Alkalmazott mesterséges intelligencia (AMI)

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

12. ea. (2019 ősz)

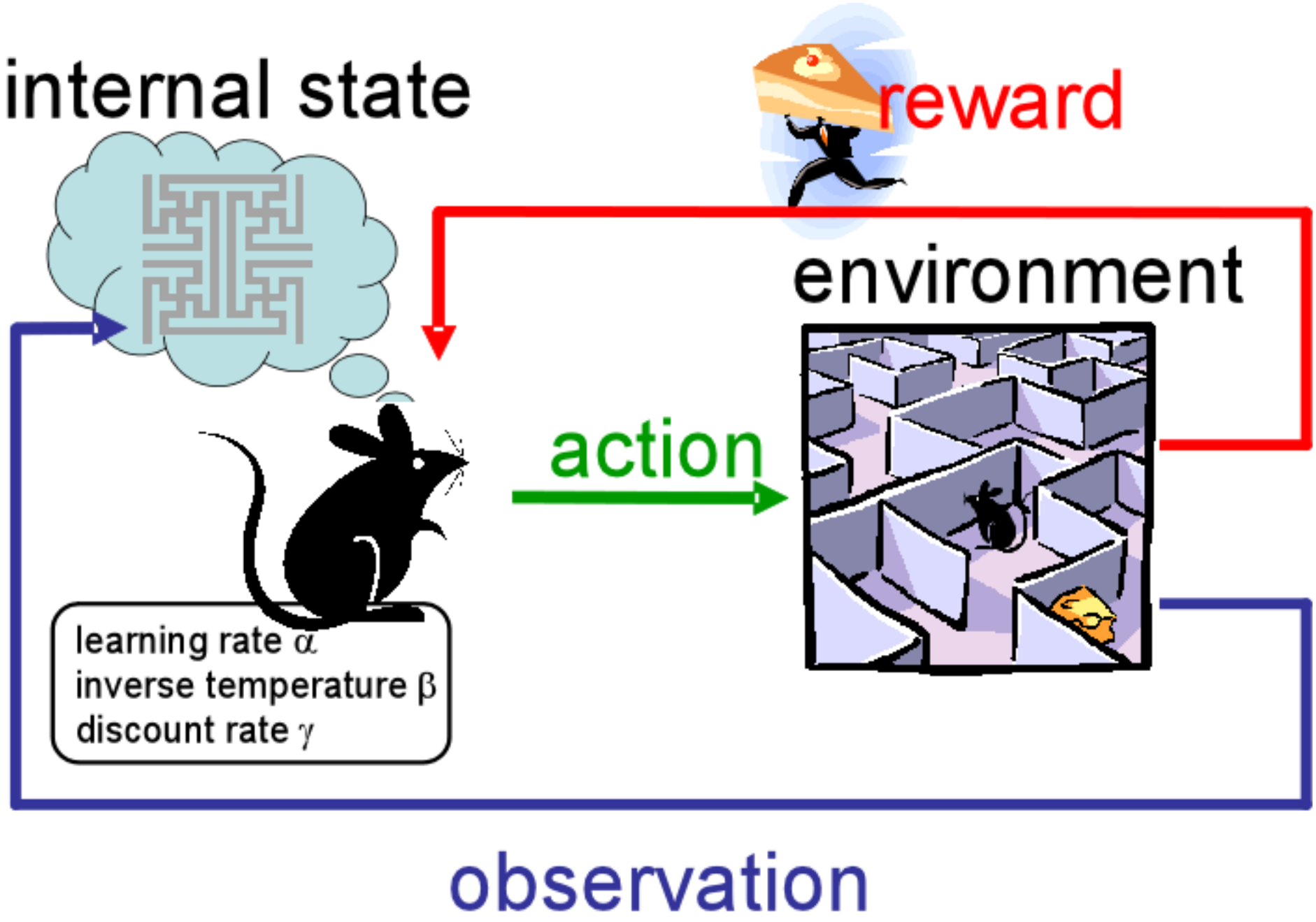
Megerősítéses tanulás

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

Előadó: Pataki Béla

a fóliák
Dobrowiecki Tadeusz és
Hullám Gábor anyagainak
felhasználásával készültek

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?

A feladat:

(ritka) jutalmakból megtanulni egy sikeres ágens-függvényt (optimális eljárásmód: melyik állapot, 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, 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, *előre rögzített eljárásmód*, nem mérlegel, előre kódoltan cselekszik

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

Ágens kialakítása: megerősítés → hasznosság leképezés

$R(s) \rightarrow U(s)$

Múlt óra – szekvenciális döntések

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

- valamilyen hasznosságot tulajdonítunk *az állapotnak*
  (az ebből az állapotból kiindulva összegyűjthető összjutalom, ha optimális cselekvéseket választunk)
- a cselekvéseink arra irányulnak, hogy minél nagyobb valószínűséggel egy hasznos állapotba jussunk

(Ehhez szükséges a **környezet modellje,** környezet: **T(*s*,*a*,*s*')**)

Q(a, s) **cselekvésérték függvény** (állapot-cselekvés párok) tanulása,

- valamilyen várható hasznot tulajdonítva *egy adott helyzetben egy adott cselekvés*nek
- ha egy állapotban vagyunk, akkor az adott s állapotban a legnagyobb Q(a,s) értékkel rendelkező cselekvést választjuk

(Ehhez **nem szükséges** a **környezet** modell **(model-free) – Q tanulás)**

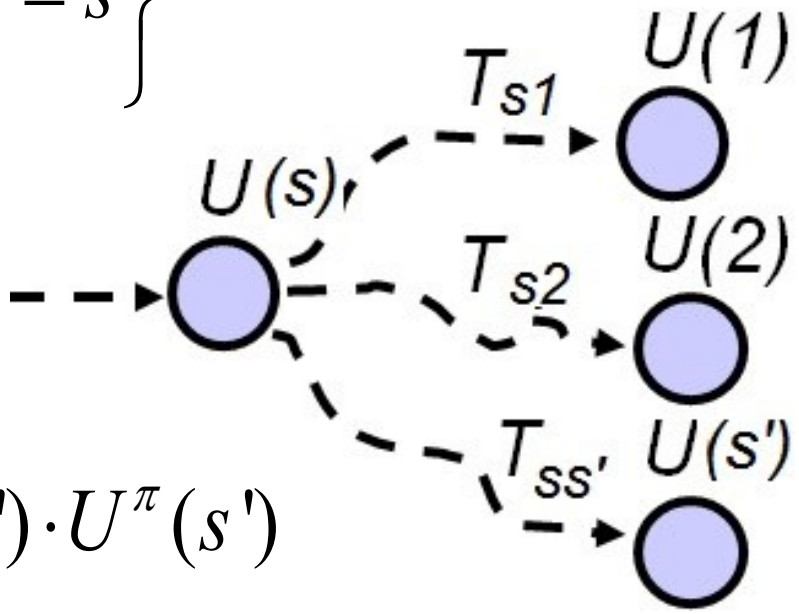
Fontos egyszerűsítés: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság **additív)**

