Alkalmazott mesterséges intelligencia (AMI)

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

9. ea. (2023 ősz)

Megerősítéses tanulás

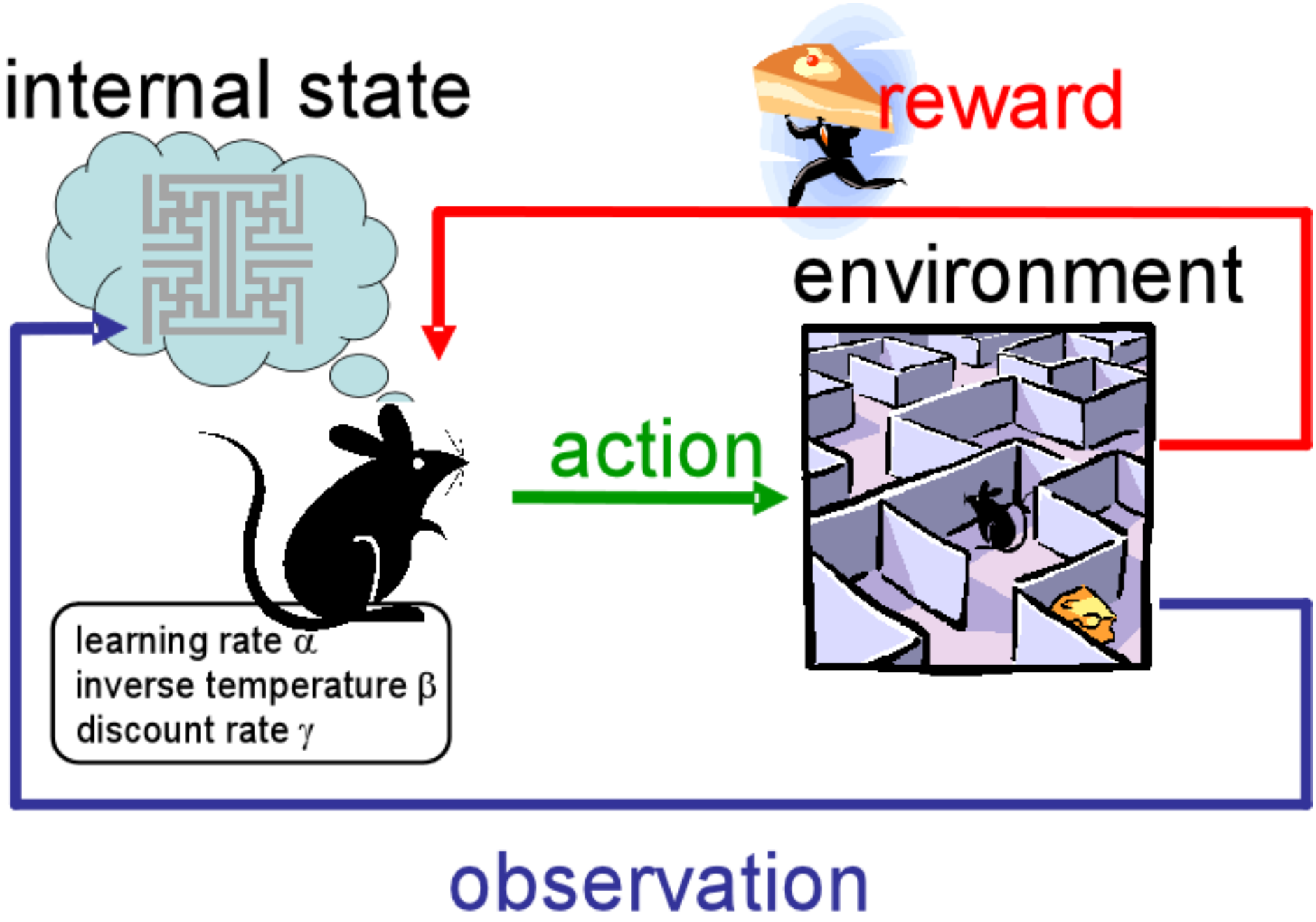
http://http://mialmanach.mit.bme.hu/aima/ch21 21. fejezet

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


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

Megerősítéses tanulás

Induktív tanulás: minden mintához megvan a jó válasz, egy „tanító”mindig megmondja, hogy jót vagy rosszat válaszoltunk, rendszerint azt is, hogy mennyire jó vagy rossz a válasz.

Megerősítéses tanulás: nincs tanító, csak időnként kapunk visszajelzést, hogy az eddigi lépéssorozat pillanatnyi eredménye jó vagy rossz-e.

Például a sakkjáték, tanító nélkül is van visszacsatolás a játék végén: **nyert** vagy **vesztett** = +/- **jutalom**, azaz pozitív vagy negatív **megerősítés**. (Pozitív: összességében jó, amit csináltunk, negatív: összességében rossz.)

Megerősítés: milyen gyakran? milyen erős? milyen értékű?

Különbség a szekvenciális döntéshez képest: **nem ismerjük a minket körülvevő világot (T(s,a,s’)**-t)! Fel kell fedeznünk, miközben igyekszünk **maximális összjutalmat gyűjteni**! **Egymásnak ellentmondó célok!**

Megerősítéses tanulás

A feladat:

A (ritka) jutalmakból megtanulni egy sikeres ágens-függvényt (optimális eljárásmód: melyik állapot hasznos, melyik állapotban melyik cselekvés hasznos).

Nehéz: az információhiány miatt az ágens sohasem tudja, hogy **mik a jó lépések**, sőt azt sem, hogy **melyik jutalom melyik cselekvés(ek)ből származik.**

Ágens tudása:

Induláskor: vagy **ismeri már** a környezetet és a cselekvéseinek hatását (ekkor szekvenciális döntési problémának neveztük), vagy pedig még ezt is **meg kell tanulnia**.

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

Ágens:

passzív tanuló: figyeli a világ alakulását és tanul, de *előre rögzített eljárásmód*, nem mérlegel, előre kódoltan cselekszik. A cél „csupán” a világ megismerése.

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

Ágens kialakítása: 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*) **hasznosságfüggvény** tanulása, ennek alapján a cselekvések eldöntése úgy, hogy az elérhető hasznosság várható értéke maximális legyen. (**Környezet modellje szükséges,** vagy adott, vagy ezt is tanulja az ágens**.** 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**/ágensmodell **nem szükséges**) – **Q tanulás (model-free)**

A cselekvésérték függvény tanulása

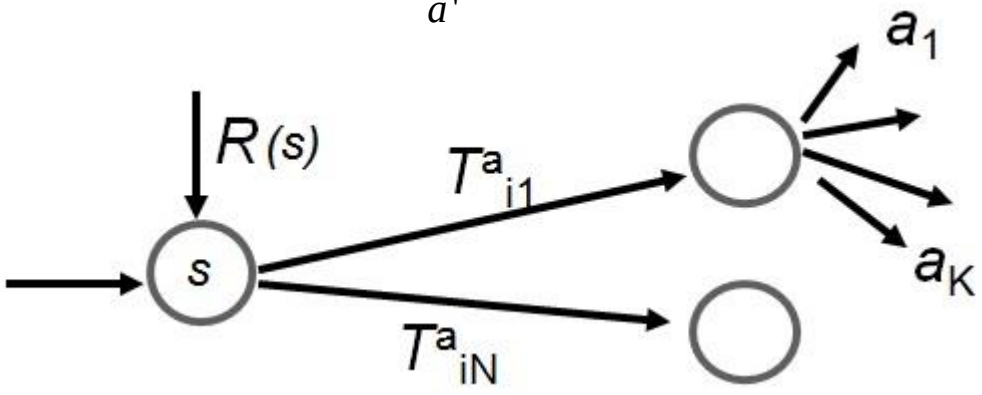
A cselekvés-érték függvény egy adott állapotban választott adott cselekvéshez egy várható hasznosságot rendel: **Q-érték**

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

A Q-értékek fontossága:
Látni fogjuk, hogy lehetővé teszik a döntést **modell használata** nélkül.
Közvetlenül a jutalom visszacsatolásával **tanulhatók**.

