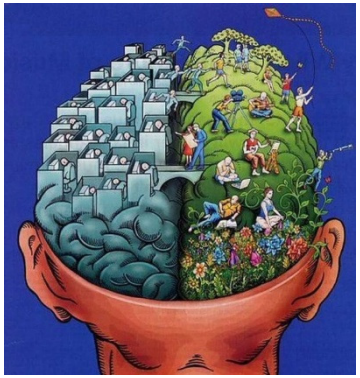
Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

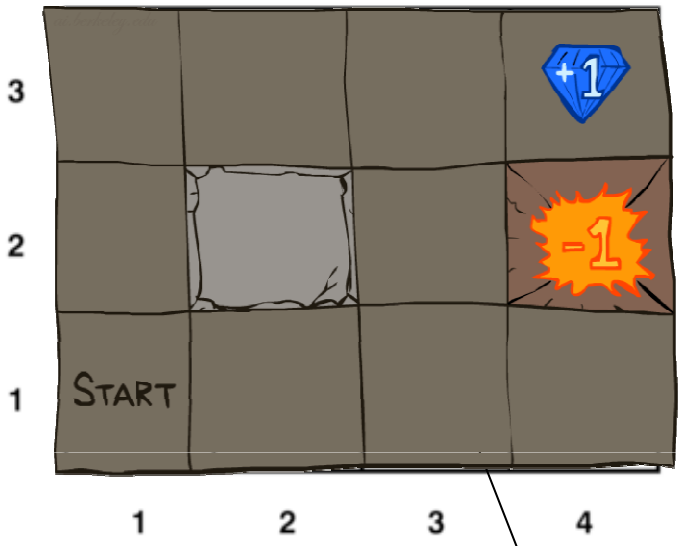


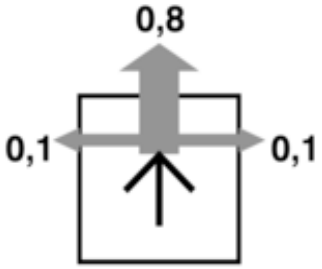
Markov döntési folyamat

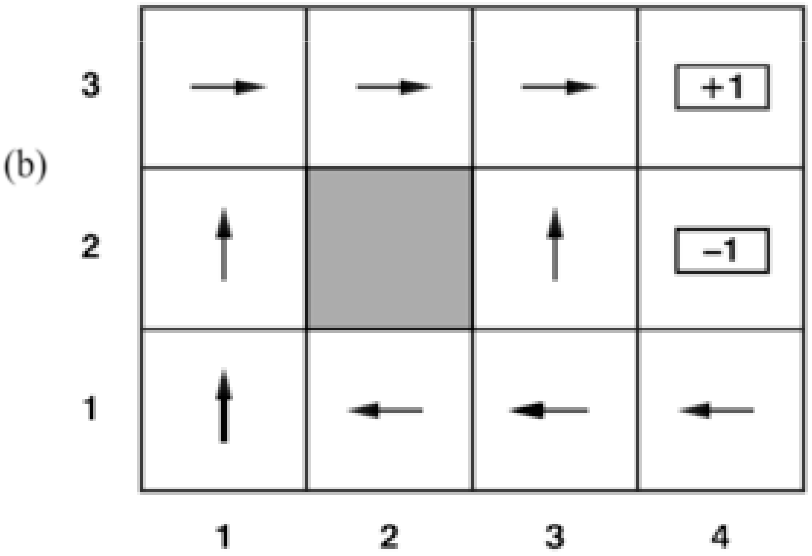
Előadó: Dr. Hullám Gábor

Előadás anyaga: Dr. Dobrowiecki Tadeusz

Szekvenciális döntési probléma







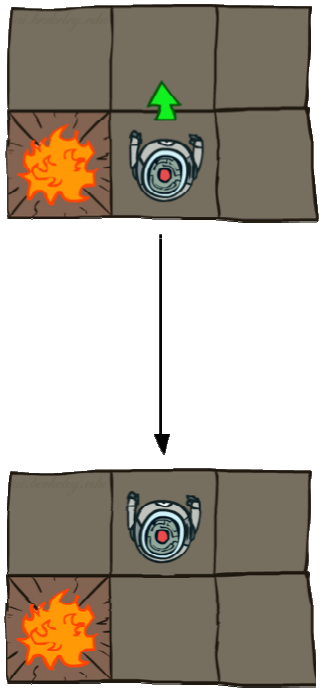
- 0.04

Fix út: fel, fel, jobbra, jobbra, jobbra?

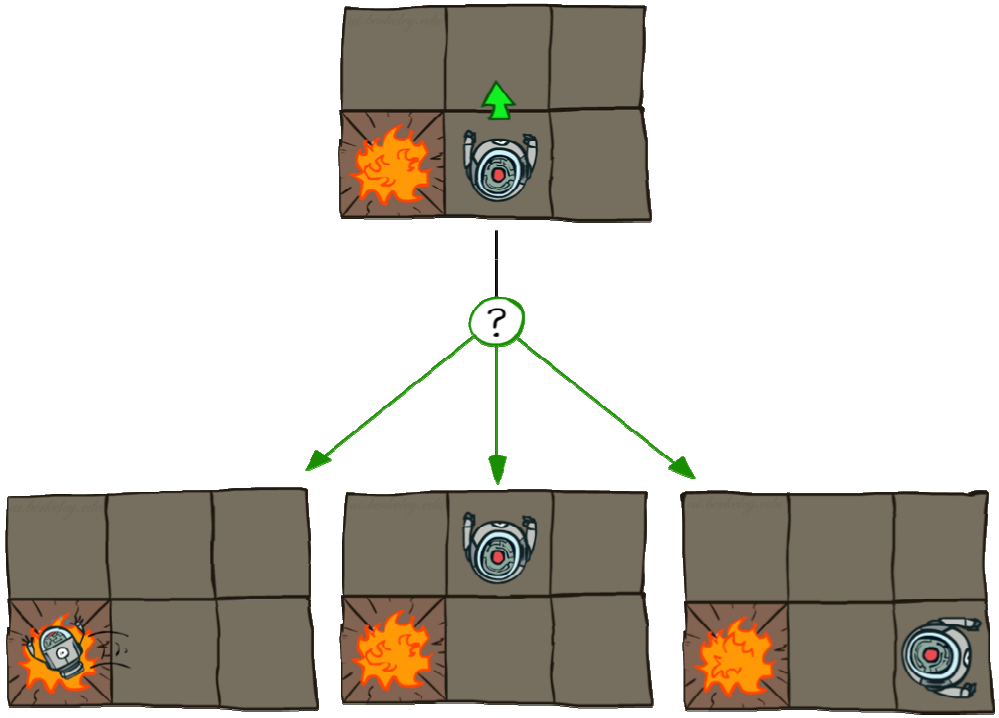
(optimális) Eljárásmód: $\pi(s) = a$

Grid World Cselekvések

Determinisztikus Grid World



Sztochasztikus Grid World



Szekvenciális döntési probléma

Markov döntési folyamat

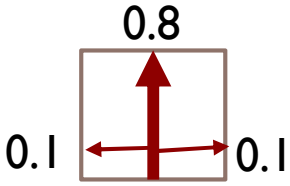
Kezdőállapot:	S0
Állapotátmenet-modell:	T(s, a, s′)
Jutalomfüggvény:	R(s), vagy R(s, a, s′)

Optimális eljárásmód = optimális mozgás, döntés cselekvés megválasztására, de nem elég egyszer, folyamatosan kell, amíg nincs vége (a problémának).