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

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

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



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

Passzív megerősítéses tanulás

Markov döntési folyamat (MDF) szekvenciális döntés rögzített stratégia esetén – nincs max! (múlt óra alapján)

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

Az egyes állapotokban a cselekvéseket az ágens $a = \pi(s)$ stratégiája határozza meg. A hasznosságot az optimális stratégiával értelmezzük:

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

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.

Passzív megerősítéses tanulás

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 **jutalomfüggvényt** (**reward function**), $R(s)$-et, amely minden állapothoz megadja az ott elnyerhető jutalmat.

Egyszerűbb jelölések (amik itt talán jobban kifejezik, hogy mit és hogyan tanul az ágens):

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

$T(s, \pi(s), s')$ helyett $T(s, s')$ hiszen $\pi(s)$ rögzített

Passzív tanulás - Közvetlen hasznosságbecslés

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 az adott lépéssorozatban 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 következik!

Kvíz 12.1. Passzív megerősítéses tanulást végzünk **közvetlen hasznosságbecslés** módszerrel. Az alábbi legújabb lépéssorozat után módosítunk: **S1 ⇒ S13 ⇒ S8 ⇒ S11 ⇒ S18 ⇒ S14 ⇒ S8 ⇒ S5 ⇒ S7 ⇒ S10**

A lépéssorozatot megelőzően a hasznosságbecslések a következők voltak, és ez előtt 10-szer jártunk már az s11 állapotban?

Us1 = 5	**Us2 = 6**	**Us3 = 7**	**Us4 = 8**	**Us5 = 7**	**Us6 = 8**
Us7 = 9	**Us8 = 8**	**Us9 = -6**	**Us10 = +21**	**Us11 = +10**	**Us12 = -7**
Us13 = -9	**Us14 = -4**	**Us15 = -6**	**Us16 = -7**	**Us17 = -7**	**Us18 = -5**

Csak két állapotban van jutalom, az S10-es (vég)állapotban R10=+21, az S13-as állapotban R13=-15.

Mi lesz a 11-es állapot hasznosságértékének új becslése a módosítás után, ha a leszámítolási tényező* 1, *illetve, ha* 0,1*?

A. **γ=1 esetén Us11=10,01** **γ=0,1 esetén Us11=10**

B. **γ=1 esetén Us11=31** **γ=0,1 esetén Us11=10,2**

C. **γ=1 esetén Us11=11** **γ=0,1 esetén Us11=10**

D. **γ=1 esetén Us11=10,2** **γ=0,1 esetén Us11=10,02**

Passzív tanulás - Adaptív Dinamikus Programozás (ADP)

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:

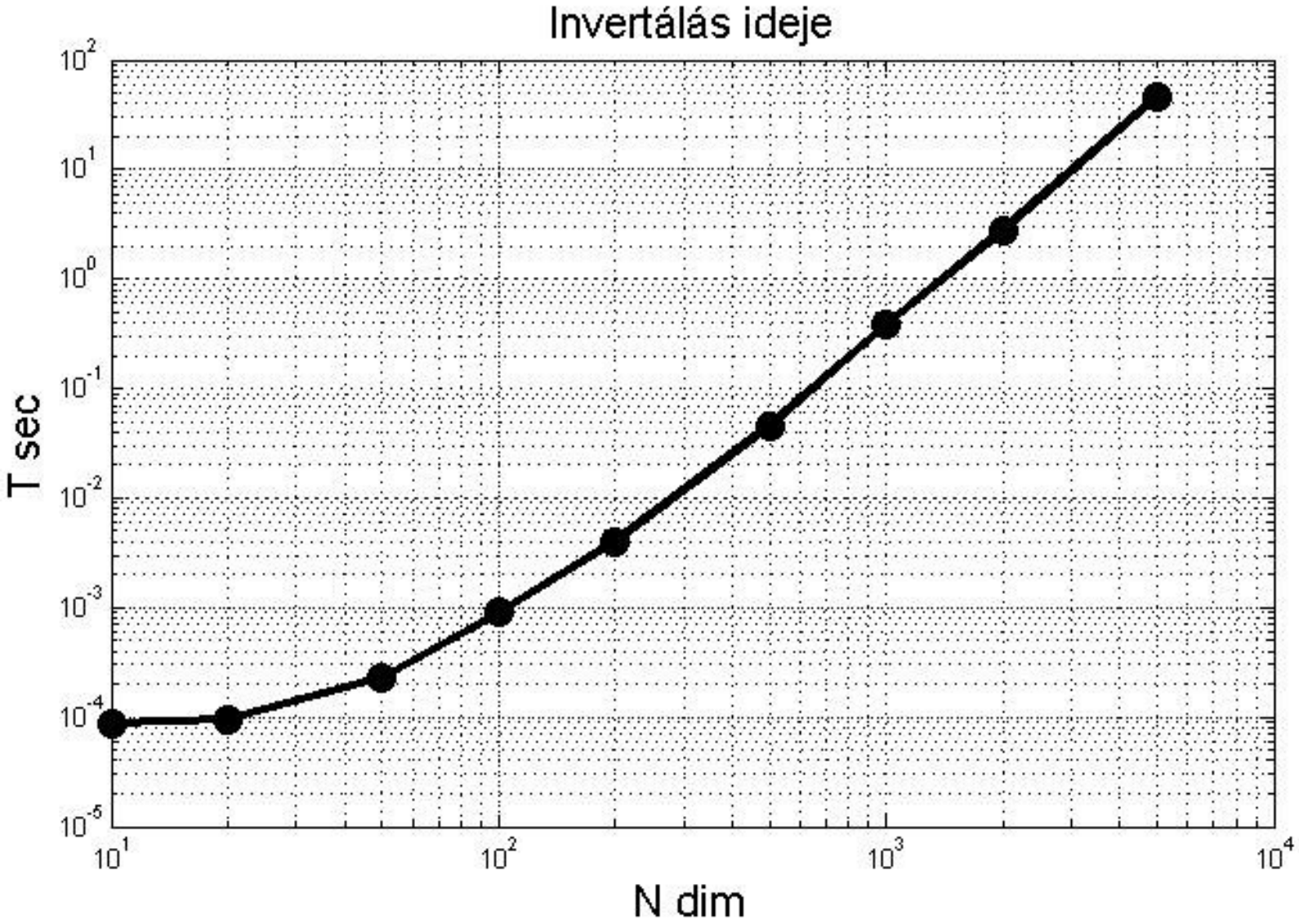
$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s,s') \cdot U^{\pi}(s')$$