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

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

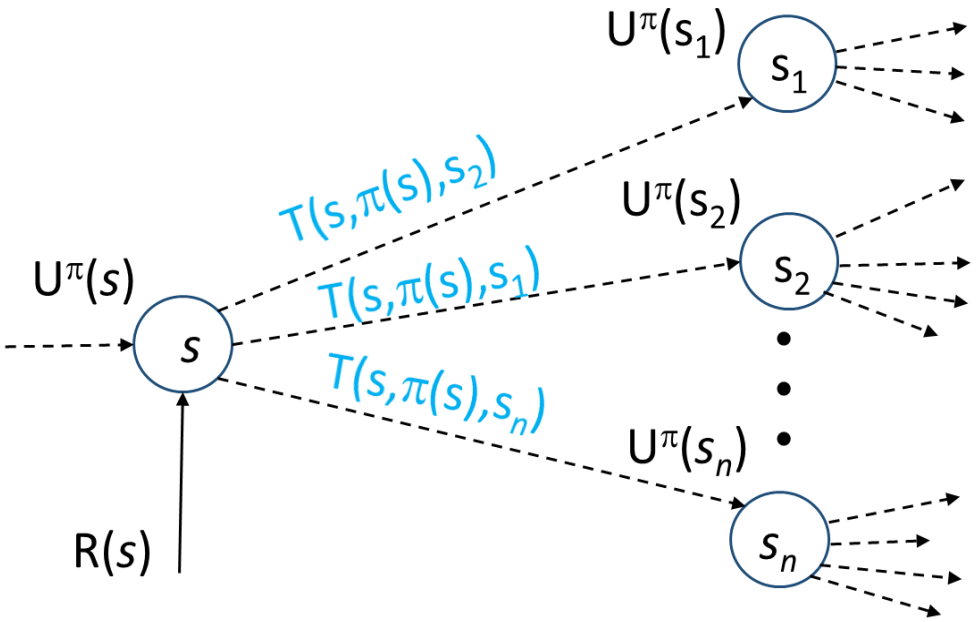


Fontos egyszerűsítés: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság **additív)**

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

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

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



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

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 környezet - T(s,a,s') – és 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 (ami talán jobban kifejezi, hogy mit tanul az ágens):

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

Passzív tanulás - Közvetlen hasznosságbecslés

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

Sok kísérletet végzünk, és feljegyezzük, hogy hányszor voltunk az s állapotban , és 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.

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

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

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 **<u>lineáris</u> egyenletrendszer megoldásával** kaphatók meg:

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

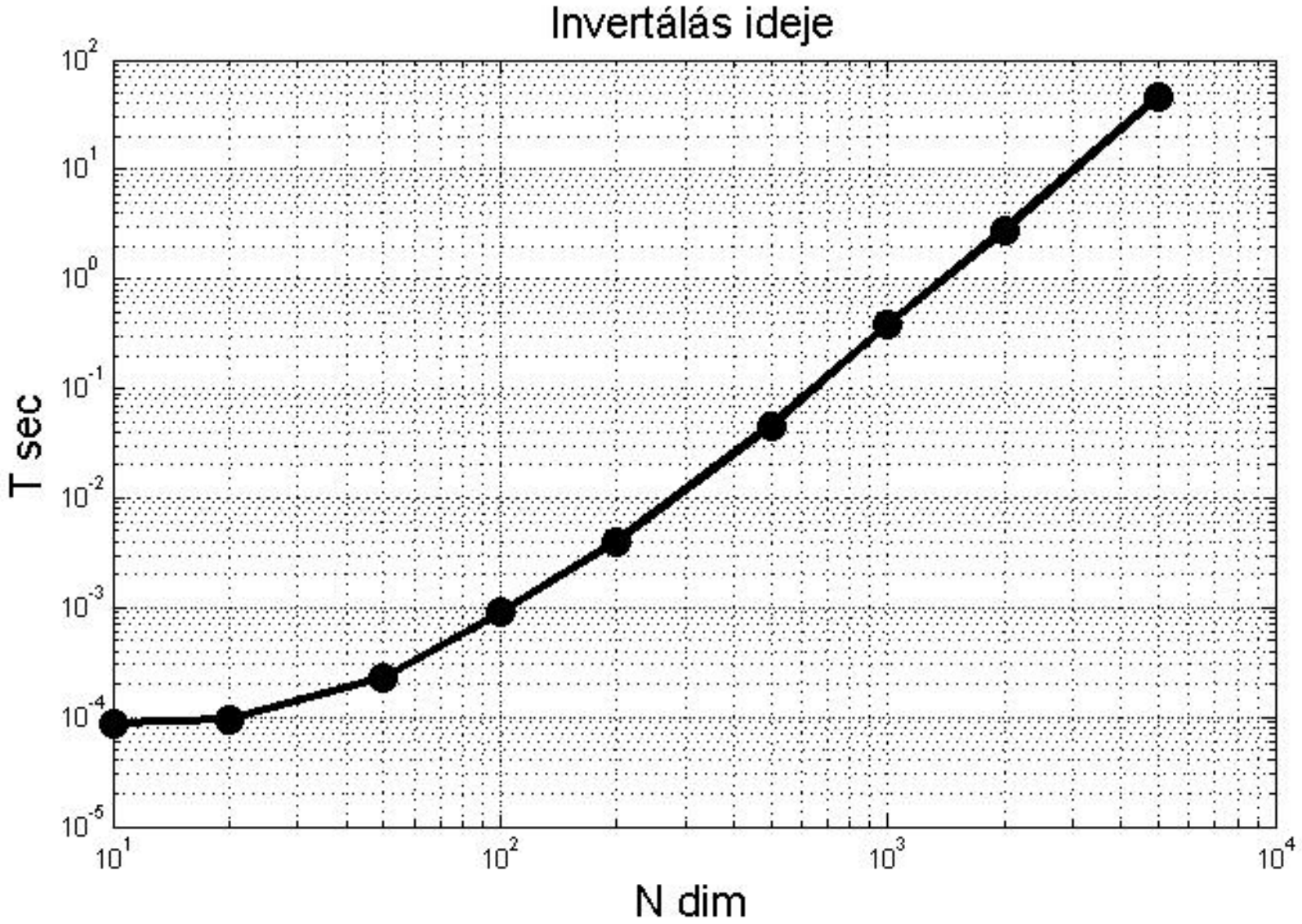
megtanult

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?



Nem a konkrét értékek fontosak (ezek változnak a hardver fejlődésével), hanem a **jelleg**! (*logaritmikus skála*!) Nagy állapotterekben nem megy. Sajnos 10^4 állapot még elég kicsinek számít!

