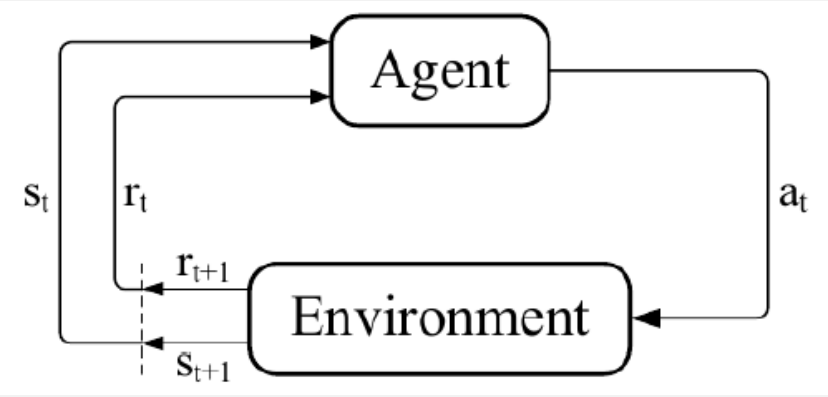
Tanulás elosztott rendszerekben/3

Multi Agent Reinforcement Learning
Többágenses megerősítéses tanulás (MARL)



Kezdjünk egy ágenssel.

Legyenek a környezeti **állapotai** *s*-ek, **cselekvései** *a*-k, az ágens cselekvéseit meghatározó **eljárásmód** π, ill. ágens **cselekvés-érték** függvénye $Q(s,a)$.

Az állapotok és a cselekvések közötti kapcsolatot az un. **Markov döntési folyamat (MDF)** írja le, $T(s,a,s')$ **átmenet-valószínűségel**. Egyes állapotokban ágens $r(s,a,s')$ közvetlen **megerősítést** kap.

Ágens célja megállapítani azt az optimális eljárásmódot, ami a **diszkont hátralévő jutalmat** (az s_k állapottól végtelen jövőbe) maximálja, ahol γ a diszkont faktor és *r* a megerősítés.

$$R_k = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{k+j+1}\right\}$$

Adott eljárásmód mellett az ágens **cselekvés-érték** függvényt tanul

$$Q^{\pi}(s,a) = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{k+j+1} \mid s_k = s, a_k = a, \pi\right\}$$

Multi Agent Reinforcement Learning
Többágenses megerősítéses tanulás (MARL)

A lehető legjobb eredmény az optimális cselekvés-érték függvény:

$$Q^*(s,a) = \max_\pi Q^\pi(s,a)$$

ami teljesíti az un. **Bellman egyenletet**:

$$Q^*(s,a) = \sum_{s' \in S} T(s,a,s') \left[r(s,a,s') + \gamma \max_{a'} Q^*(s',a') \right]$$

Az ágens eljárásmódja **mohó**: ami optimális, ha Q is optimális.

$$\bar{\pi}(s) = \arg\max_a Q(s,a)$$

A Bellman-egyenlet ismeretlen *r* és *T* mellett az un. **Q-tanulással** oldható meg (jelen formában **időkülönbség Q-tanulással**):

$$\begin{aligned} Q^{új}(s_k,a_k) &= Q(s_k,a_k) + \alpha \left[r + \gamma \max_{a'} Q(s_{k+1},a') - Q(s_k,a_k) \right] \\ &= (1-\alpha) Q(s_k,a_k) + \alpha \left[r + \gamma \max_{a'} Q(s_{k+1},a') \right] \end{aligned}$$

Multi Agent Reinforcement Learning
Többágenses megerősítéses tanulás (MARL)

Q-tanulás bizonyos feltételek mellett optimális Q-hoz konvergál.

A legfontosabb feltétel, hogy a **tanuló ágensnek nem nulla valószínűséggel ki kell próbálnia minden létező cselekvését**.

Nem tud tehát csak mohó lenni, a mohóságát **felfedezési** igénnyel kell vegyítenie.