megtanult — megfigyelt

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

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

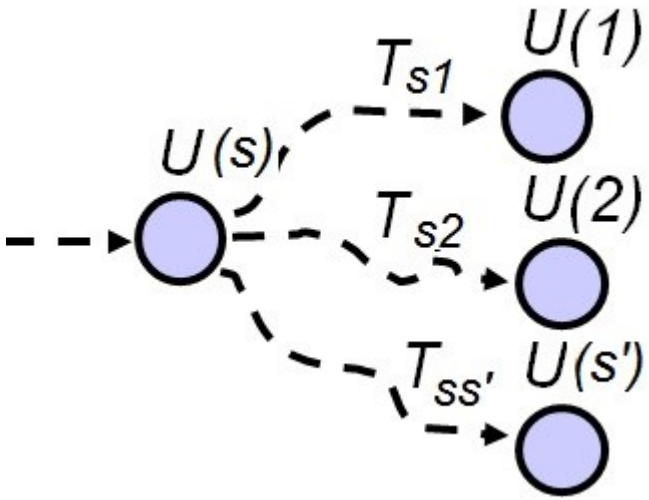
….és ha nem megy?



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



Ha csak az aktuális utat néznénk ($s \rightarrow sk$ átmenet):

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

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

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

kvíz következik!

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

A jutalom becslése az éppen elvégzett lépés alapján :

$$\hat{U}(s) = R(s) + \gamma U(sk)$$

Az előző becsléstől való eltérés:

$$\delta(s) = \hat{U}(s) - U(s) = R(s) + \gamma U(sk) - U(s)$$

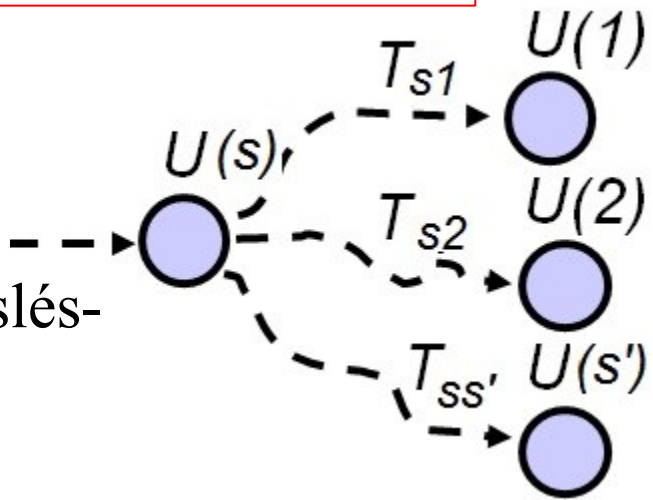
Frissítés:

$$U(s) \leftarrow U(s) + \alpha\,\delta(s) = U(s) + \alpha\left[R(s) + \gamma U(sk) - U(s)\right]$$

α - **bátorsági faktor,** tanulási tényező

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

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



Kvíz 12.2. Passzív megerősítéses tanulást végzünk **időbeli különbség** módszerrel. Az alábbi legújabb lépéssorozat után mi lesz a becslésünk?

S1 ⇒ S13 ⇒ S8 ⇒ S11 ⇒ S18 ⇒ S14 ⇒ S8 ⇒ S5 ⇒ S11 ⇒ S10

A lépéssorozatot megelőzően a hasznosságbecslések a következők voltak, ez előtt 10-szer jártunk már az s11 állapotban, és a tanulás bátorsági tényezője α=0,1.

Us1 = 5	**Us2 = 6**	**Us3 = 7**	**Us4 = 8**	**Us5 = 7**	**Us6 = 8**
Us7 = 9	**Us8 = 8**	**Us9 = -6**	**Us10 = +21**	**Us11 = +10**	**Us12 = -7**
Us13 = -7	**Us14 = -4**	**Us15 = -9**	**Us16 = -7**	**Us17 = -7**	**Us18 = -5**

Csak két állapotban van jutalom, az S10-es (vég)állapotban R10=+21, az S13-as állapotban R13=-15.

Mi lesz a 11-es állapot hasznosságértékének új becslése a módosítás után, ha a leszámítolási tényező* 1, *illetve, ha* 0,1*?

A. **γ=1 esetén Us11=11,1** **γ=0,1 esetén Us11=9,21**

B. **γ=1 esetén Us11=9,75** **γ=0,1 esetén Us11=8,265**

C. **γ=1 esetén Us11=19,49** **γ=0,1 esetén Us11=9,95**

D. **γ=1 esetén Us11=8,5** **γ=0,1 esetén Us11=9,95**

Az összes időbeli különbség eljárásmód alapötlete (kicsit más megfogalmazásban)

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 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:

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

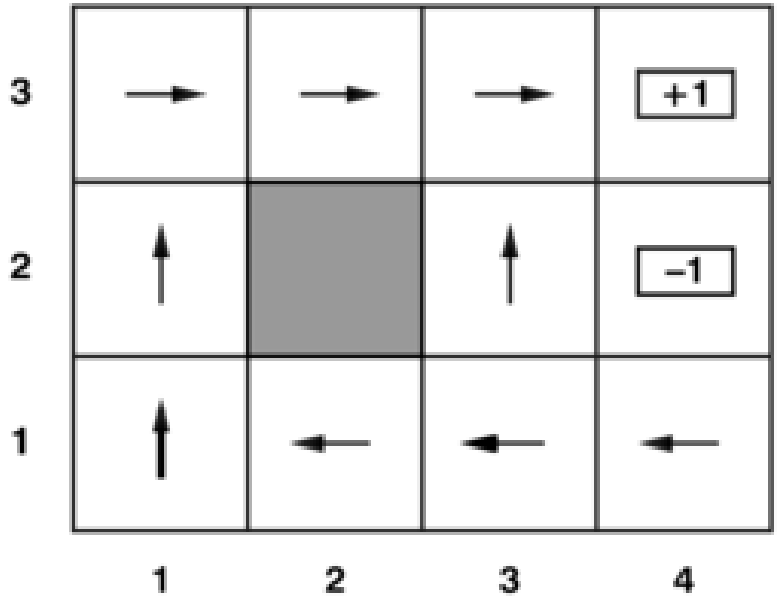
Megjegyzés: Az **ADP** módszer felhasználta a **modell teljes** ismeretét. Az **IK** a szomszédos (egy lépésben elérhető) állapotok közt fennálló kapcsolatokra vonatkozó információt használja, de csak azt, amely az **aktuális tanulási sorozatból származik**.