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

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

Alapötlet: Minden egyes lépésben vegyük úgy, mintha ez a lépéssorozat mindig ugyanígy zajlana le! (Azt feltételezzük, hogy T(*s*,*s*')=1!) A hátralevő összjutalom helyett tehát használjuk fel a megfigyelt állapotátmeneteknél a hasznosság *egyetlen, az aktuális lépés* alapján becsült értékét:

Ha csak az aktuális utat néznénk (s$\rightarrow$s' átmenet), akkor a frissítés, amellyel egyensúlyba lehet hozni a 3 mennyiséget:

$$U(s) \leftarrow R(s) + \gamma \cdot U(s')$$

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

Ha csak az aktuális utat néznénk (s→s' átmenet): $U(s) \leftarrow R(s) + \gamma \cdot U(s')$

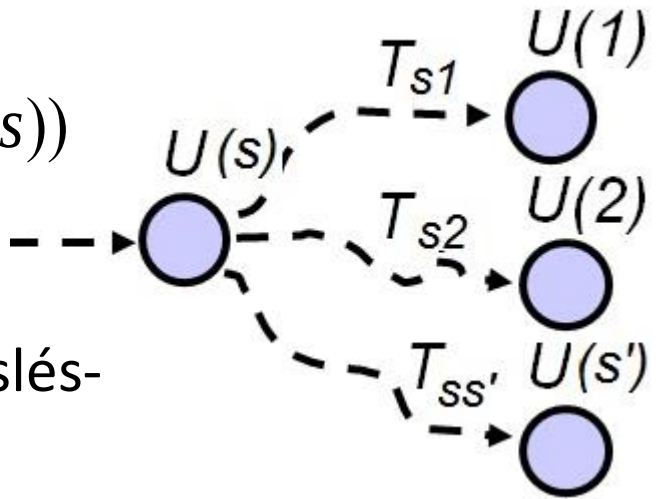
Alapötlet: a hátralevő összjutalom helyett használjuk fel a megfigyelt állapotátmeneteknél a hasznosság *egyetlen, az aktuális lépés* alapján becsült értékét:

A hasznosság új 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 hasznosságbecslés-különbsége **időbeli különbség (IK)** .

Az összes időbeli különbség eljárásmód alapötlete

1. korrekt hasznosság-értékek esetén **lokálisan** (az s állapotra) **fennálló feltételek rendszere**:

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

Ha egy adott s$\rightarrow$s' átmenetet tapasztalunk a jelenlegi sorozatban, akkor úgy vesszük, mintha T(*s*,*s*')=1 lenne (tehát a többi 0), így fenn kéne álljon az $U^{\pi}(s) = R(s) + \gamma U^{\pi}(s')$

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

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

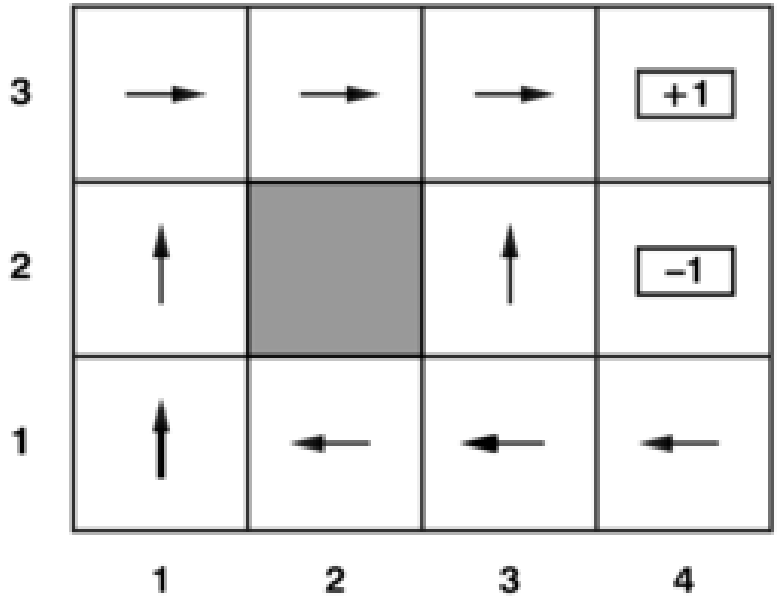
Megjegyzés: Az **ADP** módszer felhasználta a **modell teljes** ismeretét. Az **IK** a szomszédos (egy lépésben elérhető) állapotok közt fennálló kapcsolatokra vonatkozó információt használja, de csak azt, amely az **aktuális tanulási sorozatból származik**.

Az **IK** változatlanul **működik előzetesen ismeretlen környezet esetén is**. (*Nem használtuk ki a T(s,s') ismeretét*!)

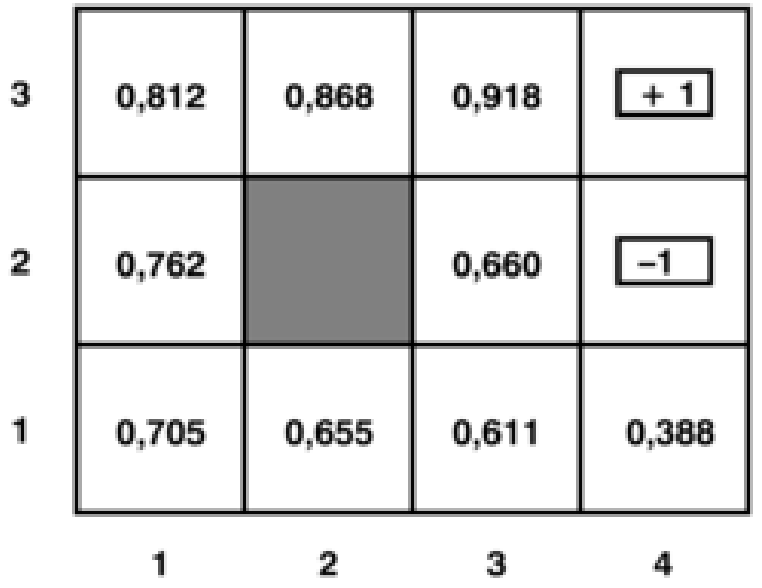
Az egyszerű demóproblémánk (Russell-Norvig)

- végállapotok: (4,2) és (4,3)
- $\gamma=1$
- minden állapotban, amelyik nem végállapot $R(s)= -0,04$
  (pl. így lehet modellezni a lépésköltséget)

0,8 valószínűséggel a szándékolt irányba lép, 0,1-0,1 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