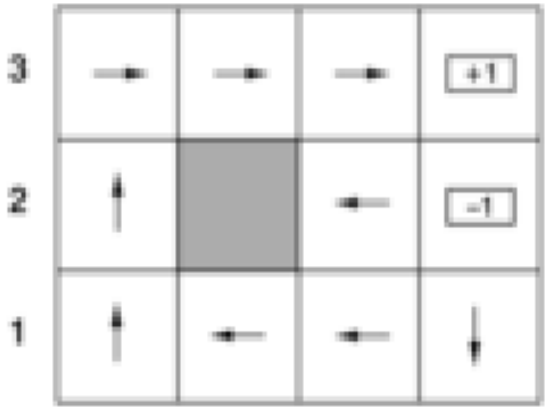
$\pi(s) = a$

$\pi^*(s) = a$

Szekvenciális döntési probléma

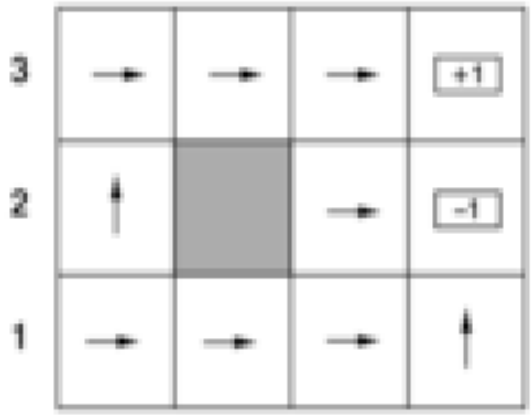


$-0{,}0221 < R(s) < 0$



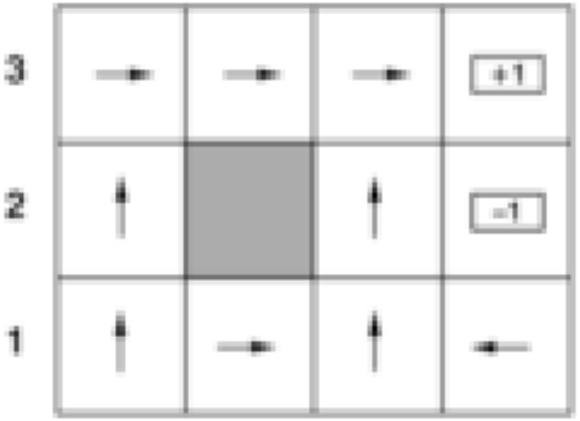
az élet csak kevéssé bánatos,
ne legyen kockázat!

$R(s) \le -1{,}6284$



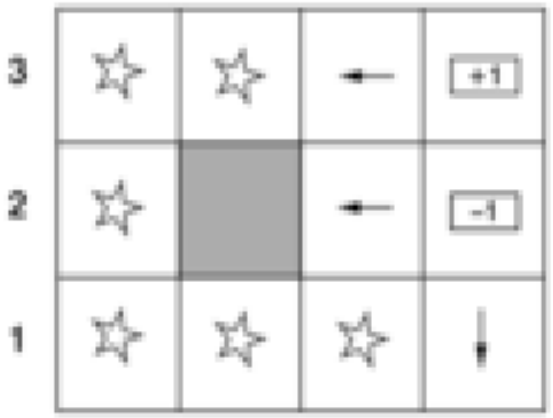
élet elviselhetetlen, ki!

$-0{,}4278 \le R(s) \le -0{,}0850$



élet kellemetlen; +1 állapot,
–1 kockázattal

$R(s) > 0$



élet kifejezetten élvezhető,
az ágens benn akar maradni

Szekvenciális döntési probléma

Optimalis szekvenciális döntési probléma

	optimális eljárásmód
végtelen horizont	stacionárius
véges horizont	nem-stacionárius

Többattribútumú hasznosságelmélet:
ágens preferenciái az állapotsorozatok között stacionáriusok

additív jutalmak $U_h([s_0, s_1, s_2, \ldots]) = R(s_0) + R(s_1) + R(s_2) + \ldots$

leszámított jutalmak

$$U_h([s_0, s_1, s_2, \ldots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots$$

Leszámítolás

- Ésszerű a jutalmak maximalizálására törekedni
- Ésszerű egy jelenlegi jutalmat preferálni egy jövőbeli jutalom helyett
- Egy megoldás: a jutalmak értékei exponenciálisan csökkennek

1

Hasznosság most

γ

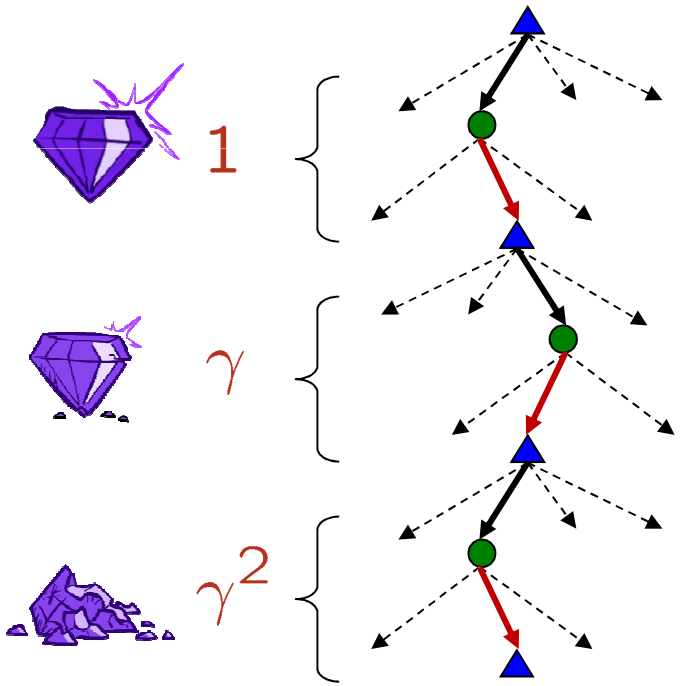
Hasznosság a következő lépésben

γ^2

Hasznosság 2 lépés múlva

Leszámítolás

- Ésszerű a jutalmak maximalizálására törekedni
- Ésszerű egy jelenlegi jutalmat preferálni egy jövőbeli jutalom helyett
- Egy megoldás: a jutalmak értékei exponenciálisan csökkennek



Szekvenciális döntési probléma

Leszámított jutalmak, egy végtelen sorozat hasznossága

$$U_h([s_0, s_1, s_2, \ldots]) = \Sigma_{t=0\ldots\infty} \gamma^t R(s_t) =< \Sigma_{t=0\ldots\infty} \gamma^t R_{\max} = {R_{\max}}/{(1-\gamma)}$$

Alternatív esetek:

1.) Ha van végállapot, és ha garantált, hogy az ágens végül bele kerül, akkor nincs szükség végtelen sorozatok összehasonlítására.
Egy eljárásmód, ami garantáltan végállapotba juttat, véges eljárásmód, $\gamma = 1$

2.) Időegységenkénti átlagjutalom

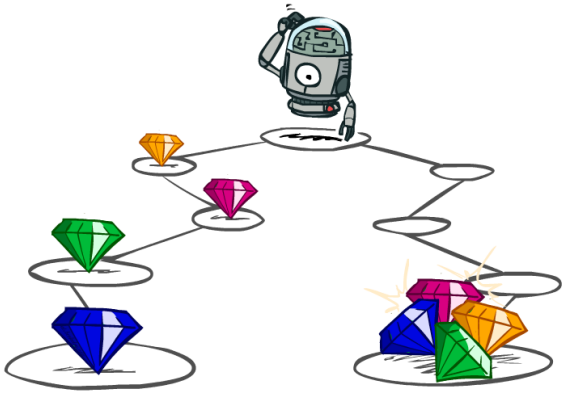
Szekvenciális döntési probléma

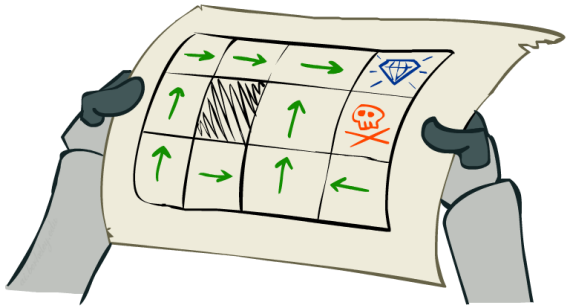
Leszámított jutalmak, egy végtelen sorozat hasznossága

$$U_h([s_0, s_1, s_2, \ldots]) = \Sigma_{t=0\ldots\infty} \gamma^t R(s_t) =< \Sigma_{t=0\ldots\infty} \gamma^t R_{\max} = {R_{\max}}/{(1-\gamma)}$$