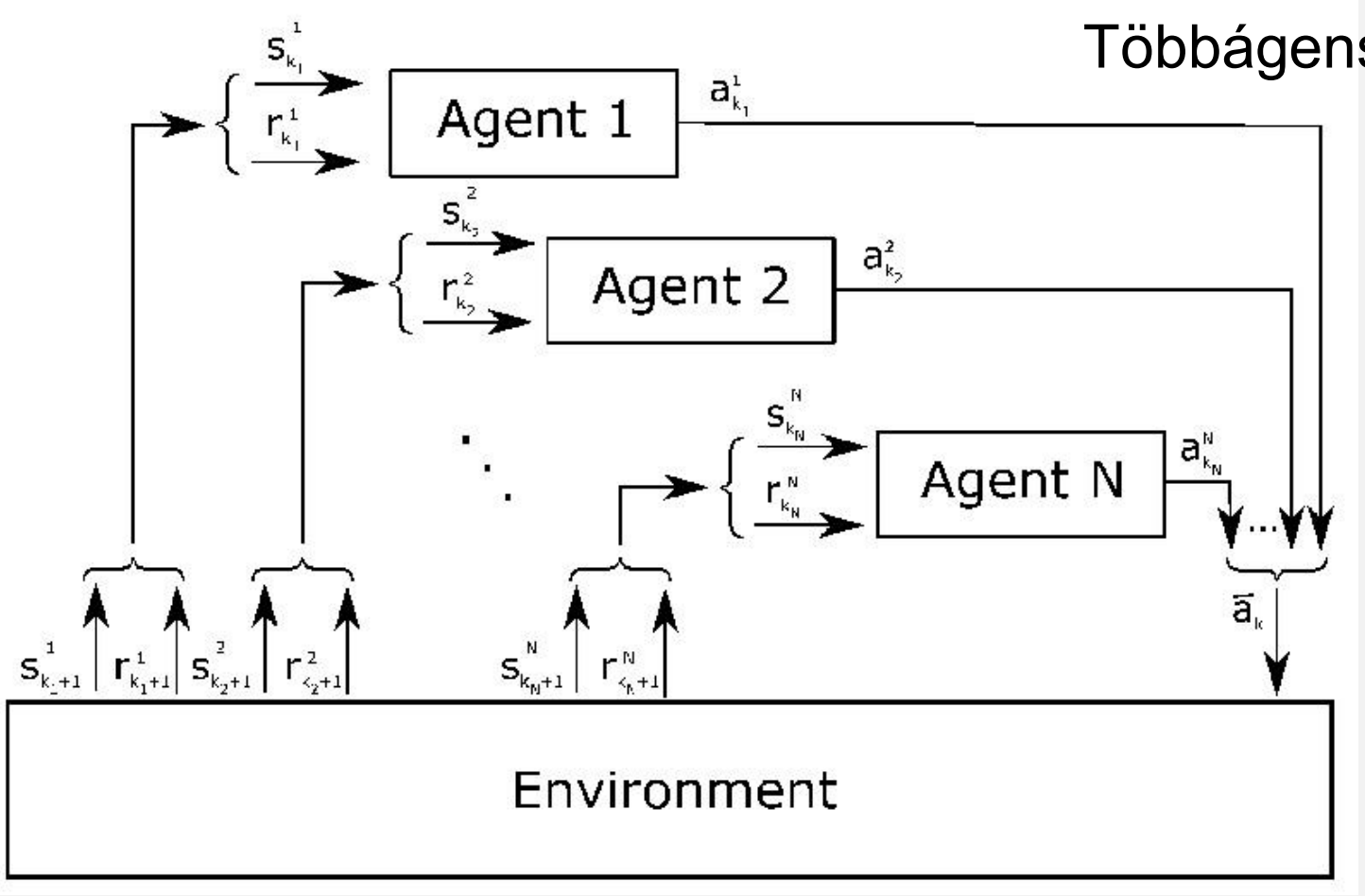
A „mohóság + felfedezés” keverékviselkedést biztosítani tudjuk:

- **ε-mohósággal**: az ágens ε valószínűséggel véletlen cselekvést választ, ill. 1-ε valószínűséggel mohó, vagy
- **Boltzmann-felfedezési modellel**, ahol egy *a* cselekvés megválasztásának valószínűsége egy *s* állapotban:

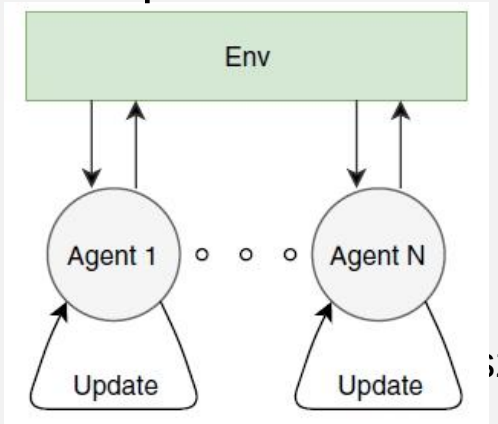
$$\pi(s,a) = \frac{e^{Q(s,a)/T}}{\sum_{a'} e^{Q(s,a')/T}}$$

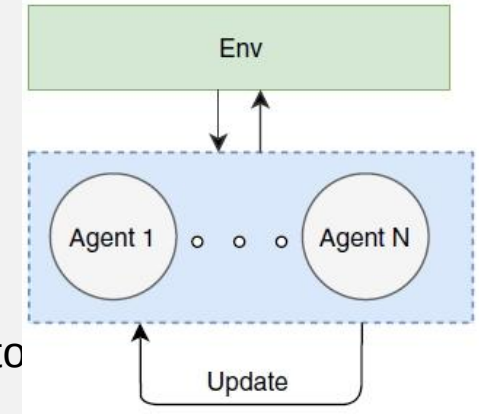
ahol a T „hőmérséklet” a két véglet között szabályoz.
ha $T \to \infty$, akkor a választás tisztán (**egyenletesen**) **véletlen**,
ha $T \to 0$, akkor a választás **mohó**.