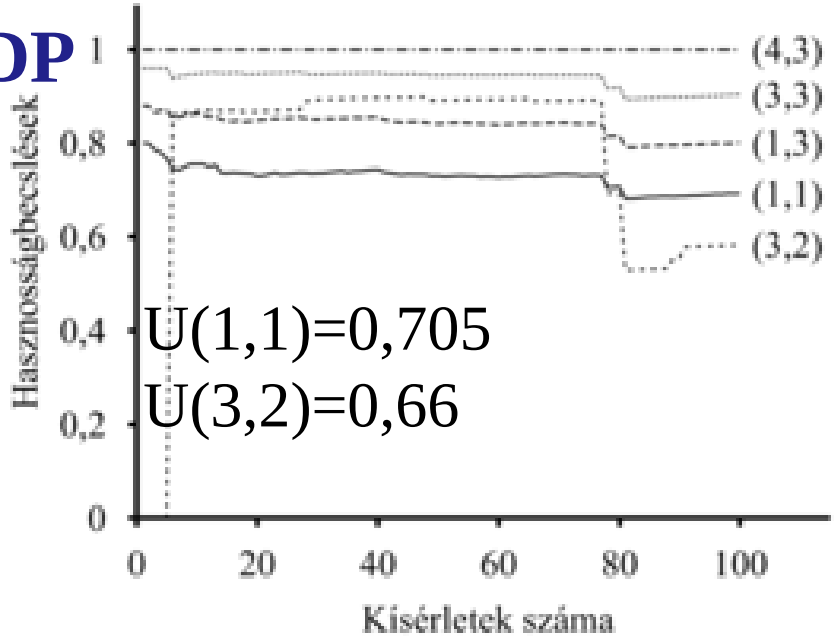


Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok

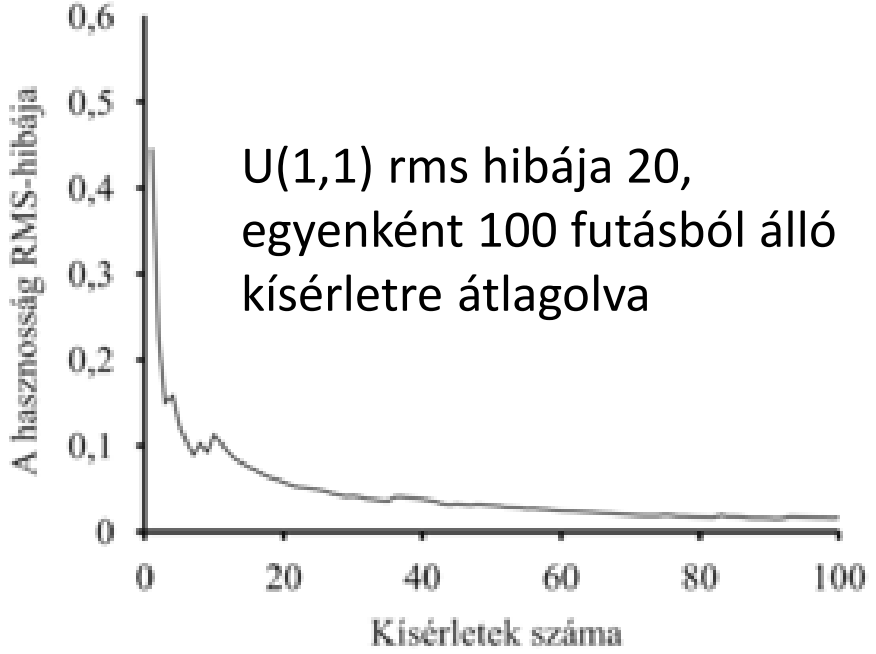
ADP



U(1,1)=0,705

U(3,2)=0,66

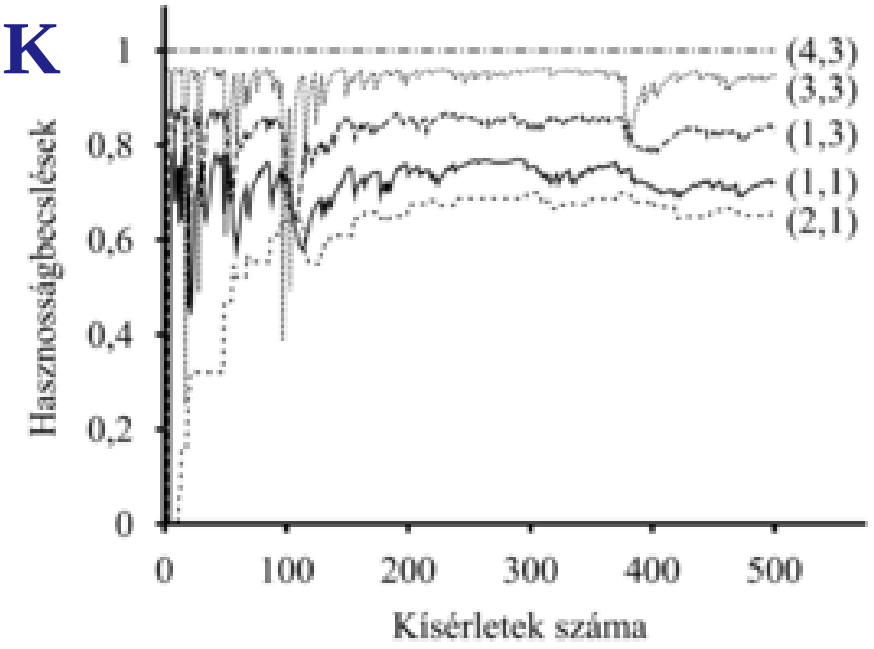
(a)

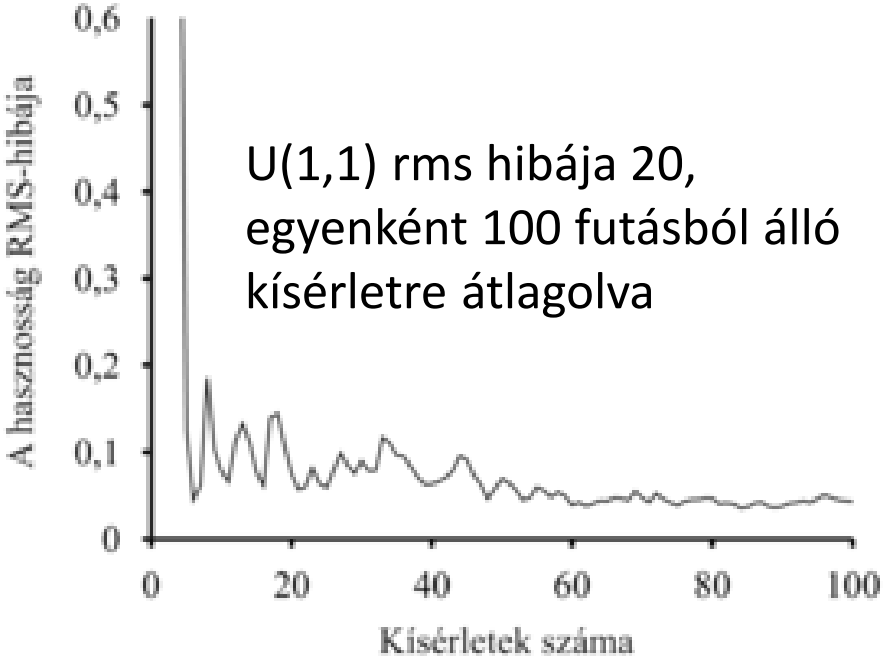


U(1,1) rms hibája 20, egyenként 100 futásból álló kísérletre átlagolva

(b)

IK





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

Ismeretlen környezetben végzett **Aktív megerősítéses tanulás**

Döntés: Melyik cselekvést? 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

Helyes-e (mohón) 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

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.**

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

Két szélsőséges megközelítés a cselekvés kiválasztásában:
„**Hóbortos, Felfedező**”: véletlen módon cselekszik, annak reményében, hogy végül is felfedezi az egész környezetet. A tudását nem használja arra, hogy jutalmakat gyűjtsön, ő *csak felfedezni szeret, a jutalom nem érdekli*.
„**Mohó**”: *a jelenlegi becslésre alapozva maximalizálja a hasznot*. Amikor eljut egy olyan ismeretszintre, hogy már tud valamennyi jutalmat gyűjteni, *elveszti érdeklődését a felfedezés iránt*.