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

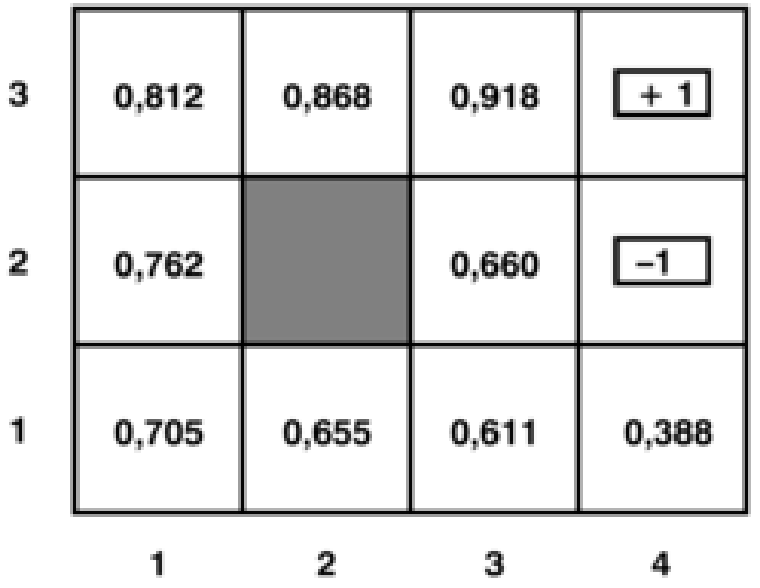
Egy egyszerű demóprobléma

- 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, 0,2 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

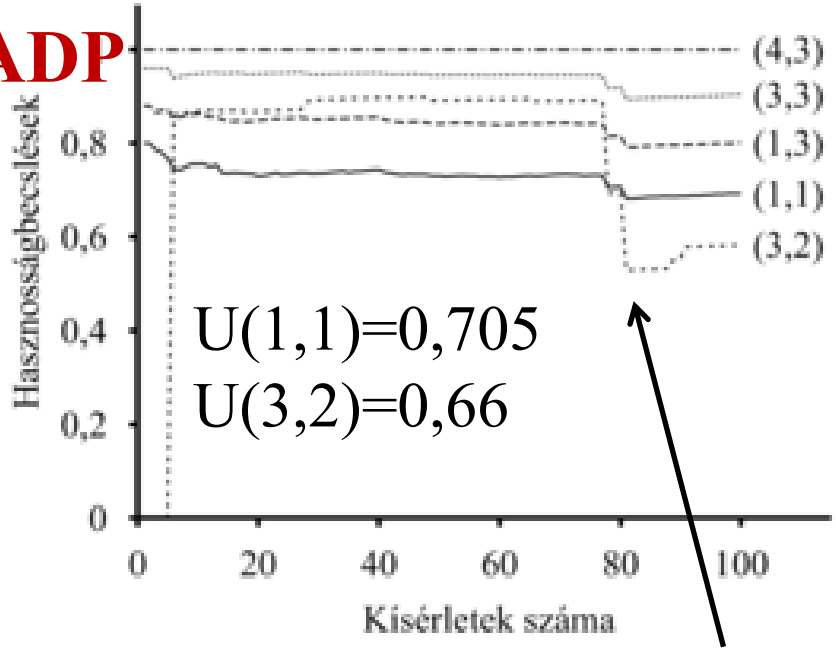


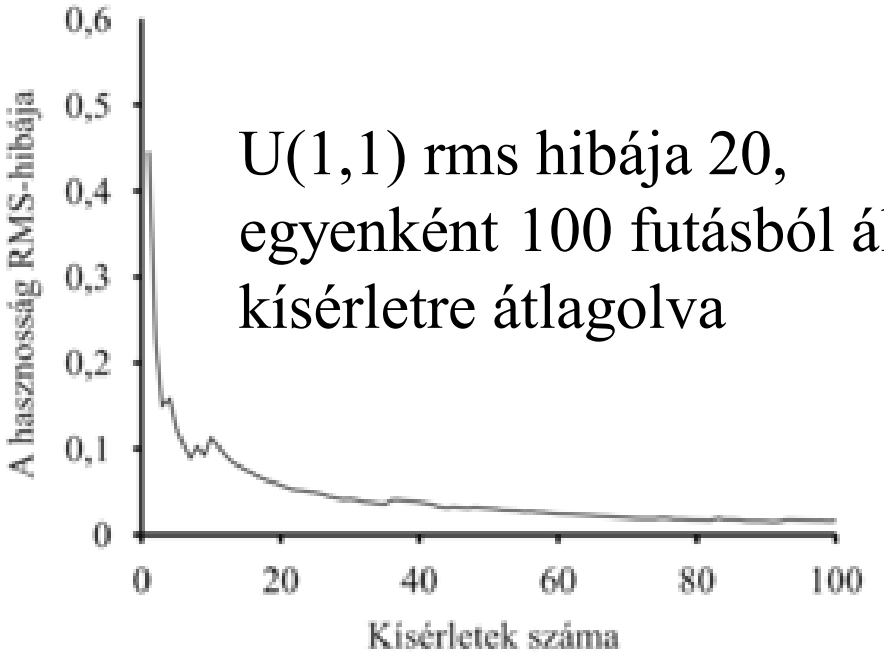
Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok

ADP

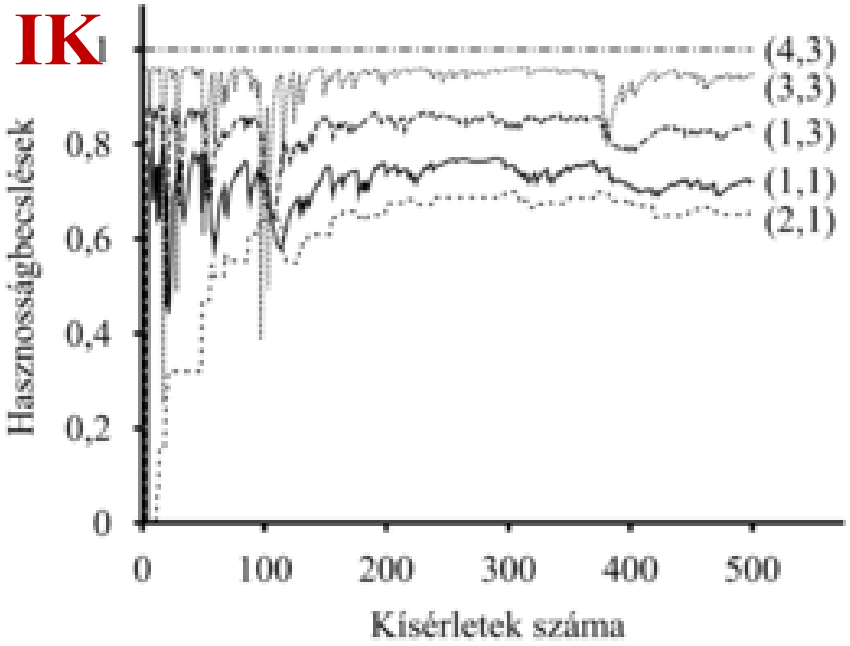


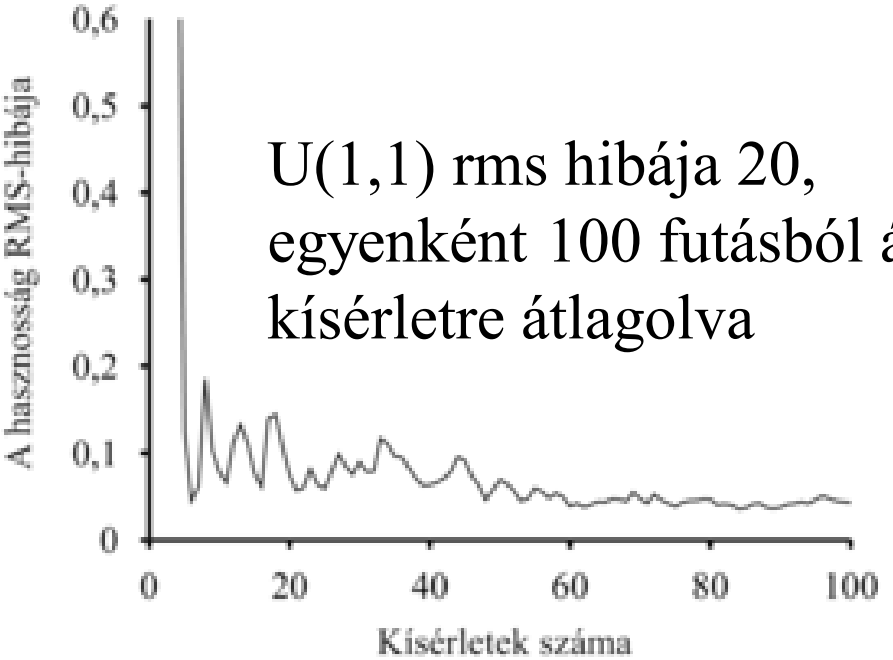


U(1,1) rms hibája 20, egyenként 100 futásból álló kísérletre átlagolva

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 ⇔ jutalmak begyűjtése

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

Nehéz probléma

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

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.

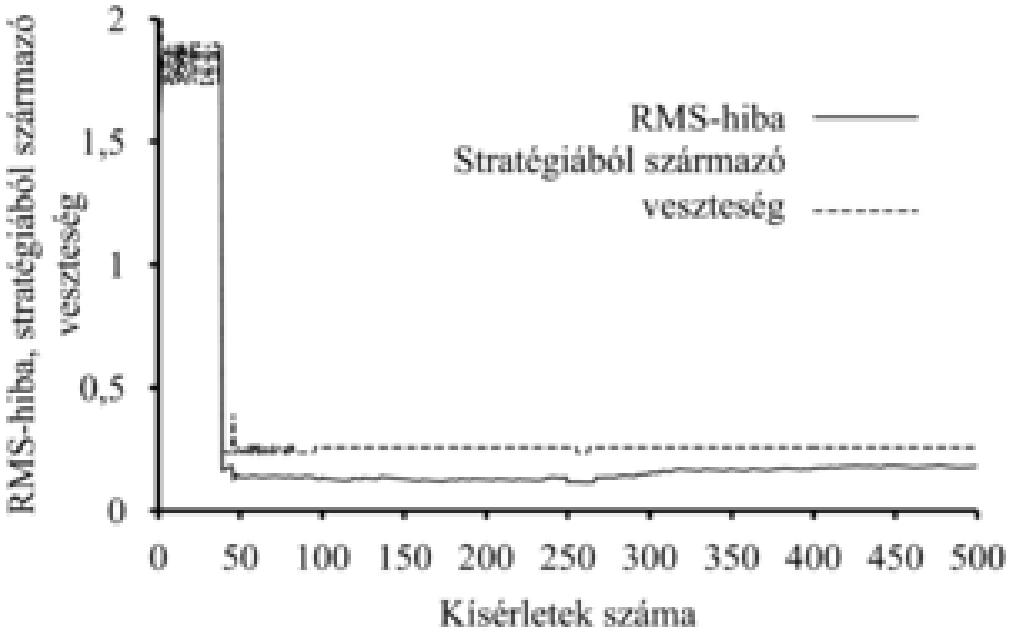
Hóbortos: képes jó hasznosság-becsléseket megtanulni az összes állapotra. Sohasem sikerül fejlődnie az optimális jutalom elérésében.