Optimális eljárásmód

$$\pi^* = \arg\max_{\pi} E[\Sigma_{t=0\ldots\infty} \gamma^t R(s_t) \,|\, \pi]$$





Optimális eljárásmód meghatározása - Értékiteráció

Egy **állapot hasznossága –** a belőle kiinduló állapotsorozatok várható hasznossága

Az állapotsorozatok függnek a végrehajtott eljárásmódtól, így elsőként egy adott π eljárásmódra definiáljuk a hasznosságot:

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^{t} R(s_t) \,|\, \pi, s_0 = s\right]$$

Optimális eljárásmód

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

Optimális eljárásmód meghatározása - Értékiteráció

Az **állapot hasznossága** - az állapotban tartózkodás közvetlen jutalmának és a következő állapot várható leszámított hasznosságának az összege, feltéve, hogy az ágens az optimális cselekvést választja

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

Bellman egyensúlyi **egyenlet**

Legyen $\gamma = 1$ és a nem végállapotoknál $R(s) = -0{,}04$

Nézzük meg a 4 × 3-as világ Bellman-egyenleteinek egyikét.

Az (1, 1) állapothoz tartozó egyenlet:

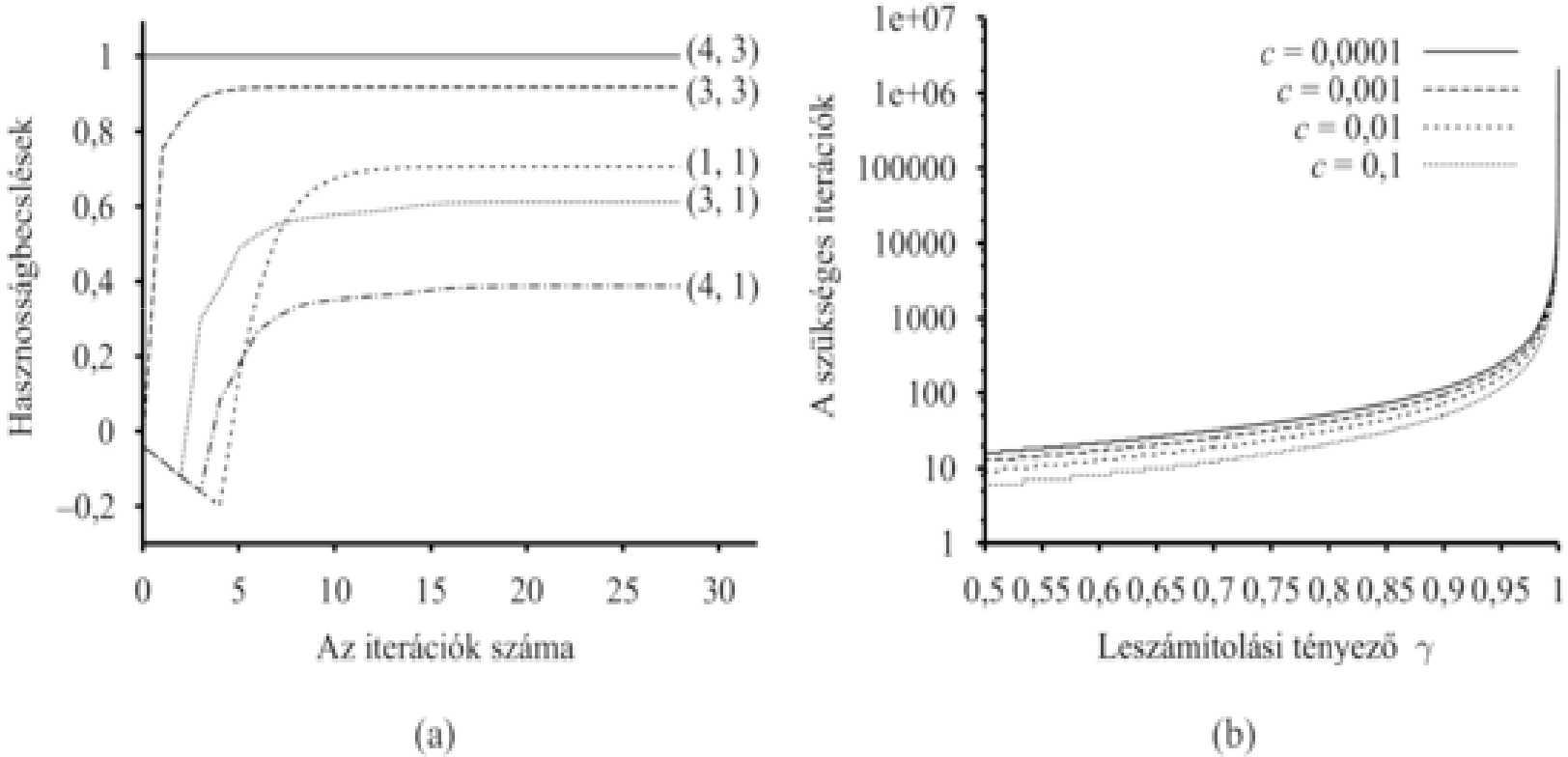
$$U(1,1) = -0{,}04 + \gamma \max\{$$
$$0{,}8\,U(1,2) + 0{,}1\,U(2,1) + 0{,}1\,U(1,1) \quad \text{(Fel)},$$
$$0{,}8\,U(2,1) + 0{,}1\,U(1,2) + 0{,}1\,U(1,1) \quad \text{(Jobbra)},$$
$$0{,}9\,U(1,1) + 0{,}1\,U(1,2) \quad \text{(Balra)},$$
$$0{,}9\,U(1,1) + 0{,}1\,U(2,1)\} \quad \text{(Le)}$$

Sajnos nemlineáris – iteráció!

3	0,812	0,868	0,918	+1
2	0,762		0,660	−1
1	0,705	0,655	0,611	0,388
	1	2	3	4

Bellman-frissítés:

$$U_{t+1}(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s') U_t(s')$$



A hasznosságok fejlődése

A szükséges értékiterációk száma, hogy a hiba garantáltan legfeljebb $\varepsilon = c\, R_{max}$ legyen

Az értékiteráció konvergenciája – **kontrakció** $U_{i+1} = B\,U_i$

$$\|U\| = \max_s |U(s)|$$

$$\|BU_i - U^*\| \le \gamma \|U_i - U^*\|, \qquad U^* \text{ az igazi: } B(U^*) = U^*$$

Az összes állapot hasznossága korlátos $\pm R_{max}/(1-\gamma)$ értékkel
A maximális kezdeti hiba $\|U_0 - U^*\| \le 2R_{max}/(1-\gamma)$

Ha $\|U_{i+1} - U_i\| < \varepsilon(1-\gamma)/\gamma$, akkor $\|U_{i+1} - U^*\| < \varepsilon$

Eljárásmód-iteráció

(optimális eljárásmódot is kaphatunk, ha a hasznosság becslése pontatlan - ha egy cselekvés egyértelműen jobb, akkor a releváns állapotok pontos hasznosságát nem szükséges precízen tudnunk)

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

Eljárásmód-értékelés

Egy adott π_i eljárásmódnál számítsuk ki $U_i = U^{\pi i}$-t, az egyes állapotok hasznosságát mintha π_i volna végrehajtva.

lineáris!