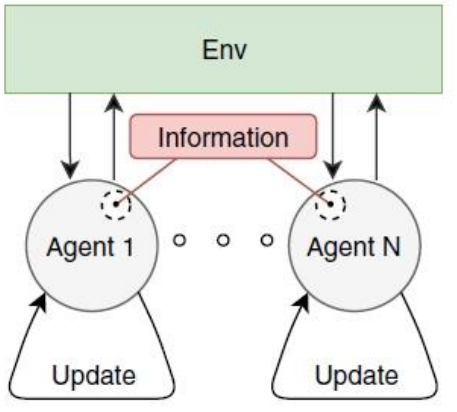
Többágenses eset



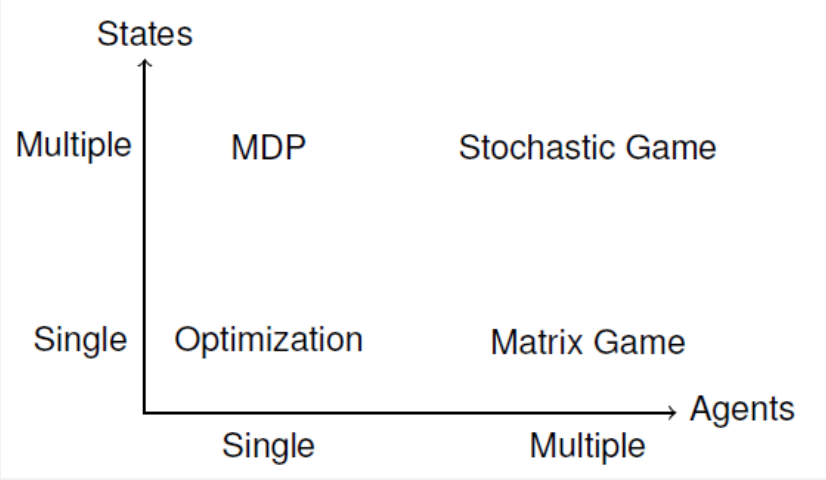
Elosztott/központosított tanulás - Elosztott/központosított végrehajtás







Többágenses eset:



- (**matrix**)**játék**, hasznossági mátrixxal definiált,
- **ismételt játék**, minden fordulóban ugyanazt a mátrixjátékot játsszák,
- **sztochasztikus játék**, a MDF többágenses kiterjesztése. Az állapotátmeneteket és a kapott megerősítést az összes ágens **együttes cselekvése** határozza meg, és az egyedi ágensek eljárásmódjai mellett beszélünk az **együttes eljárásmódról** is.
  Mindegyik állapotban az ágensek új mátrix játékot játszanak, aminek mátrixát a tanult hasznosságok határozzák meg.

$$\left(N, S, A = \underset{k \in N}{\times} A_k, T, \{R_k\}\right), \quad T : S \times A \times S \to [0,1], R_k : S \times A \to \Re$$

Megjegyzés:
- Egy mátrix játékban mindegyik ágens megerősítése/ hasznossága függ az állapottól és az összes ágens együttes cselekvésétől (joint action, joint learners)
- MDF a sztochasztikus játék egyágenses esete
- Ismételt játék a sztochasztikus játék egyetlenegy állapotú esete.

$R_i(s, \mathbf{a})$

Játék lehet **modell-alapú**. Akkor az ágens először megtanulja az ellenfél stratégiáját, majd talál rá a legjobb választ.

Lehet **model-nélküli** is, amikor az ágens az ellenfelélre jó választ adó stratégiát tanulja meg anélkül, hogy az ellenfél stratégiáját explicite kitanulná.

Jelölje egy-egy ágens megerősítését generáló függvényt ρ_i. Beszélhetünk akkor

- **teljesen kooperatív** ágensrendszerekről $\rho_1 = \text{---} = \rho_n$

-**teljesen versengő** ágensrendszerekről, ill. $\rho_1 = -\rho_2$ (két ágens, zérus összegű)

- $\rho_1 + \rho_2 + \ldots \rho_n = 0$, több ágens esetén

- **vegyes** ágensrendszerekről (általános összegű, ahol semmilyen feltétel nem adható).

$$\rho_1 + \rho_2 + \ldots + \rho_n > 0$$

Minden zérus-összegű mátrixjátéknak van NE-ja tiszta stratégiákban. (Neumann)

Minden általános összegű mátrixjátéknak van NE-ja. (Nash)

Minden teljesen versengő sztochasztikus játéknak van NE-ja (Shapley)

Minden általános (vegyes) sztochasztikus játéknak van NE-ja. (Fink)

Többágenses megerősítéses tanulás problémái:

Alapvető problémák:
nem stacionárius – a szokásos (egy ágenses) bizonyítható konvergencia lehetetlen.
koordinálás igénye (pl. több NE esetén)

Mi legyen a tanulás célja?

(1) **Stabilitás** - Konvergencia stratégiában valamilyen egyensúlyhoz (pl. NE), ha a saját maga ellen játszik (**self-play**, minden ágens u.a-t az algoritmust használja)

(2) **Adaptivítás** - Az ellenfél stratégiájának sikeres megtanulása.

(3) Egy bizonyos hasznossági szintet **túlhaladó** nyereségek megszerzése.

Milyen tulajdonságokkal rendelkezzen egy tanulási algoritmus?

(1) **Biztonságos** – garantáljon legalább a minimax szintű nyerességet.

(2) **Konzisztens** – legalább ilyen jó, mint az egyensúlyi esetre számított legjobb válasz (**best response**).

(**P1**) **Konvergencia** - Konvergáljon egy stacionárius eljárásmódhoz.

(**P2**) **Racionalitás** - Ha az ellenfél egy stacionárius stratégiához konvergál, a tanulónak a legjobb válaszhoz kell konvergálnia.