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

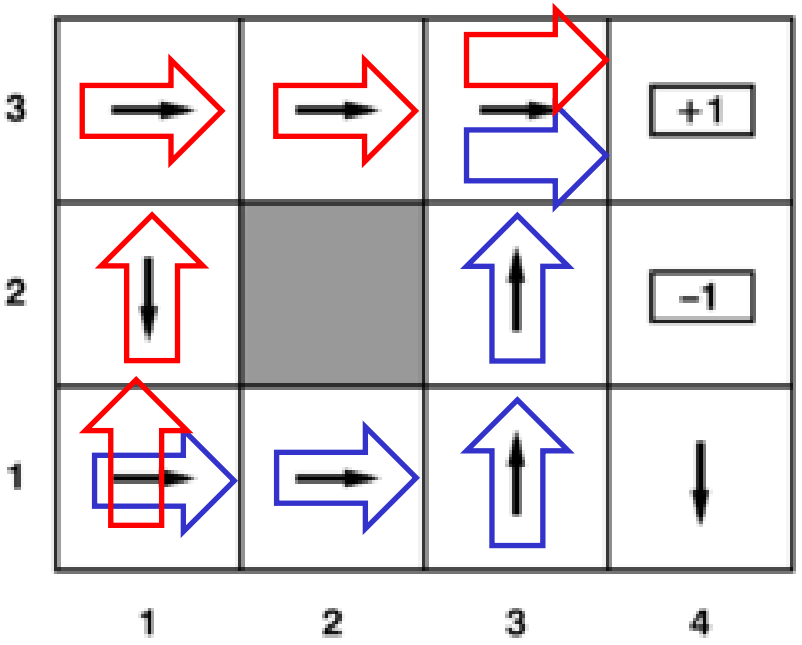
A **mohó ágens** által talált stratégia ⟶

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 már 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>

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

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.

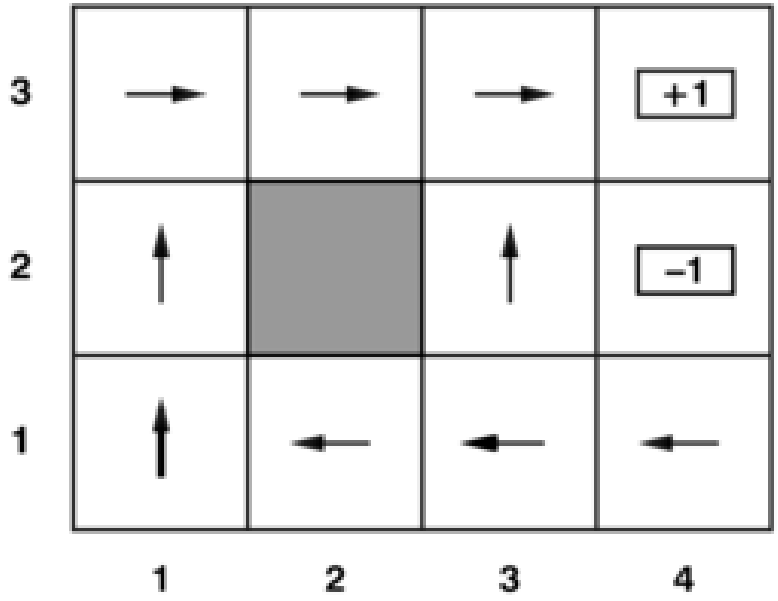
mohóság ↔ kíváncsiság
$f(u,n)$: u-ban monoton növekvő (mohóság),
n-ben monoton csökkenő (a próbálkozások számával csökkenő kíváncsiság)

Ismételjük meg a régi fóliát

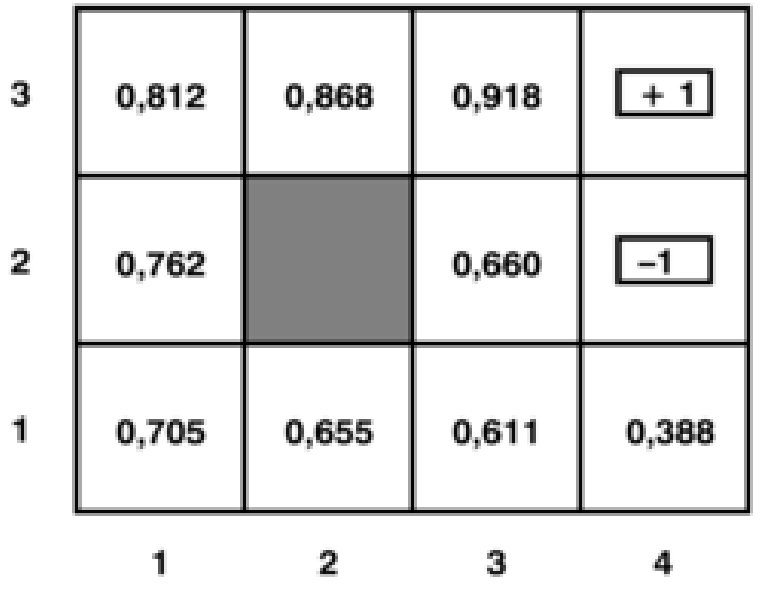
Egy egyszerű demóprobléma:

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

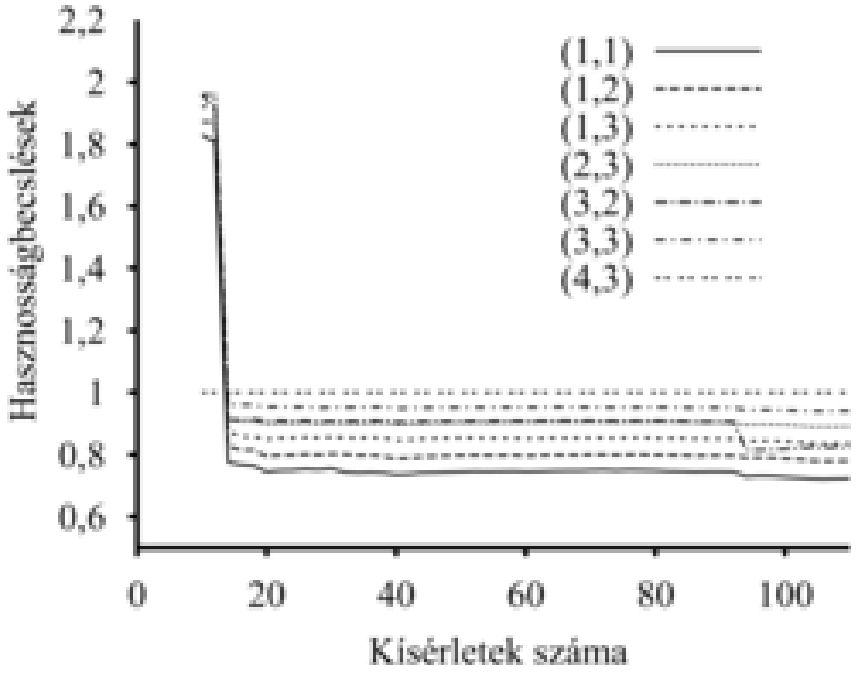
0,8 valószínűséggel a szándékolt irányba lép, 0,1-0,1 valószínűséggel a választott irányhoz képest oldalra. (Falba ütközve nem mozog.)