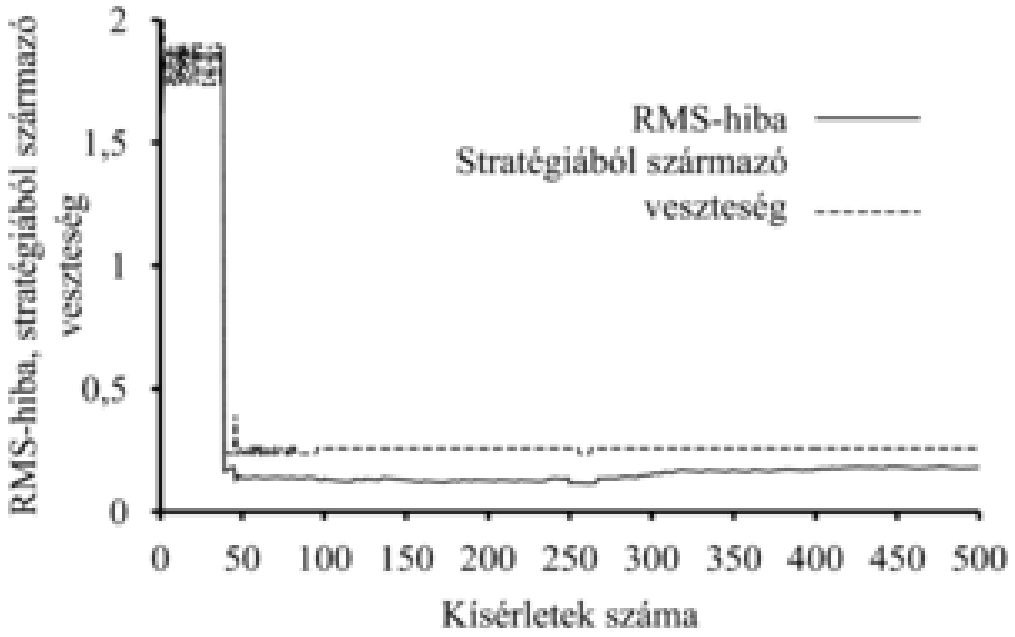
Hóbortos: képes jó hasznosság-becsléseket megtanulni az összes állapotra. Sohasem sikerül fejlődnie az optimális jutalom elérésében.

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

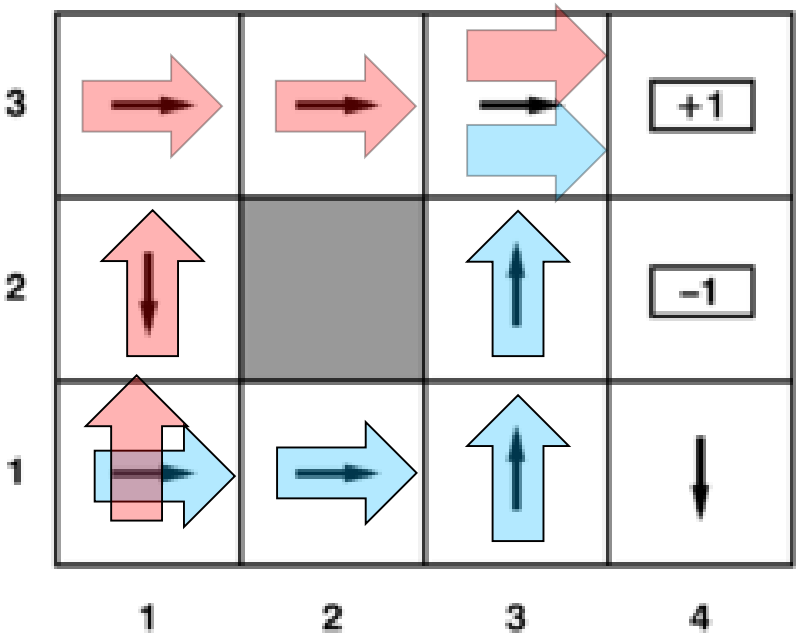
A **mohó ágens** által talált stratégia ⟶

A mohó ágens által talált optimális út

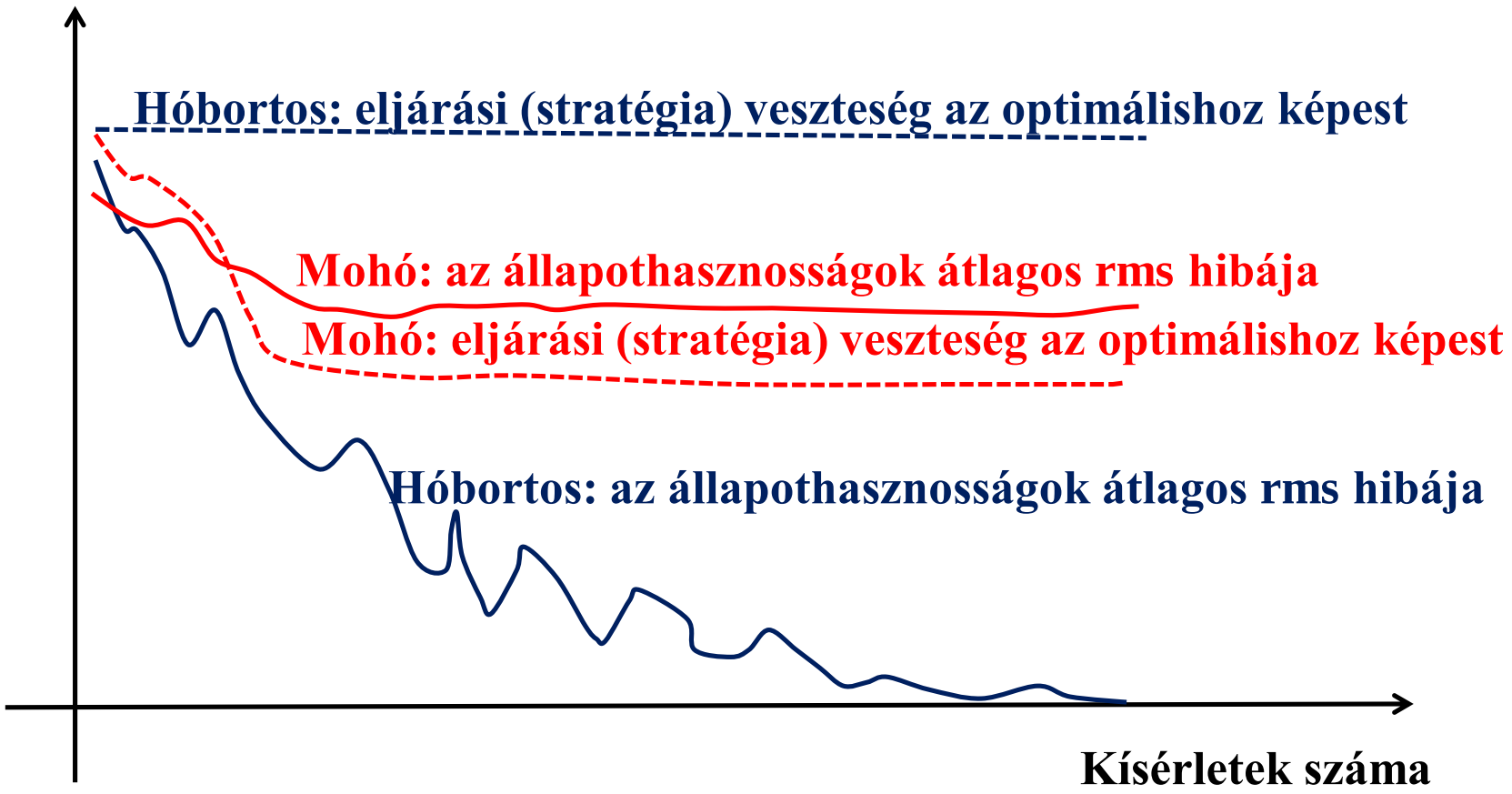
A valódi optimális út



(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 már valósághoz közeli modellel rendelkezik.