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

$$U_{t+1}(s) \leftarrow R(s) + \gamma \sum_{s'} T(s,\pi_t(s),s')U_t(s')$$

Eljárásmód-javítás

Módosított eljárásmód-iteráció

for each s állapotra **in** S **do**

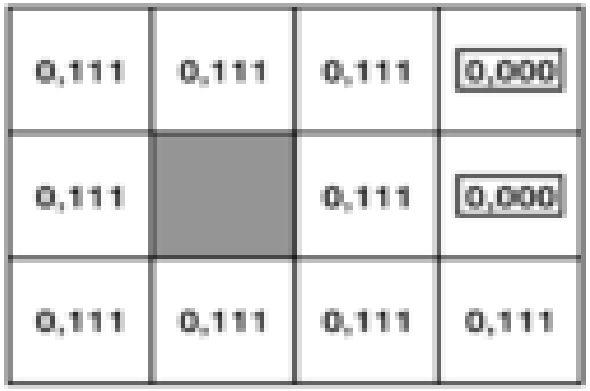
if $\max_a \sum_{s'} T(s,a,s')U[s'] > \sum_{s'} T(s,\pi[s],s')U[s']$ **then**

$$\pi[s] \leftarrow \text{argmax}_a \sum_{s'} T(s,a,s')U[s']$$

Részlegesen megfigyelhető Markov döntési folyamat

Kezdőállapot: S0
Állapotátmenet-modell: T(s, a, s′)
Jutalomfüggvény: R(s), v. R(s, a, s′)
Megfigyelési modell,
az s állapotban az o megfigyelés érzékelésének a valószínűsége
O(s, o)

Hiedelmi állapot = *b*(*s*) = eloszlás állapotok felett



$$\left\langle \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, 0, 0 \right\rangle$$

$$b'(s') = \alpha O(s', o) \sum_{s} T(s, a, s') b(s)$$

(szűrés)

$$P(o \mid a,b) = \sum_{s'} P(o \mid a,s',b)\, P(s' \mid a,b)$$

$$= \sum_{s'} O(s',o) P(s' \mid a,b) = \sum_{s'} O(s',o) \sum_{s} T(s,a,s') b(s)$$

$$\tau(b,a,b') = P(b' \mid a,b) = \sum_{o} P(b' \mid o,a,b)\, P(o \mid a,b)$$

$$= \sum_{o} P(b' \mid o,a,b) \sum_{s'} O(s',o) \sum_{s} T(s,a,s') b(s)$$

$$\rho(b) = \sum_{s} b(s) R(s)$$

RMMDF megoldása a fizikai (véges) állapottérben redukálható egy MDF megoldására a hozzá tartozó hiedelmi állapot térben (val. eloszlások folytonos terében)

Szekvenciális döntési problémából indulunk

Markov döntési folyamat (MDF)

S0
T(s, a, s′)
R(s), v. R(s, a, s′)

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

$$\pi(s) \leftarrow U(s)$$

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s,a,s')U_i(s') \qquad \pi(s) \leftarrow U_i(s)$$

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

Megerősítéses tanulás

Kezdőállapot:	ismert
Állapotátmenet-modell:	**nem ismert, legfeljebb a tapasztalat**
Jutalomfüggvény:	**nem ismert, legfeljebb ahogy jön**
Optimális eljárásmód	**megtanulni, mindennek ellenére?!**

Megerősítéses tanulás

Pl. sakkjáték: tanító nélkül is van visszacsatolás a játék végén:
nyert vagy **vesztett** = +/- **jutalom**, azaz **megerősítés**.

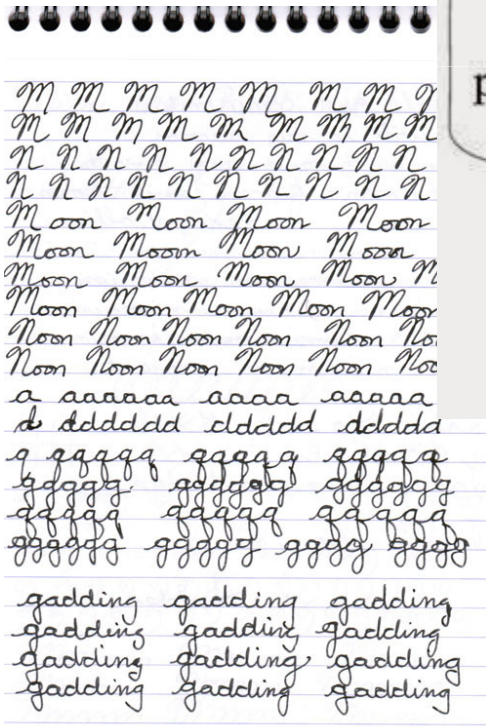
... a játék végén = a cselekvés szekvencia végén

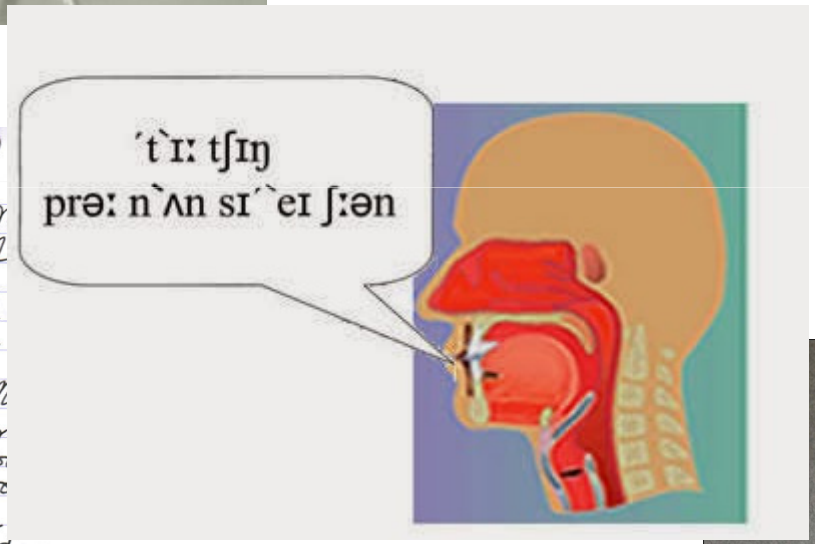
Megerősítés: milyen gyakran? milyen erős? milyen értékű?

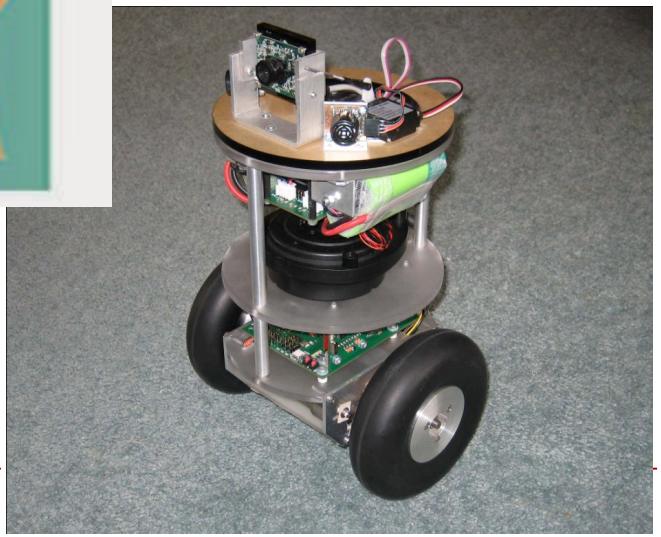
A feladat:

(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**, azt sem, **melyik jutalom melyik cselekvésből ered.**


Moon
Noon
a
d
q
b
p
gadding


ˈtˌɪː tʃɪŋ
prəː nˌʌn sɪˈˌeɪ ʃːən



Ágens tudása:

indulásakor **tudja már** a környezetet és a cselekvéseinek hatását, vagy pedig még ezt is **meg kell tanulnia**.

Megerősítés:

csak a **végállapotban**, vagy menet közben **bármelyikben**.