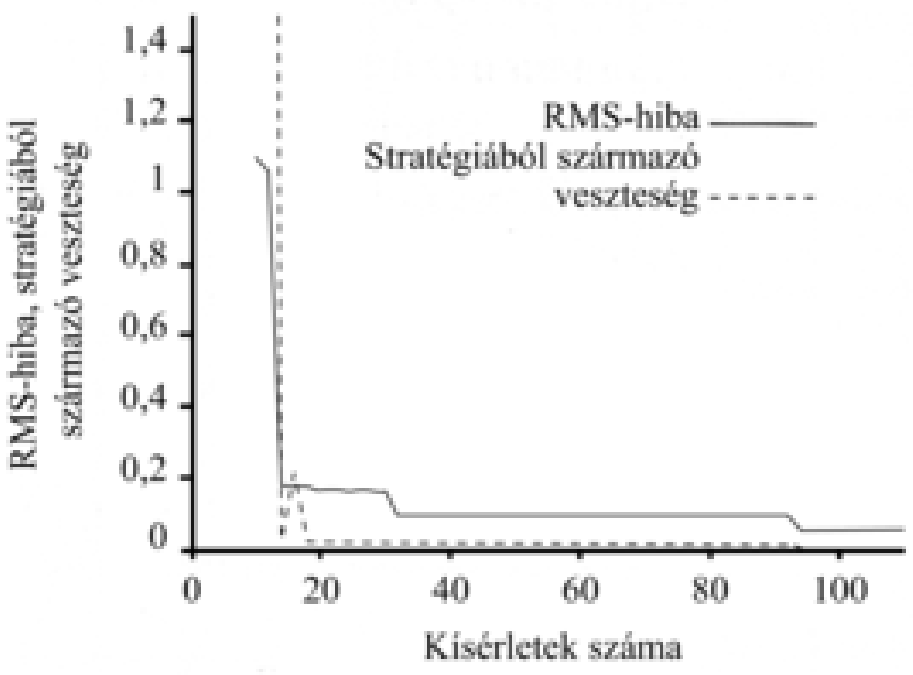


Optimális π^* stratégia (eljárásmód)



Az ezzel kapható hasznosságok





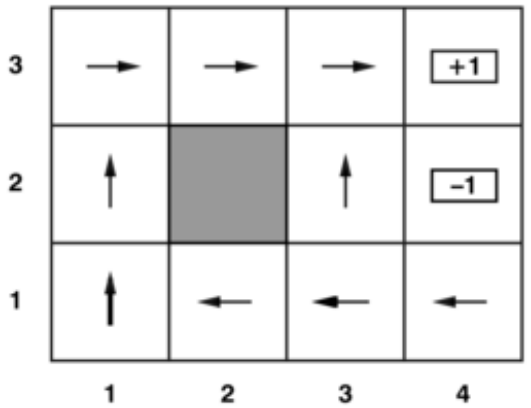
(b)

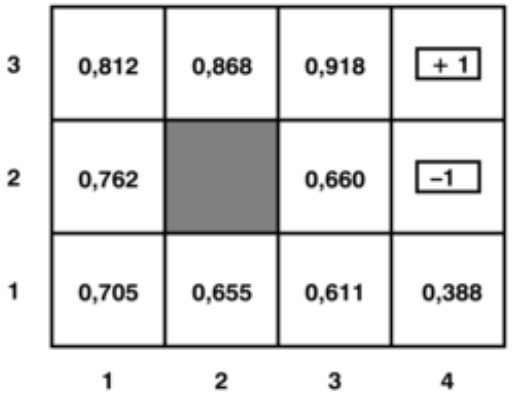
Felfedezési függvény: $R^+=2$, $N_e=5$

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

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

Valós hasznosságok, opt. stratégia





1. 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

3. Felfedezési stratégia - Boltzmann-felfedezési modell

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 $T \to \infty$, akkor a választás tisztán (egyenletesen) véletlen,

ha $T \to 0$, akkor a választás mohó. **Felfedezési stratégia**

A cselekvésérték függvény tanulása

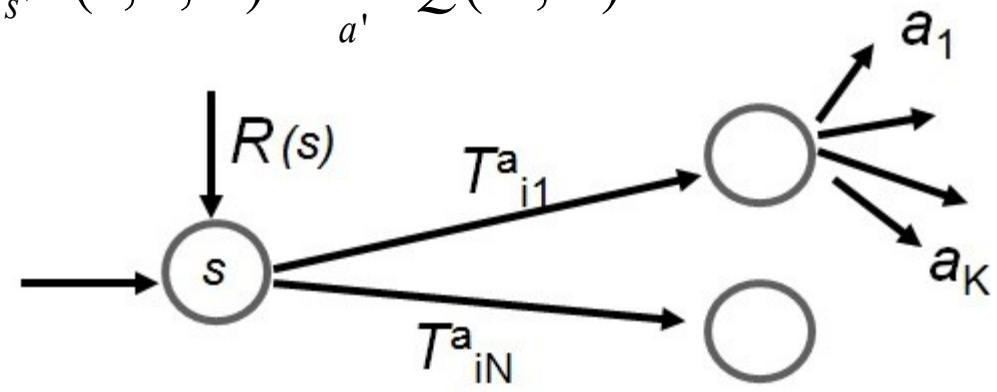
A cselekvésérték függvény egy adott állapotban választott adott cselekvéshez egy várható hasznosságot rendel: **Q-érték**

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

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:

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



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. Boltzmann felfedezési modellből, felfedezési függvényből stb. jöhet

Itt is használható ugyanaz a **felfedezési-függvény** **kvíz következik!**

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 és cselekvéssel indexelünk Q[*s*,*a*]

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

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** $s,a,r \leftarrow$nulla

else $s,a,r \leftarrow s'$, $\operatorname{argmax}_{a'} \boldsymbol{f}(\mathbf{Q}[a',s'], \boldsymbol{N_{sa}}[a',s']), r'$

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.)