Létezik-e optimális felfedezési stratégia?

Az ágens

- előnyben 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 → felfedezési függvény

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

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

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

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

$U^+(i)$: az i állapothoz rendelt hasznosság ***optimista*** becslése
$N(a,i)$: az i állapotban hányszor próbálkoztunk az a cselekvéssel
Az a cselekvés, amely felderítetlen területek *felé* vezet, nagyobb súlyt kap.

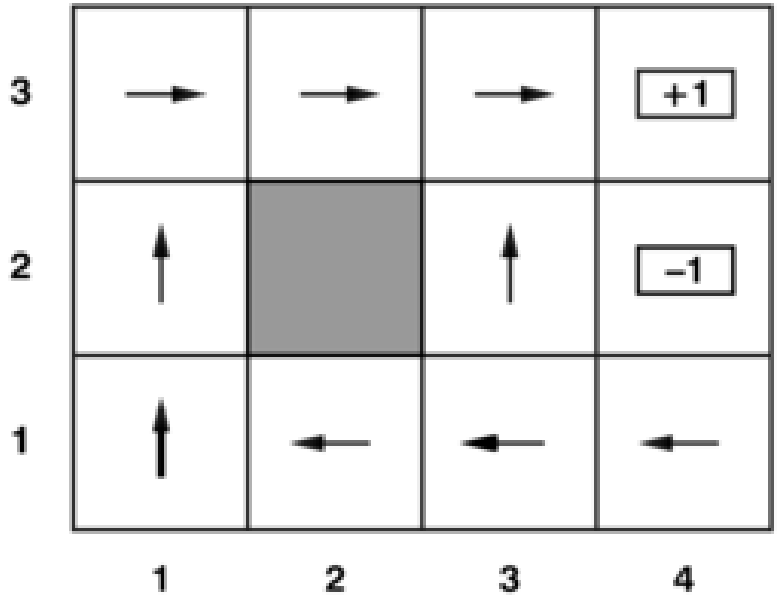
mohóság ↔ kíváncsiság
$f(u,n)$: u-ban monoton növekvő (mohóság),
n-ben monoton csökkenő (a próbálkozások számával csökkenő kíváncsiság)

Ismételjük meg a régi fóliát

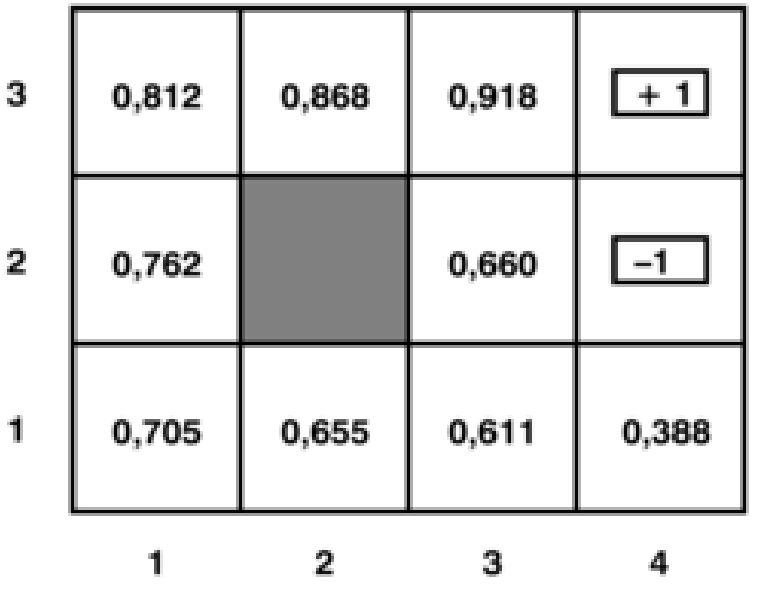
Egy egyszerű demóprobléma:

- végállapotok: (4,2) és (4,3)
- γ=1
- minden állapotban, amelyik nem végállapot R(*s*)= - 0,04
  (pl. így lehet modellezni a lépésköltséget)

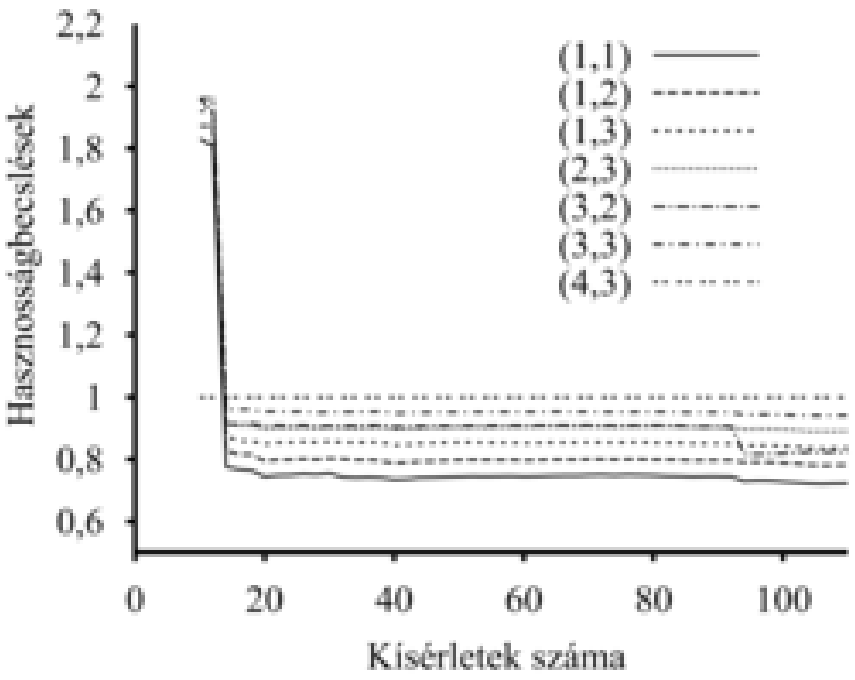
0,8 valószínűséggel a szándékolt irányba lép; 0,1-0,1 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