Mechanizmus A legjobb válasz tanulása
Egyensúly tanulása

Egyensúlyok pro és kontra

Pro

Egy egyensúly feltételeket fogalmaz meg, hogy a tanulásnak mikor le kell, le kellene állnia.

Egyensúlyban egyszerűbb a játék (kevesebb erőforrás, számítás)

Kontra

NE nem egy előírás jellegű.

Több egyensúly is lehet (koordinálás).

Valamilyen konkrét egyensúlyra vonatkozó ötletadás nem lehetséges (vagy csalás)

Az ellenfél nem biztos, hogy az egyensúlyra vágyik (és játszik).

NE egyensúly számításának komplexitása általános játékok esetén tiltó lehet.

Egy ágenstől több ágensig

$$Q^{új}(s_k, a_k) = Q(s_k, a_k) + \alpha \left[r + \gamma \max_{a'} Q(s_{k+1}, a') - Q(s_k, a_k) \right]$$

Mások cselekvései is

$$Q^{új}(s_k, \mathbf{a}_k) = Q(s_k, \mathbf{a}_k) + \alpha \left[r + \gamma \max_{\mathbf{a}'} Q(s_{k+1}, \mathbf{a}') - Q(s_k, \mathbf{a}_k) \right]$$

Valami más, ami a cselekvések baráti, vagy adverz jellegére utal és eszerint számítja ki a várható jövőt.

$$Q^{új}(s_k, a_k) = Q(s_k, \mathbf{a}_k) + \alpha \left[r + \gamma [XYZ] - Q(s_k, \mathbf{a}_k) \right]$$

1 ágens ⇨ 2 ágens ⇨ N ágens

Teljes együttműködés $\rho_1 = \ldots = \rho_n$

Optimális együttes Q értékek parallel tanulása

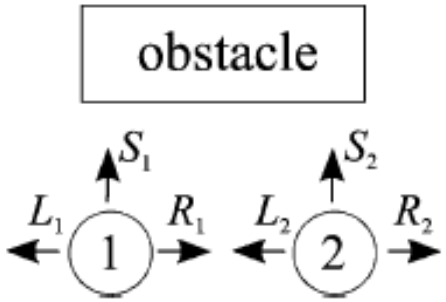
$$Q^{új}(s_k, \mathbf{a}_k) = Q(s_k, \mathbf{a}_k) + \alpha \left[r + \gamma \max_{\mathbf{a}'} Q(s_{k+1}, \mathbf{a}') - Q(s_k, \mathbf{a}_k) \right]$$

és belőle egyenkénti optimális eljárásmód származtatása

$$\bar{\pi}_i^*(s) = a = \arg\max_{a_i} \max_{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n} Q^*(s, \mathbf{a})$$

Együttműködés ellenére probléma a koordinálás.

Példa: formáció-mozgás



Q	L_2	S_2	R_2
L_1	10	-5	0
S_1	-5	-10	-5
R_1	-10	-5	10

A két optimális helyzet ellenére, koordinálás hiányában ágensek szuboptimális helyzetben $Q(L_1, R_2)$ végezhetnek. (ha a Q érték közös, mindkét optimális eset egy Nash egyensúly)

$$Q(L_1, L_2) = Q(R_1, R_2) = 10$$

Koordinálás kérdése

Koordinálás-mentes

pl. **Team-Q**: optimális együttes cselekvés egyedi jellegét tételezi fel

Distributed-Q-Learning: lokális *Q* és π tanulás, de az egyedi *Q* frissítése csak akkor, ha növekszik (a közös optimális Q értéket is el fogja kapni).

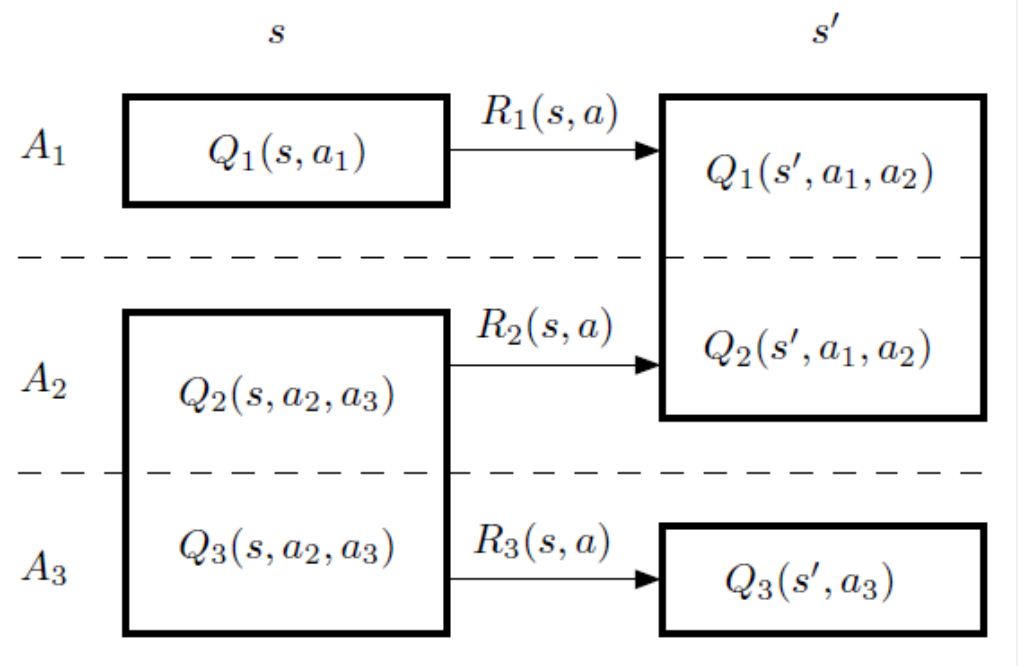
A stratégia frissítése csak akkor, ha a Q érték növekszik.

