének négyzete

$$E_j(s) = \frac{1}{2}\left(\hat{U}_\theta(s) - u_j(s)\right)^2$$

Frissítés $$\theta_i \leftarrow \theta_i - \alpha \frac{\partial E_j(s)}{\partial \theta_i} = \theta_i + \alpha\left(u_j(s) - \hat{U}_\theta(s)\right)\frac{\partial \hat{U}_\theta(s)}{\partial \theta_i}$$

$$\theta_0 \leftarrow \theta_0 + \alpha\left(u_j(s) - \hat{U}_\theta(s)\right)$$

$$\theta_1 \leftarrow \theta_1 + \alpha\left(u_j(s) - \hat{U}_\theta(s)\right)x$$

$$\theta_2 \leftarrow \theta_2 + \alpha\left(u_j(s) - \hat{U}_\theta(s)\right)y$$

Gyakorlófeladat-Aktív Q-tanulás, SARSA, (IK módszer)

$a1$=FEL

0,8	0,8	0,8	+1
0,7	xxx	0,65	-1
0,68	0,6	0,5	-0,89

$a2$=LE

0,6	0,35	0,2	+1
0,3	xxx	0,10	-1
0,4	0,3	0,0	0,0

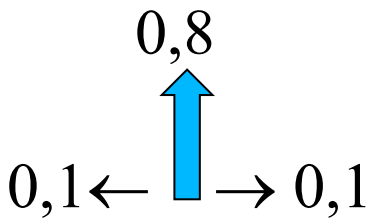
$a3$=JOBBRA

0,83	0,88	0,95	+1
0,71	xxx	-0,8	-1
0,2	0,1	0,1	-0,2

$a4$=BALRA

0,5	0,6	0,7	+1
0,5	xxx	0,4	-1
0,47	0,42	0,32	0,3

Megerősítés csak a végállapotokban!



Feladat:

1A. Az (1,1) – bal alsó sarok – állapotból indultunk, és a=FEL-t választottunk. A jelenlegi (a fenti táblázatokban megadott) $\hat{Q}(a,s)$ becsléseink alapján melyik cselekvést milyen valószínűséggel választjuk, ha az (1,2), illetve ha a (2,1) állapotba jutottunk, és

- mohó eljárással „biztosítjuk a felfedezést”
- hóbortos eljárással biztosítjuk a felfedezést
- ε-mohó eljárással biztosítjuk a felfedezést és ε=0,12
- Boltzmann lehűtéssel biztosítjuk a felfedezést, és a jelenlegi iterációnál T=20? Adja meg a valószínűségeket T=5, T=0,2 és T=0,01 esetre is!

(a következő véletlen számokat dobja a véletlenszám-generátorunk: 0,4231 0,7813)

1B. Tegyük fel, hogy az (1,2)-re léptünk. Hogyan alakul az IK-tanulás alapján a

$$\hat{Q}(FEL,(1,1)),\quad \hat{Q}(JOBBRA,(1,1)),\quad \hat{Q}(LE,(1,1)),\quad \hat{Q}(BALRA,(1,1))$$

becslésünk, ha a bátorsági (vagy tanulási) faktor) 0,1; a leszámítolási tényező 0,5 és az (1,2) állapotban a Boltzmann módszerrel választunk cselekvést, T=100?

Az alábbi véletlen számsorozatot használhatjuk a valószínűségi jellegű lépéseknél:

0,8147 0,9058 0,1270 0,9134 0,6324 0,0975 0,2785

2. Passzív megerősítéses tanulást végzünk időbeli különbség módszerrel. Az alábbi lépéssorozat mentén módosítunk, a tanulás bátorsági faktora 0,1; a leszámítolási tényező 0,8.

S1 ⇒ S9 ⇒ S12 ⇒ S11 ⇒ S8 ⇒ S17 ⇒ S18 ⇒ S14 ⇒ S11 ⇒ S5 ⇒ S7 ⇒ S10

Mi lesz a 11-es állapot hasznosságértékének új becslése a lépéssorozat után, ha a lépéssorozatot megelőzően a hasznosságbecslések a következők voltak:

U1 = 5	**U2 = 6**	**U3 = 7**	**U4 = 8**	**U5 = 7**	**U6 = 8**
U7 = 9	**U8 = 8**	**U9 = -6**	**U10 = +12**	**U11 = +10**	**U12 = -9**
U13 = -9	**U14 = -4**	**U15 = -9**	**U16 = -7**	**U17 = -7**	**U18 = -7**

Csak 2 állapotban van jutalom, az S10-es (vég)állapotban R10=+12, az S13-as állapotban R13=-13.

(Ne törődjön azzal, hogy előzetesen kialakulhatott-e ez a hasznosságbecslés!)

3. 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 T(s → 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 T(s → 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

Az állapotokhoz tartozó jutalmak $R(S_1) = -0,5$; $R(S_2) = +0,4$; $R(S_3)=+1,2$; $R(S_4)=+0,5$. Értékiterációs algoritmust alkalmazunk, az eddigi iterációk eredményeképp a *t*. lépésben a becsült hasznosságok $U_t(S_1) = 0$; $U_t(S_2) = 1$; $U_t(S_3)= R(S_3)= 1,2$; $U_t(S_4) = 1$.

Megegyeznek-e az iteráció során kapott értékek a valós hasznosságokkal? Előfordulhat-e nullánál nagyobb valószínűséggel, hogy a rendszer soha nem jut el a végállapotába (S_3-ba), ha mindig az a_2 cselekvést választjuk?