Ágens:

passzív tanuló: figyeli a világ alakulását és tanul.

aktív tanuló: a megtanult információ birtokában cselekednie is kell.

felfedező …

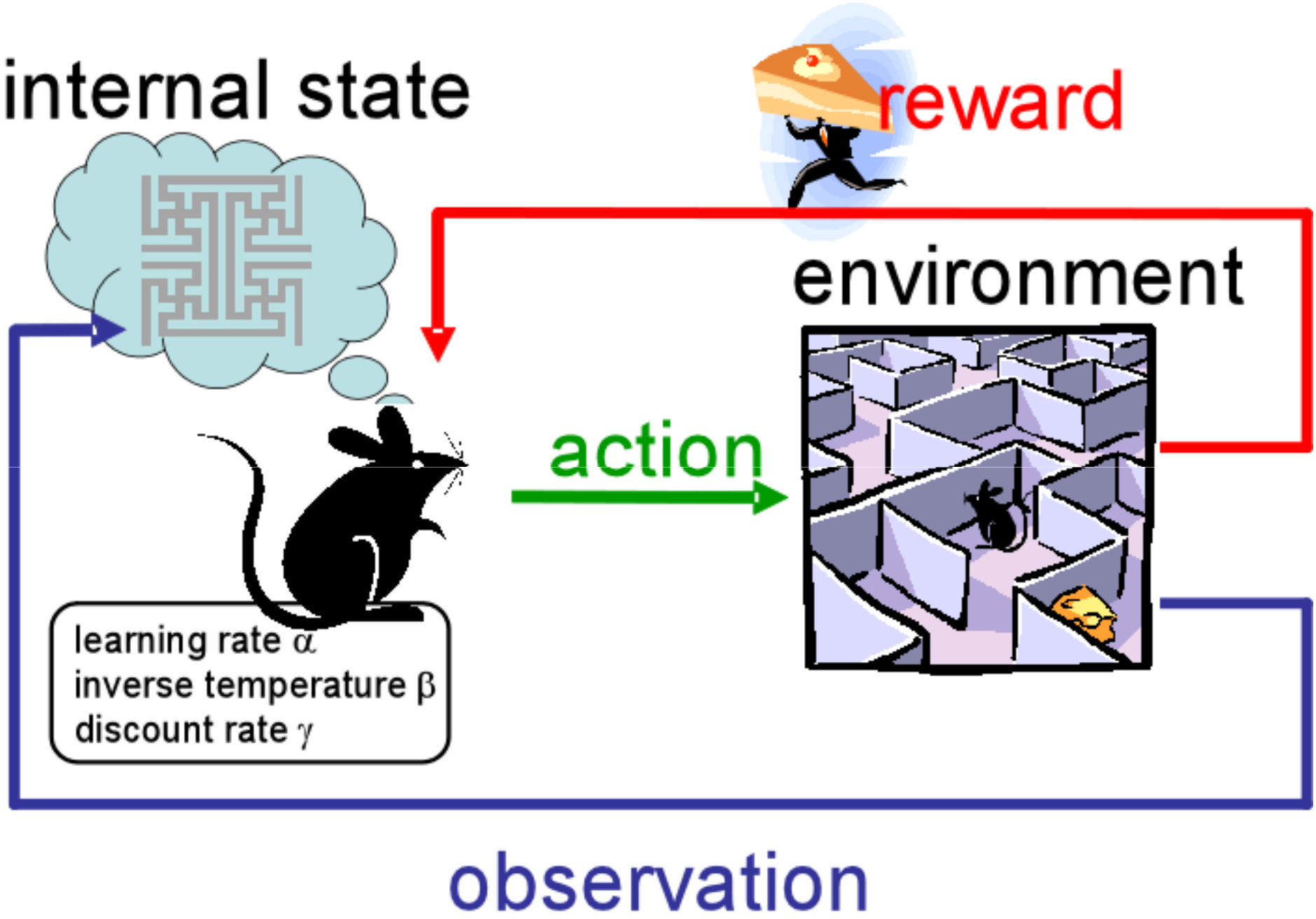
Ágens kialakítása: megerősítés → hasznosság leképezés

U(s) **hasznosság** függvény tanulása, ennek alapján a cselekvések eldöntése, hogy az elérhető hasznosság várható értéke max legyen

(környezet/ágens modellje **szükséges**)

Q(a, s) **cselekvés érték** függvény (állapot-cselekvés párok) tanulása, valamilyen várható hasznot tulajdonítva egy adott helyzetben egy adott cselekvésnek (környezet/ágens modell **nem szükséges**, közben tanult)

– **Q tanulás (model-free)**

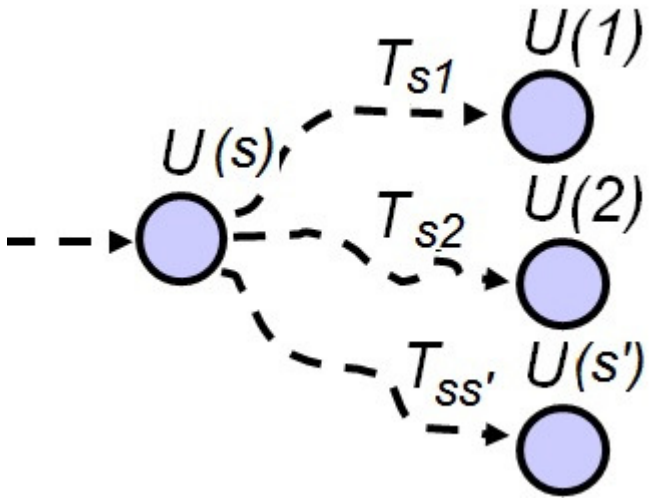

internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

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

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

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

$$U^{\pi}(s) = E[\sum_t \gamma^t R(s_t) \,|\, \pi, s_0 = s]$$



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

Adaptív Dinamikus Programozás

Az állapotátmeneti valószínűségek T*ij* a megfigyelt gyakoriságokkal becsülhetők.
Amint az ágens megfigyelte az összes állapothoz tartozó jutalom értéket, a hasznosság-értékek a következő **egyenletrendszer megoldásával** kaphatók meg:

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

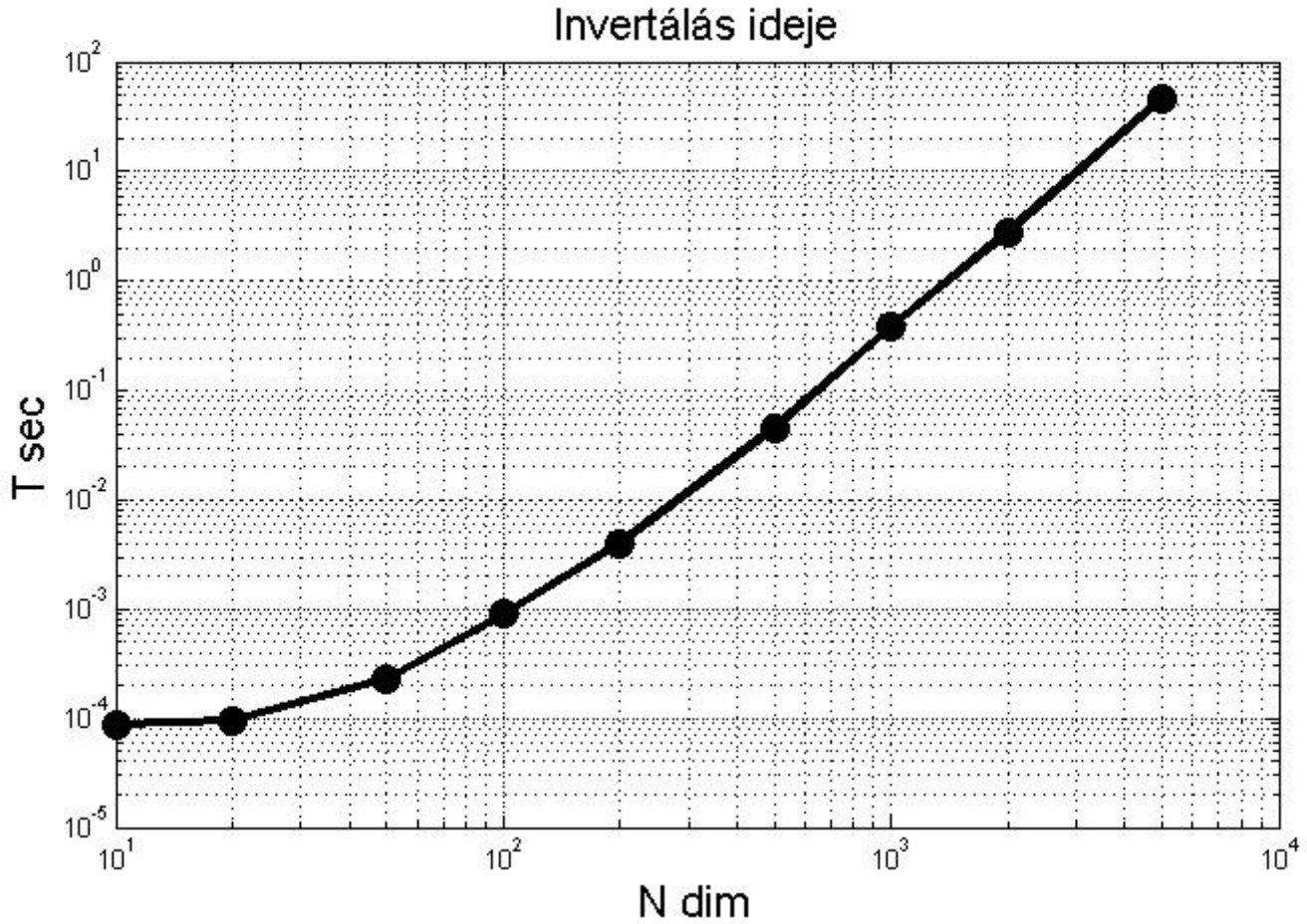
megfigyelt

megtanult

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

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

és ha nem megy?


Invertálás ideje
T sec
N dim
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{1}
10^{2}
10^{3}
10^{4}

Az időbeli különbség (IK) tanulás (TD – Time Difference)

$$U(s) \leftarrow U(s) + \alpha(R^{\circ}(s) - U(s))$$

Monte Carlo mintavétel: használjuk fel a megfigyelt állapotátmeneteket a megfigyelt visszaterjesztett megerősítésekkel
- ki kell várni az epizód végét.

Alapötlet: használjuk fel a megfigyelt állapotátmeneteknél a jutalom becsült értékét:

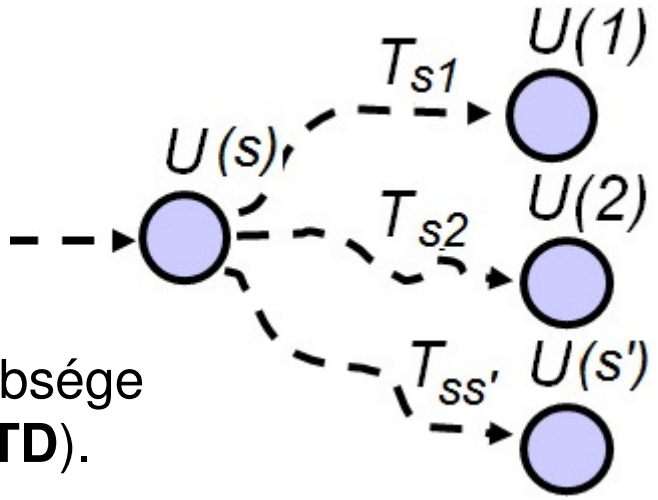
$$R(s) + \gamma U(s')$$

TD(0)-különbség (a jutalom becslése)
TD(0) -hiba:

$$\delta(s) = R(s) + \gamma U(s') - U(s)$$

$$U(s) \leftarrow (1-\alpha)\, U(s) + \alpha\, \delta(s) =$$

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



α - **bátorsági faktor,** tanulási tényező
γ - **leszámoltatási** tényező (ld. MDF)

Az egymást követő állapotok hasznosság-különbsége **időbeli különbség** (**IK**) (**temporal difference**, **TD**).

Az összes időbeli különbség eljárásmód alapötlete

1. korrekt hasznosság-értékek esetén **lokálisan fennálló feltételek rendszere**:

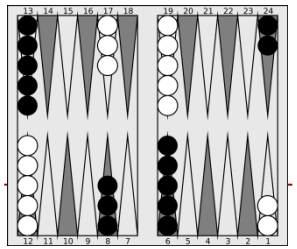
$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s') U^{\pi}(s')$$

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

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

3. Csak a **(A)DP** módszer használta fel a **modell teljes** ismeretét.
Az **IK** az 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**.



Ismeretlen környezetben végzett aktív 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**.

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

Az állapottér felderítése - milyen cselekvést válasszunk? **Nehéz probléma**

Helyes-e azt a cselekvést választani, amelynek a jelenlegi hasznosság-becslés alapján legnagyobb a várható hasznossága?

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

Döntésnek kétféle hatása van:

1. Jutalmat eredményez a **jelenlegi** szekvenciában.
2. Befolyásolja az észleléseket, és ezáltal az ágens **tanulási képességét** – így jutalmat eredményezhet a **jövőbeni** szekvenciákban.

Az állapottér felderítése

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

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

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

Létezik-e optimális felfedezési stratégia?

Ágens súlyt kell adjon azoknak a cselekvéseknek, amelyeket még nem nagyon gyakran próbált, a kis hasznosságúnak gondolt cselekvéseket elkerülje.

felfedezési függvény

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

mohóság ↔ kíváncsiság

f(*u*,*n*) *u*-ban monoton növekvő,
n-ben monoton csökkenő

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

$U^+(i)$: 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.

ε-mohóság: az ágens ε valószínűséggel véletlen cselekvést választ, ill. 1-ε valószínűséggel mohó

Boltzmann-felfedezési modell

egy a cselekvés megválasztásának valószínűsége
egy s állapotban:

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

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.

A cselekvés-érték függvény tanulása

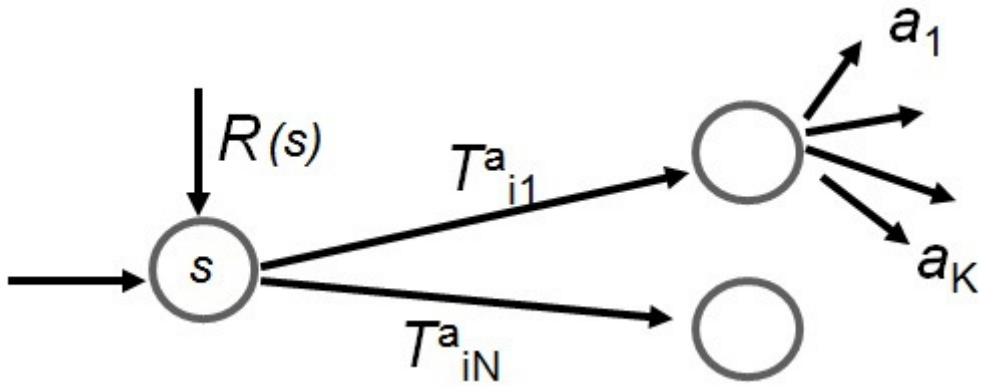
A cselekvés-érték függvény egy adott állapotban választott adott cselekvéshez egy várható hasznosságot rendel: **Q-érték**

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

A Q-értékek fontossága:
a feltétel-cselekvés szabályokhoz hasonlóan lehetővé teszik a döntést **modell használata** nélkül,
ellentétben a feltétel-cselekvés szabályokkal, **közvetlenül** a jutalom visszacsatolásával **tanulhatók**.

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

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



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.