$$Q_i^{új}(s_k, a_{i,k}) = \max\left\{ Q_i(s_k, a_{i,k}), r + \gamma \max_{a'_i} Q_i(s_{k+1}, a'_i) \right\}$$

Koordinálás kérdése

Koordinálás-alapú pl. együttes Q dekomponálása kisebb csoportosulások szerint. (**koordinációs gráfok**)

$$Q(s,\mathbf{a}) = Q_1(s,a_1) + Q_2(s,a_2,a_3) + Q_3(s,a_2,a_3)$$
$$Q(s',\mathbf{a}) = Q_1(s',a_1,a_2) + Q_2(s',a_1,a_2) + Q_3(s',a_3)$$



Koordinálás kérdése

Indirekt koordinálás

pl. tanulva mások (stacionárius probabilisztikus) empirikus modelljét, hogy egyes cselekvéseit milyen gyakorisággal használják:

JAL – Joint Action Learner

$$\hat{\sigma}_j^i(a_j) = \frac{C_j^i(a_j)}{\sum_{\tilde{a}_j \in A_j} C_j^i(\tilde{a}_j)}$$

$C^i_j(a_j)$ – i-edik ágens hányszor látta, hogy az j-edik ágens egy a_j cselekvéséhez folyamodik, erre a legjobb válasz számítása

Frequency Maximum Q-value heurisztika: mely cselekvések jók voltak a múltban? $r_{max}(a_i)$ - az a_i–re eddig kapott max megerősítés, $C^i_{max}(a_i)$ – ennek a gyakorisága és $C^i(a_i)$ – az a_i cselekvés (önmagában) gyakorisága (az a_i -re több megerősítés is jöhetett, szórás csakis a mások cselekvései miatt, determinisztikus probléma).
A számított Q értéket a Boltzmann-felfedezés képletében használja (*v* egy súly).

$$\tilde{Q}_i(a_i) = Q_i(a_i) + v \frac{C^i_{max}(a_i)}{C^i(a_i)} r_{max}(a_i)$$

Explicit koordinálás

pl. társadalmi szabályok
normatívák, törvények
kommunikáció
szerepek

pl. ágens 1 < ágens 2
L < R < S
döntés (L_1, L_2)

Teljes versengés

Minimax Q-tanulás (1. ágens esete, $Q = Q_1 = -Q_2$)

$$Q^{új}(s_k, a_{1,k}, a_{2,k}) = Q(s_k, a_{1,k}, a_{2,k}) + \alpha\left[r + \gamma \mathbf{m}_1(Q, s_{k+1}) - Q(s_k, a_{1,k}, a_{2,k})\right]$$

$$\mathbf{m}_1(Q,s) = \max_{\pi_1(s,\cdot)} \min_{a_2} \sum_{a_1} \pi_1(s,a_1) Q(s,a_1,a_2)$$

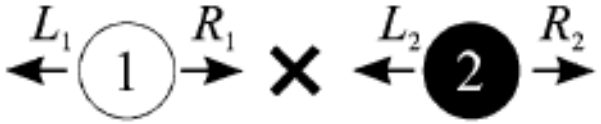
Az 1. ágens minimax haszna (Lineáris Programozással)

$$\pi_1(s_k,\cdot) = \arg \mathbf{m}_1(Q, s_k)$$

Garantált konvergencia (NE-hez) (self-play), de nem racionális.

Teljes versengés

Minimax Q-tanulás, példa



Q	L_2	R_2
L_1	0	1
R_1	-10	10

Az 1. ágens szeretne elfoglalni a keresztet és elmenekülni.
A 2. ágens szeretne elkapni az 1. ágenst.

A Q táblázat az 1. ágens perspektíváját mutatja, a 2. ágens Q függvénye ennek (-1)-szerese.

A minimax megoldás az 1. ágensre:

Ha L_1-et lép, akkor a 2. minimalizálva L_2-et lép, eredményben 0.
Ha R_1-et lép, akkor a 2. minimalizálva szintén L_2-et lép, eredményben -10.

Az 1. ágensnek tehát L_1-et kell lépnie (minimumok maximuma),
mert így legfeljebb 0-val megússza.

(L_1, L_2) egyben egy NE is.

Vegyes feladatok

Nincsenek feltételek megerősítésekre. Valamilyen egyensúly felé mégis húzni kell. Lehet pl. Nash-egyensúly, de mi van, ha több van ilyen?