Kvíz 12.3. - Aktív Q-tanulás, IK módszer

Az (1,1) – bal alsó sarok – állapotból indultunk, és *a*=FEL-t választottunk. A jelenlegi (a következő dián lévő táblázatokban megadott) $\hat{Q}(a,s)$, $N(a,s)$ becsléseink alapján **melyik cselekvést választjuk, ha az (1,2), illetve ha a (2,1) állapotba jutottunk**? A szokásos felfedezési függvényt használjuk, N_{LIMIT}=2, R^+=5.

$$f(q,n) = \begin{cases} q & \text{ha } n \geq N_{LIM} \\ R^+ & \text{egyébként} \end{cases}$$

(1,3)	(2,3)	(3,3)	(4,3)
(1,2)	xxx	(3,2)	(4,2)
(1,1)	(2,1)	(3,1)	(4,1)

Gyakorlófeladat - Aktív Q-tanulás, IK módszer

a1=FEL, Q(FEL,*s*)

0,8	0,8	0,8	+1
0,75	xxx	0,65	-1
0,68	0,6	0,5	-0,89

***N*(FEL,*s*)**

2	2	2	2
3	xxx	1	1
3	2	1	0

a2=LE , Q(LE,*s*)

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

***N*(LE,*s*)**

1	1	2	2
2	xxx	3	2
4	2	3	1

a3=JOBBRA, Q(JOBBRA,*s*)

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

***N*(JOBBRA,*s*)**

1	1	4	2
1	xxx	5	4
6	4	3	2

a4=BALRA, Q(BALRA,*s*)

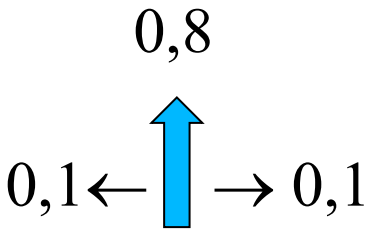
0,5	0,6	0,7	+1
0,5	xxx	0,4	-1
0,47	0,42	0,32	0,3

***N*(BALRA,*s*)**

1	1	2	3
2	xxx	2	1
4	2	2	0

Jutalom csak a végállapotokban! (4,2) és (4,3)

Ha egy adott irányba mutató parancsot adunk, akkor a következő valószínűségek mentén mozdulunk el:



$\gamma=0$, $\alpha=0,1$

Gyakorlófeladat - Aktív hasznosság-tanulás, IK módszer

a1=FEL

0,8	0,8	0,8	+1
0,75	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

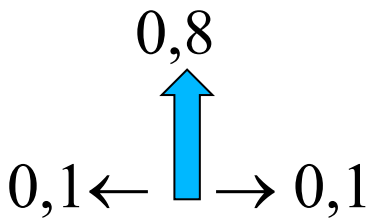
a3=JOBBRA

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:

2A. 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!

(véletlenszám-generátorunk: 0,1576 0,9706 0,9572 0,4854 0,8003 0,1419)

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

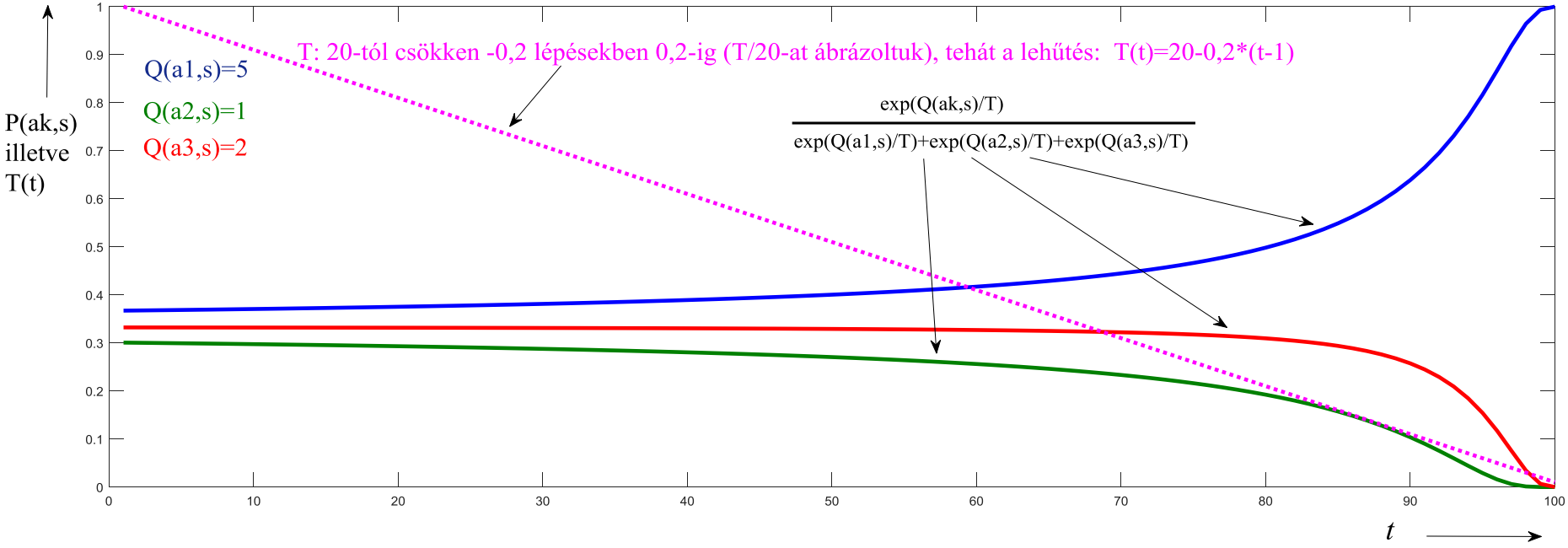
$$\hat{Q}(\ (1,1), FEL)$$

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

Boltzmann felfedezési modell, 3 lehetséges cselekvés, 3 követő állapotba visz. A az s-ben végrehajtott cselekvéseknek – a követő állapotok hasznosságából becsülhető – értéke
Q(a1,s)=5, Q(a2,s)=1, Q(a3,s)=2



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

A megerősítéses tanulás általánosító képessége

Az általánosító képesség nem a megerősítéses tanulás jellemzője – minden tanulásnál fontos. (Már a felügyelt tanulásnál is beszéltünk róla.)

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 10000 állapotból következtetni a teljes térre, sőt a 10001-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:

$$\hat{U}_\theta(x,y)=\theta_0+\theta_1 x+\theta_2 y=$$
$$=0,5+0,2\cdot 3-0,2\cdot 1=0,9$$

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

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

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