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

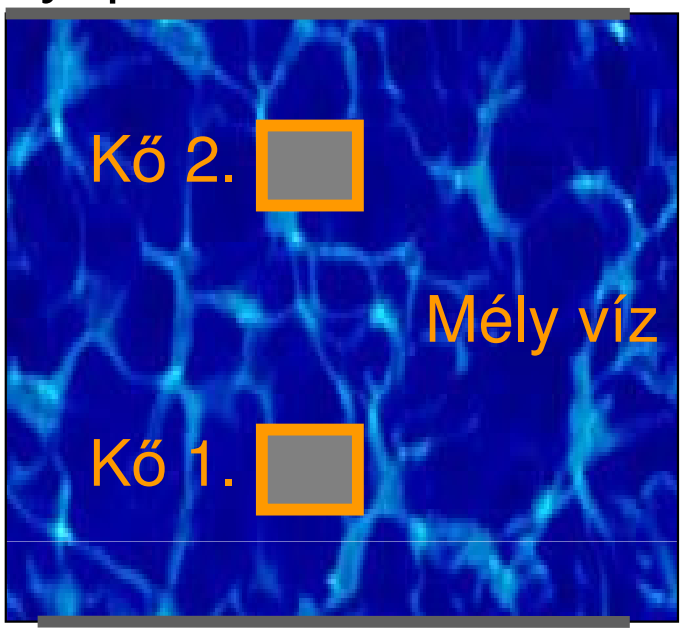
$$Q(a,s) \leftarrow Q(a,s) + \alpha\,(R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s))$$

SARSA Q-tanulás (State-Action-Reward-State-Action)

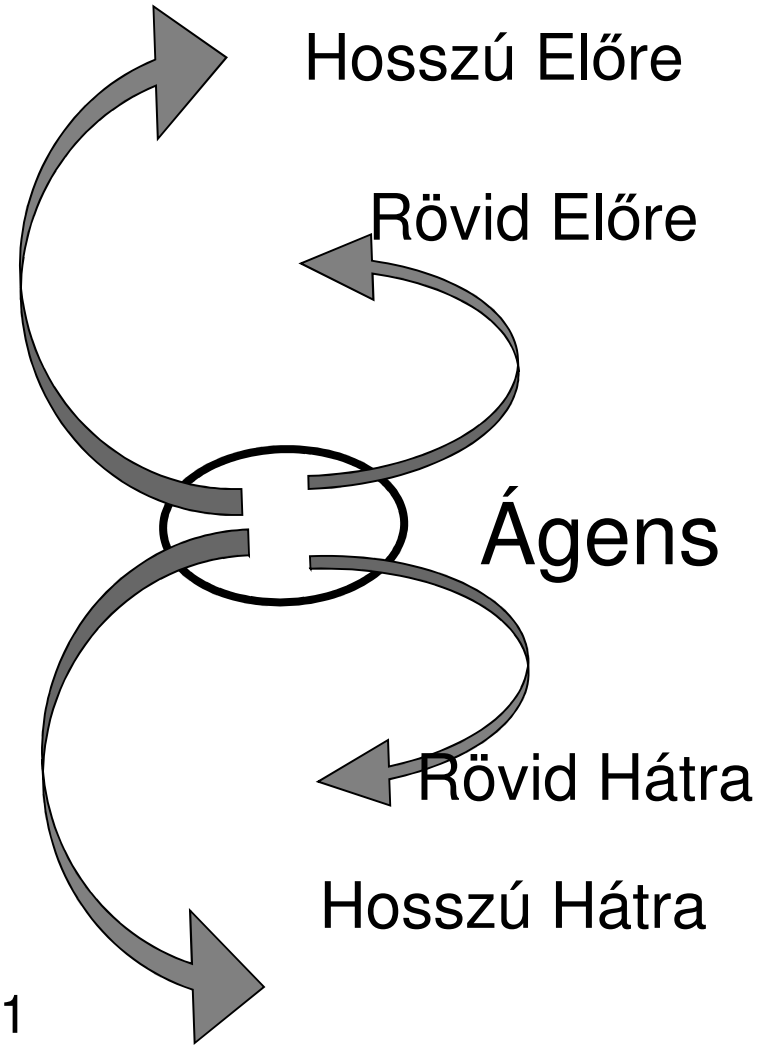
$$Q(a,s) \leftarrow Q(a,s) + \alpha[R(s) + \gamma Q(a',s') - Q(a,s)]$$

(a' megválasztása pl. Boltzmann felfedezési modellből)

Folyópart B



Folyópart A



Tanuló szekvencia:

RE→ HE→ RE→ +1 (száraz lábbal át)

RE→ HE→ HH→ HE→ HH→ RH→ HE→ -1

RE→ HE→ RH→ -1 (megfürdött)

……

Part B

Kő 2. 

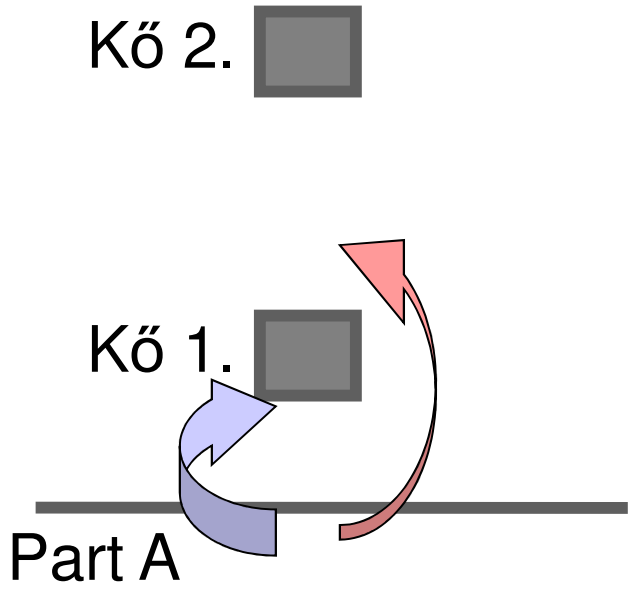
$$Q(a,s) \leftarrow Q(a,s) + \alpha\left(R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s)\right)$$

Kő 1. 

Part A

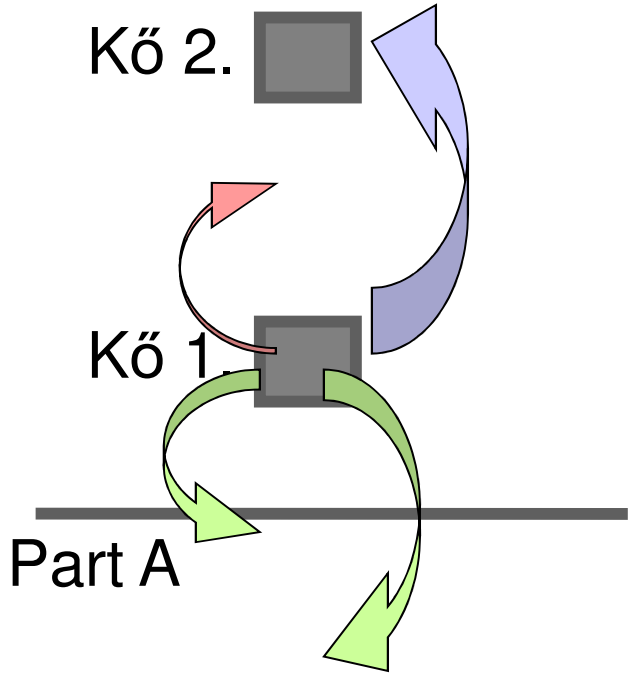
		Cselekvés			
		HH	RH	RE	HE
Állapot	Part A	0	0	-0.035	-0.877
	Kő 1.	-0.520	-0.550	-0.835	0.555
	Kő 2.	-0.204	-0.897	1.180	1.158