Egyedi ágens Q-tanulása (a többi implicite a környezeti információban rejlik)

Ágens-független módszerek (egymástól függetlenül, de feltehetően egy közös egyensúly fel), pl. **Nash-Q-tanulás**.

$$Q_i^{új}(s_k,\mathbf{a}_k) = Q_i(s_k,\mathbf{a}_k) + \alpha\left[r_i + \gamma \mathbf{eval}_i\left\{Q_{\cdot}(s_{k+1},\cdot)\right\} - Q_i(s_k,\mathbf{a}_k)\right]$$

Q értékek által definiált kifizetési mátrix, m-edik pillanatban $Q_{\cdot,m}(s_k,\cdot)$

$$\mathbf{eval}_i\left\{Q_{\cdot}(s,\cdot)\right\} = V_i(s, NE\left\{Q_{\cdot}(s,\cdot)\right\})$$

$$\mathbf{solve}_i\left\{Q_{\cdot}(s,\cdot)\right\} = NE_i\left\{Q_{\cdot}(s,\cdot)\right\}$$

$$\pi_i(s,\cdot) = \mathbf{solve}_i\left\{Q_{\cdot}(s_k,\cdot)\right\}$$

NE az egyensúly kiszámítása, NE_i az i-ik ágens stratégiája az egyensúlyban és V_i az ágens várható haszna s-ben az egyensúlyban. Bizonyos feltételekkel NE-hoz konvergál. Mindegyik ágens számon tartja mások Q értékeit is!

Vegyes feladatok

Nash-Q egyszerübben: **FoF Friend-or-Foe tanulás**

Fel kell ismerni, hogy az ellenfél kooperál, vagy ellenünk dolgozik.

Kooperál esetében: JAL (Joint Action Learner) a megoldás, a Q közös maximumára való törekvés. Ha ellenünk van legyen a minimax-Q játék.

$$Nash_1(s, Q_1, Q_2) = \max_{a_1, a_2} Q_1(s, a_1, a_2)$$

$$Nash_1(s, Q_1, Q_2) = \max_{\pi_1(s,\cdot)} \min_{a_2} \sum_{a_1} \pi_1(s, a_1) Q_1(s, a_1, a_2)$$

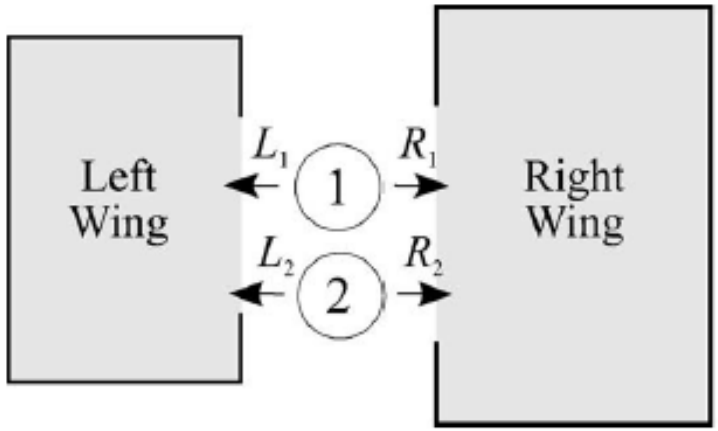
Ellenséges egyensúly (**EE**): ha mások tőle eltérnek, az ágensünk helyzete változatlan lesz, vagy javul.
Kooperatív egyensúly (**KE**): ahol az ágensek maximális haszonhoz jutnak (a jóléti megoldás).

Ha a játéknak van EE egysúlya, FOE-Q tanulás ezt megtanulja.
Ha a jatéknak van KE egyensúlya, FRIEND-Q tanulás ezt megtanulja.

Vegyes feladatok

Problémák egyensúlyi helyzetek koordinálásával itt is, pl.:

Két porszívóágens feladata a két szobából álló lakás kitakarítása.
Mindegyik jobban szeretne a bal szobát megkapni, mert ez kisebb.



Q_1	L_2	R_2
L_1	0	3
R_1	2	0

Q_2	L_2	R_2
L_1	0	2
R_1	3	0

Két Nash-egyensúly van: $(L_1, R_2), (R_1, L_2)$
de ha a két ágens között nincs koordináció (nincs megegyezés melyik NE-re játszanak, akkor mindketten ugyanabban a szobában végeznek, kisebb hasznossággal.

$$Q_1(L_1, L_2) = Q_1(R_1, R_2) = Q_2(L_1, L_2) = Q_2(R_1, R_2) = 0$$

Vegyes feladatok

Ágens-követő, ágens-tudatos módszerek (más ágensek modellezése, a modell felhasználása tanulásban: - érzékelés + stratégia-váltás)

AWESOME
(**Adapt When Everyone is Stationary, Otherwise Move to Equilibrium**)

Induláskor π_i egyensúlyi stratégiát játszik (az i-ik ágens) és mások cselekvéseit követi. Minden N-ik kör végére a megfigyelt gyakoriságokból kiszámítja az s_j-t, az j-ik ágens (feltehetően kevert) stratégiabecslését.
Ha s_j minden j-ik játékos stratégiája egyensúlyi, akkor i folytatja az egyensúlyi stratégiájának bevetését.
Különben az i-ik ágens megjátsza az s_j stratégiára kiszámított legjobb válasz stratégiáját (best response).

Ha minden játékos AWESOME játékos, akkor a közös tanulás az egyensúlyi helyzethez konvergál és nem fog tőle eltérni.