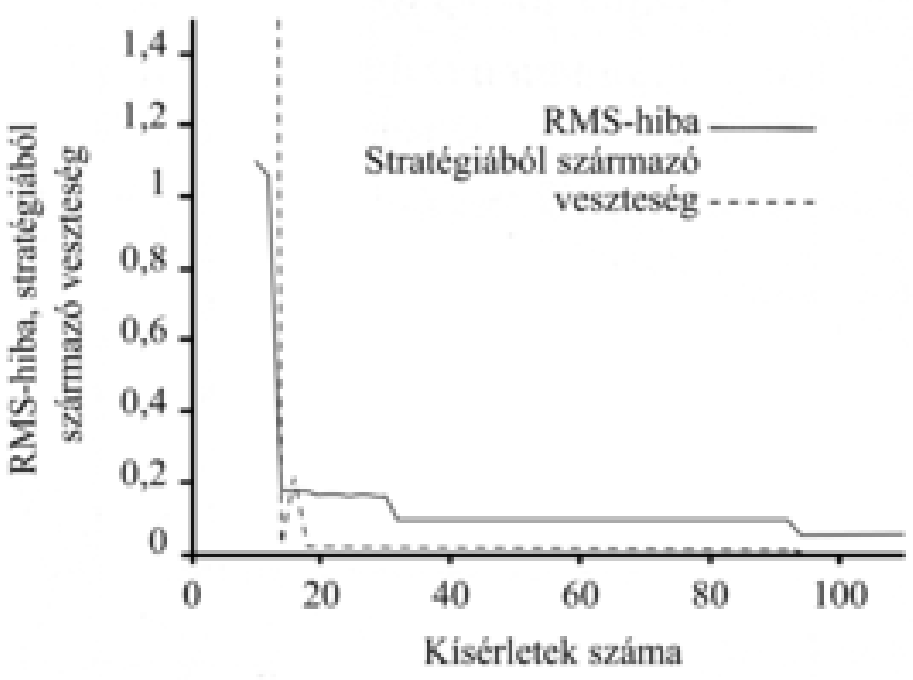


Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok





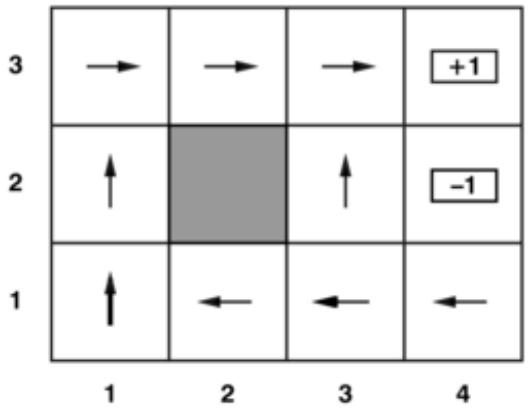
(b)

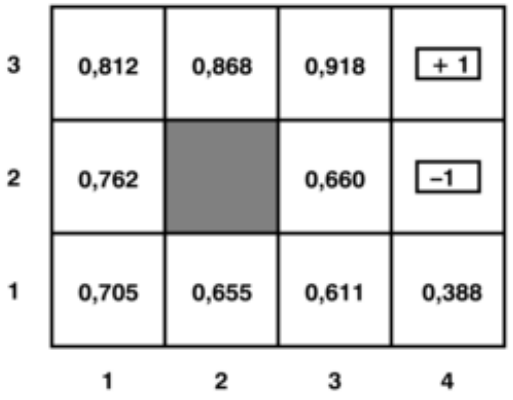
Felfedezési függvény: $R^+=2$, $N_e=5$

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

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

Valós hasznosságok, opt. stratégia





1. **Felfedezési stratégia** - **felfedezési függvény** (eddig erről volt szó)

2. Felfedezési stratégia - **ε-mohóság** (*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

A cselekvésérték függvény tanulása

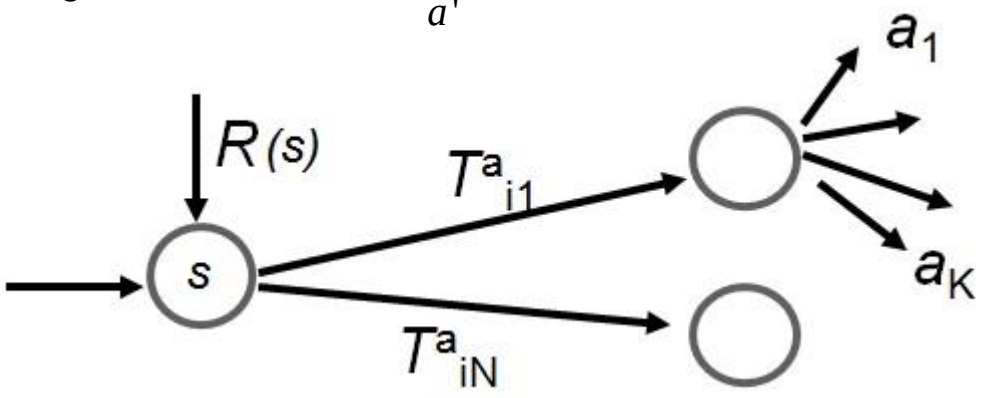
A cselekvés-érték függvény egy adott állapotban választott adott cselekvéshez egy várható hasznosságot rendel: **Q-érték**

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

A Q-értékek fontossága:
Lehetővé teszik a döntést **modell használata** nélkül. **Közvetlenül** a jutalom visszacsatolásával **tanulhatók**.

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

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



Régebbi vizsgapélda: Aktív megerősítéses tanulásnál az egyes állapotokban cselekvést kell választanunk, ezt bizonyos esetekben véletlenszerűen célszerű megtenni. Egy adott *s* állapotban 4 cselekvés közül kell választanunk: *A*1, *A*2, *A*3 és *A*4. Az egyes cselekvések becsült hasznossága ebben az állapotban: Q(*A*1,*s*)=+14,0 ; Q(*A*2,*s*)=+4,4 ; Q(*A*3,*s*)=+5,6 és Q(*A*4,*s*)=+1,2 ; eddig mindegyik cselekvést ötször-ötször választottuk az *s* állapotban. A cselekvésválasztást úgy végezzük el, hogy kiszámítjuk a négy cselekvés valószínűségét valamilyen eljárással, majd véletlenszám-generátorunktól lekérünk egy – [0,1] tartományba eső – értéket, a konkrét esetben ez *R*=0,9371-re adódott. Ez az érték választja ki számunkra a cselekvést, az alábbiak szerint:

ha $0 \le R < P(A1)$ akkor $A1$-et választjuk

ha $P(A1) \le R < P(A1) + P(A2)$ akkor $A2$-t választjuk

ha $P(A1) + P(A2) \le R < P(A1) + P(A2) + P(A3)$ akkor $A3$-at választjuk

ha $P(A1) + P(A2) + P(A3) \le R$ akkor $A4$-et választjuk

Mi lesz a választott cselekvés, ha **ε-mohó** eljárással választjuk, és ε=0,3? (*Indoklás szükséges*!)

A. $A1$ **B.** $A2$ **C.** $A3$ **D.** $A4$

$$P(A2) = P(A3) = P(A4) = \frac{\varepsilon}{3} = 0{,}1$$

$P(A1) = 1 - \varepsilon = 0{,}7$ mivel Q(A1,s) a legnagyobb,

tehát

$$P(A1) + P(A2) + P(A3) =$$

$$= 0{,}9 \le R = 0{,}9371 < P(A1) + P(A2) + P(A3) + P(A4) = 1$$

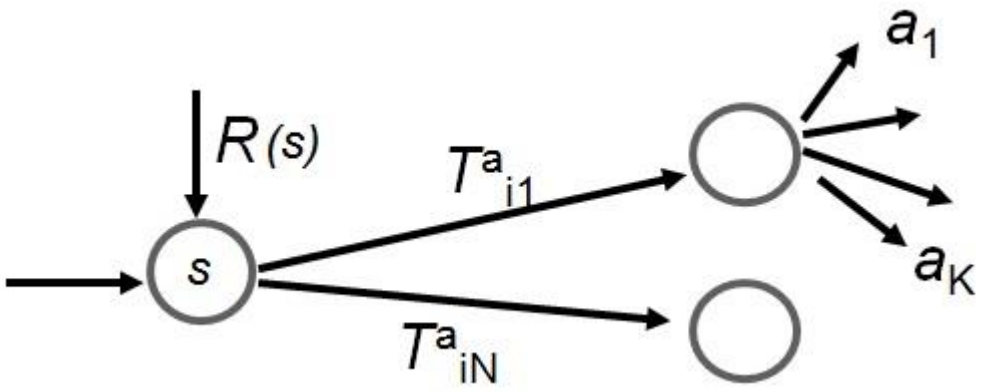
azaz $A4$-et választjuk.