Part B

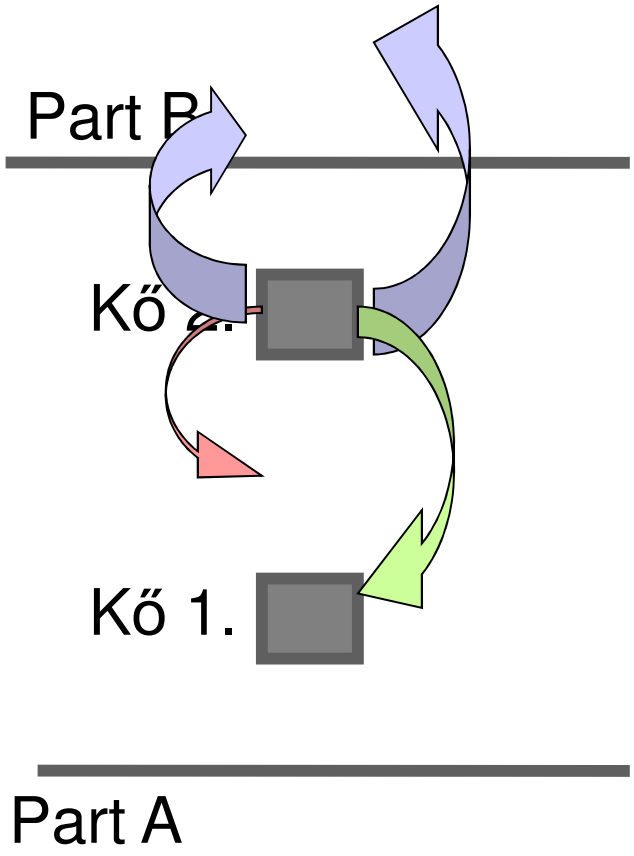


	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158

Part B



	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158



	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158

A megerősítéses tanulás általánosító képessége

Eddig: ágens által tanult hasznosság: **táblázatos** formában reprezentált
= **explicit reprezentáció**

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

Gondosan vezérelt közelítő ADP: 10000, vagy ennél is több.
De a való világhoz közelebb álló környezetek szóba se jöhetnek.
(a sakk stb. - állapottere 10^{50}-10^{120} nagyságrendű)
(az összes állapotot látni, hogy megtanuljunk játszani?!)

Egyetlen lehetőség: a függvény **implicit reprezentációja**:

Pl. egy játékban a táblaállás tulajdonságainak valamilyen halmaza $f_1,\ldots f_n$.
Az állás becsült hasznosság-függvénye:

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

C. Shannon, **Programming a Computer for Playing Chess**, Philosophical Magazine, 7th series, 41, no. 314 (March 1950): 256-75
A.L. Samuel, **Some Studies in Machine Learning Using the Game of Checkers. B.II-Recent Progress**, IBM. J. 3, 211-229 (1959), kb. 20 tulajdonság
Deep Blue (Feng-Hsiung Hsu) kb. 8000 tulajdonság, spec. sakk-csip

A hasznosság-függvény *n* értékkel jellemezhető, pl. 10^{120} érték helyett. Egy átlagos sakk kiértékelő függvény kb. 10 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

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

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

Jósolt és kísérleti tényleges

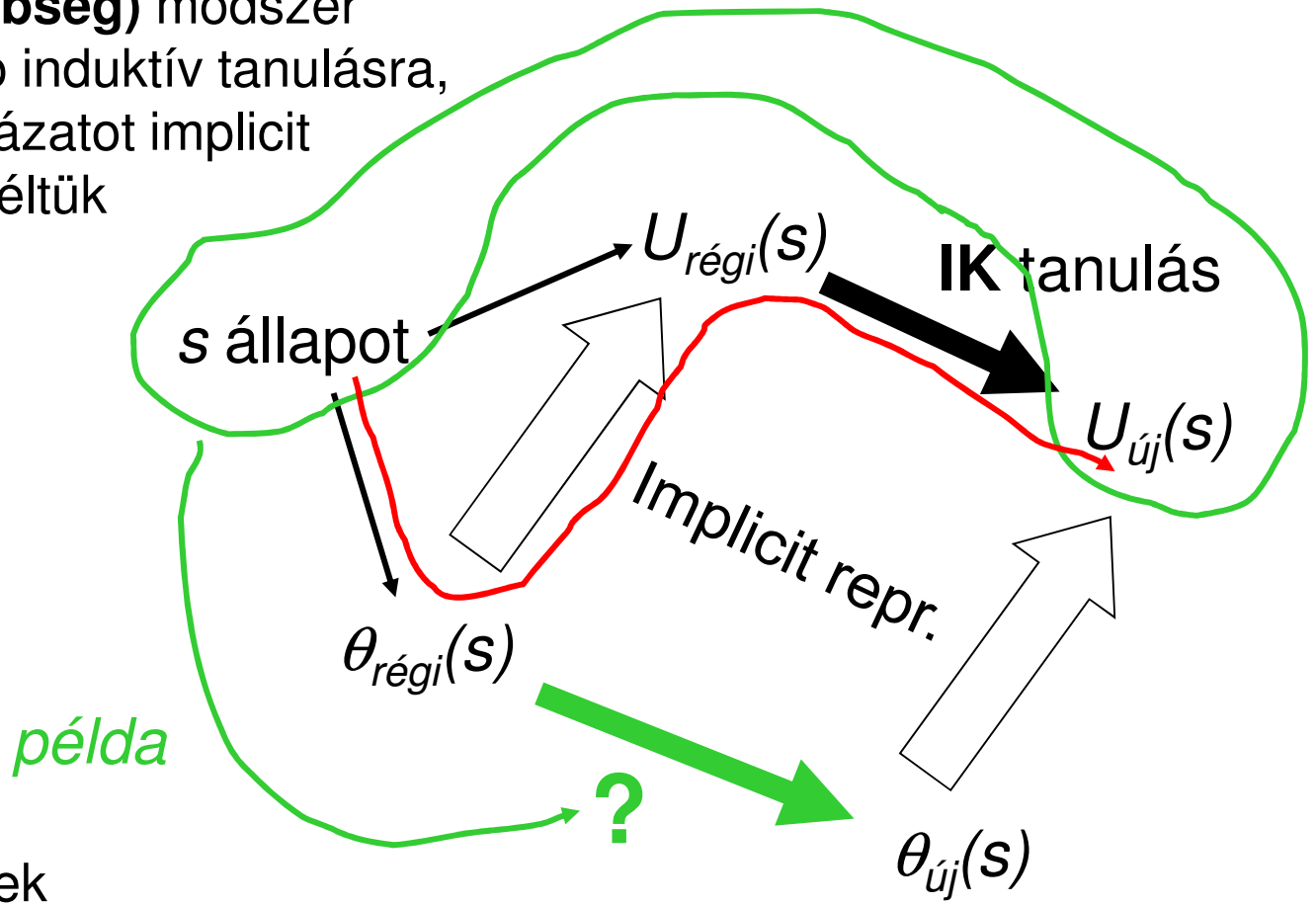
$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial E_j(s)}{\partial \theta_i} = \theta_i + \alpha\left(u_j(s) - \hat{U}_\theta(s)\right)\frac{\partial \hat{U}_\theta(s)}{\partial \theta_i}$$

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

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

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

Az **IK (időbeli különbség)** módszer szintén alkalmazható induktív tanulásra, ha az *U,* vagy *Q* táblázatot implicit reprezentációra cseréltük (pl. neurális háló).



U/*Q* tanuló gradiensek

$$\theta_i \leftarrow \theta_i + \alpha\left[R(s) + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)\right]\frac{\partial \hat{U}_\theta(s)}{\partial \theta_i}$$

$$\theta_i \leftarrow \theta_i + \alpha\left[R(s) + \gamma \max_{a'} \hat{Q}_\theta(a', s') - \hat{Q}_\theta(a, s)\right]\frac{\partial \hat{Q}_\theta(a, s)}{\partial \theta_i}$$