Vegyes feladatok

IGA (**Infinitensimal Gradient Ascent**) –
ágensek cselekvéseinek a valószínűsége az,
amit az ágensek tanulnak (2 ágens, 2-2 cselekvés),

α az 1. ágens 1. cselekvésének a valószínűsége és
β a 2. ágens 1. cselekvésének a valószínűsége:

$$\alpha_{k+1} = \alpha_k + \delta_{1,k} \frac{\partial E\{r_1 \mid \alpha, \beta\}}{\partial \alpha}$$

$$\beta_{k+1} = \beta_k + \delta_{2,k} \frac{\partial E\{r_2 \mid \alpha, \beta\}}{\partial \beta}$$

$$\delta_{1,k} = \delta_{2,k} = \delta$$

Vegyes feladatok

WoLF-IGA (**Win-or-Learn-Fast**)

- győztes helyzetben ágens óvatos, kis δ-val lassan tanul, nehogy az előnyös pozícióját elveszítse,
- vesztes esetben nagyobb δ-val gyorsan igyekszik kikerülni a jelen helyzetből.

$$\alpha_{k+1} = \alpha_k + \delta_{1,k} \frac{\partial E\{r_1 \mid \alpha, \beta\}}{\partial \alpha}$$

$$\beta_{k+1} = \beta_k + \delta_{2,k} \frac{\partial E\{r_2 \mid \alpha, \beta\}}{\partial \beta}$$

$$\delta_{1,k}, \delta_{2,k} \in [\delta_{\min}, \delta_{\max}] > 0$$

Hogyan látszik meg, melyik a győzelmi, melyik a vesztes állapot?

WoLF-elv:

Ha a várható hasznom, ahogy játszom és ahogy az ellenfél játszik, jobb, mintha a jelen játéka mellett én az egyensúlyi stratégiát játszanám, akkor győzelemre állok.

Ha rosszabbul állok az egyensúlyi stratégiámhoz képest (az ellenfél adott játéka mellett), akkor vesztésre állok.

Lucian Busoniu, Robert Babuska, and Bart De Schutter, A Comprehensive Survey of Multiagent Reinforcement Learning, IEEE Trans. on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 38, No. 2, March 2008

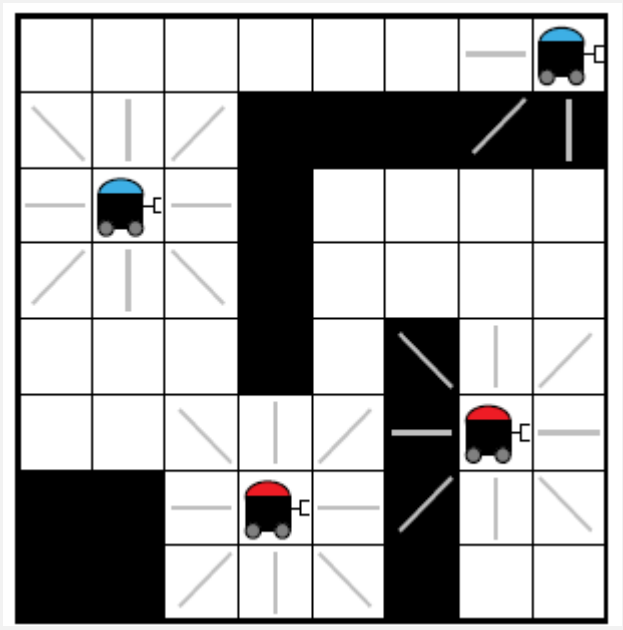
Többágenses mély megerősítéses tanulás

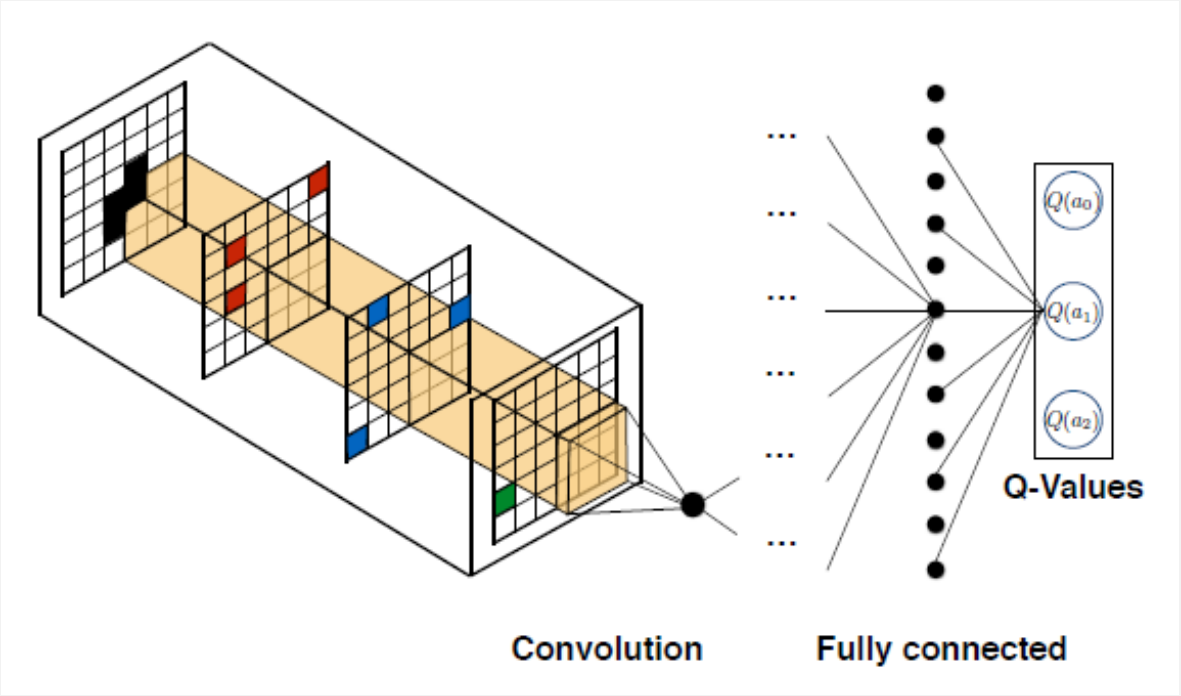
$$Q(s,\mathbf{a}) = Q(s,\mathbf{a}) + \alpha\left[r + \gamma \max_{a'} Q(s',\mathbf{a}') - Q(s,\mathbf{a})\right], \quad Q(s,\mathbf{a}|\boldsymbol{\theta})$$