A cselekvésérték függvény tanulása

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

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

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

Változatlan a passzív tanuláshoz képest!

Amennyiben el akarjuk kerülni a mohóság során szükséges max( ) kiértékelést, véletlenszerű – esetleg rossz - lépésekkel operálhatunk.
SARSA Q-tanulás (State-Action-Reward-State-Action)

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

ahol *a*' megválasztása pl. felfedezési függvénnyel, ε-mohó eljárással stb.

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

Pszeudókód:

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 (*s*) és cselekvéssel (*a*) indexelünk

$\mathbf{N}_{sa}$, az állapot-cselekvés párok gyakorisági táblája

s, *a*, *r* az előző állapot, előző cselekvés, jutalom, a és r kezdeti értéke nulla, s: valamilyen állapot

if *nem Végállapot* [s] **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 $s,a,r \leftarrow s'$, $\text{argmax}_{a'} f(\text{Q}[a',s'], N_{sa}[a',s']), r'$

else $a \leftarrow$ nulla

return *a*

(aztán a meghívó függvény végrehajtja *s'* állapotban a visszakapott *a* cselekvést, s"-be jut, újra meghívja a függvényt és megkapja az s"-ben végrehajtandó cselekvést stb.)

A megerősítéses tanulás általánosító képessége

Az **általánosító képesség** nem a megerősítéses tanulás jellemzője – minden tanulásnál fontos.

Eddig: ágens által tanult hasznosság: **táblázatos** (mátrix) formában reprezentált = **explicit reprezentáció** (T(*s*,*a*,*s*') vagy Q(*a*,*s*))

Kis állapottereknél elfogadható, de a konvergencia idő és (ADP esetén) az iterációnkénti idő gyorsan nő a tér méretével.

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?!)
Sajnos a táblázatnak nincs általánosító képessége, nem tudunk mondjuk 10.000 állapotból következtetni a teljes térre, sőt a 10.001-dikre sem…

A megerősítéses tanulás általánosító képessége

Egyetlen lehetőség: a függvény (hasznosságfüggvény, Q-függvény) **implicit reprezentációja**:

Pl. egy játékban a táblaállás *tulajdonság*ainak 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 kiértékelt állapotok hasznosságbecslése (vagy a Q[a,s] értékek) segítségével tanuljuk a Θ paramétereket – az ismeretlen állapotokra is ad (jó vagy rossz) választ $\Rightarrow$ általánosít!

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 (elég szerencsétlen, mert nemlineáris felületen helyezkednek el a hasznosságok, de példának jó):

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

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

Például (3,1)-ből indulunk, és az aktuális sorozatban $u_j(s)$=SumSor(3,1)=0,8 összjutalmat gyűjtünk.

A sorozat előtt a becslésünk:

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

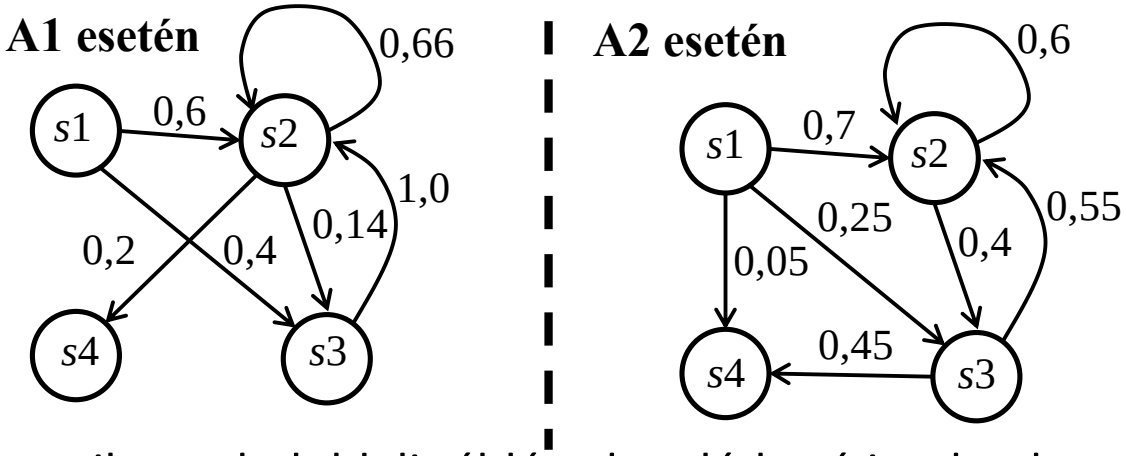
$$=0,5+0,2\cdot 3-0,2\cdot 1=0,9$$

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

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

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

Régebbi vizsgapélda: Egy probléma mindegyik állapotában két cselekvést választhatunk: A1-et vagy A2-őt. A rendszer végállapota s4, a leszámítolási tényező 0,2. A választott cselekvéstől függően az alábbi állapotátmeneti valószínűségek jellemzik a rendszert:



A rendszer egyes állapotaiban a baloldali táblázatban látható jutalmakat kapjuk. A jobboldali táblázat mutatja az optimális eljárásmódot és az egyes állapotok hasznosságát

s	s1	s2	s3	s4
R(s)	+1	-5	+2	+10

s	s1	s2	s3	s4
π*(s)	A1	A1	A2	-
U (s)	+0,5591	-5,2245	+2,3253	+10

Írja fel a Bellman egyenletet a konkrét értékekkel *optimális stratégia alkalmazása esetére* az s1 állapotra!

1. gyakorlófeladat Passzív megerősítéses tanulást végzünk *időbeli különbség* módszerrel. Az alábbi lépés-sorozat 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!)

2. Gyakorlófeladat 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

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 $U_t(S_1) = 0$; $U_t(S_2) = 1$; $U_t(S_3)= R(S_3)= 1,2$; $U_t(S_4) = 1$.

2A. Megegyeznek-e az iteráció során kapott értékek a valós hasznosságokkal?

2B. Előfordulhat-e nullánál nagyobb valószínűséggel, hogy a rendszer soha nem jut el a végállapotába (S_3-ba), ha mindig az a_2 cselekvést választjuk?

3. gyakorlófeladat - Aktív Q-tanulás, 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

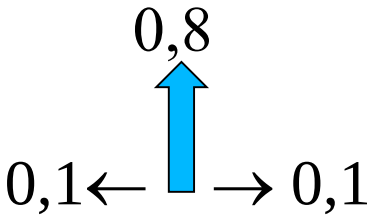
a3=JOBBRA

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

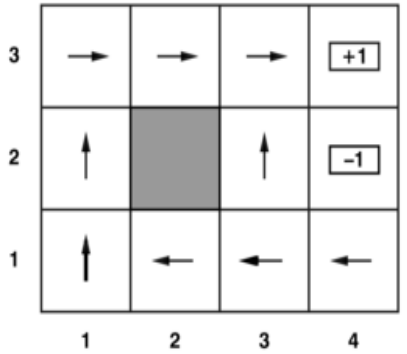
a4=BALRA

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!



emlékeztető:



Feladat:

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

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