$$\min L(s,\mathbf{a}|\boldsymbol{\theta}_i) = \left(r + \gamma \max_{a} Q(s',\mathbf{a}|\boldsymbol{\theta}_i) - Q(s,\mathbf{a}|\boldsymbol{\theta}_i)\right)^2$$

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_i)$$

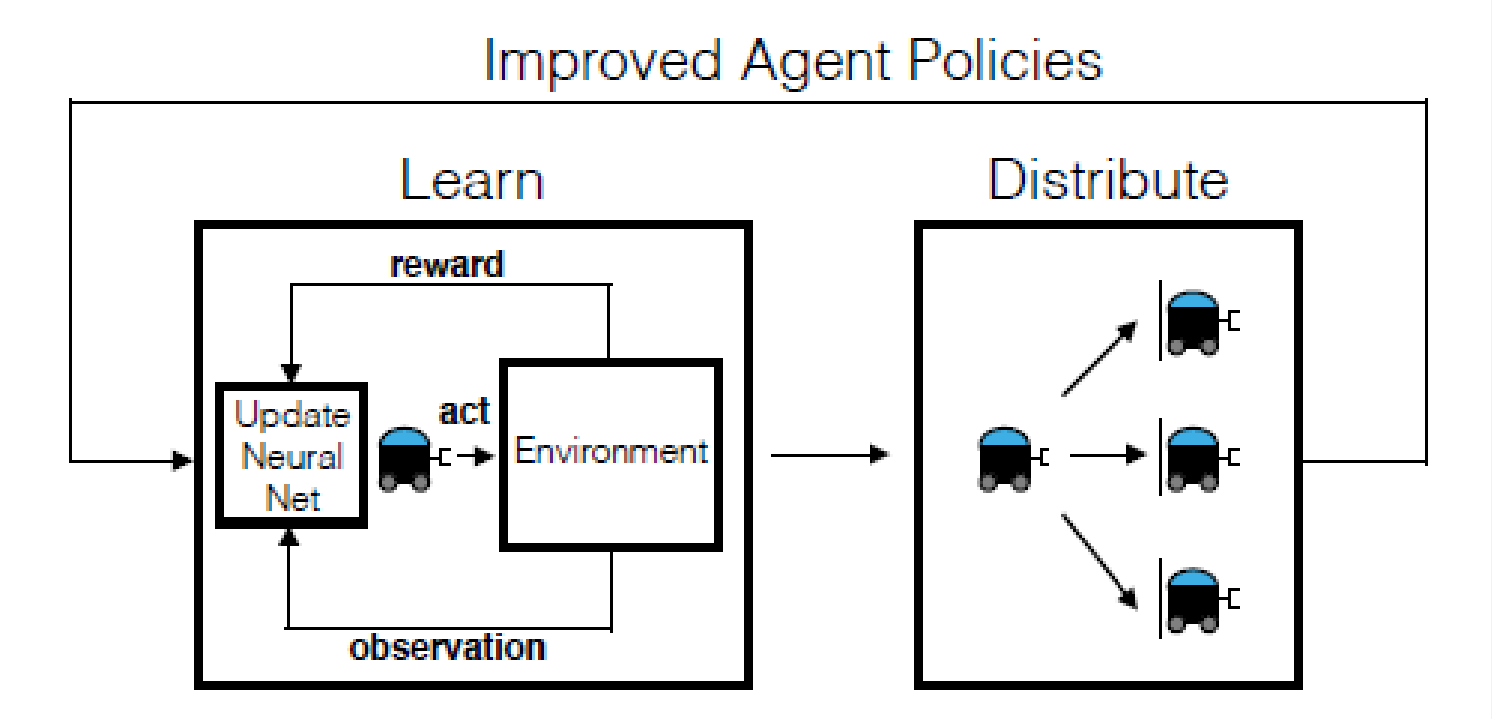
Környezet 2 dim, 2 ágens tipus, diszkrét tér/idő.
Konvolúciós NN csatornák: háttér cs. (akadályok), ellenfelek cs., szövetségesek cs., saját magam cs. : 4 x H x H.



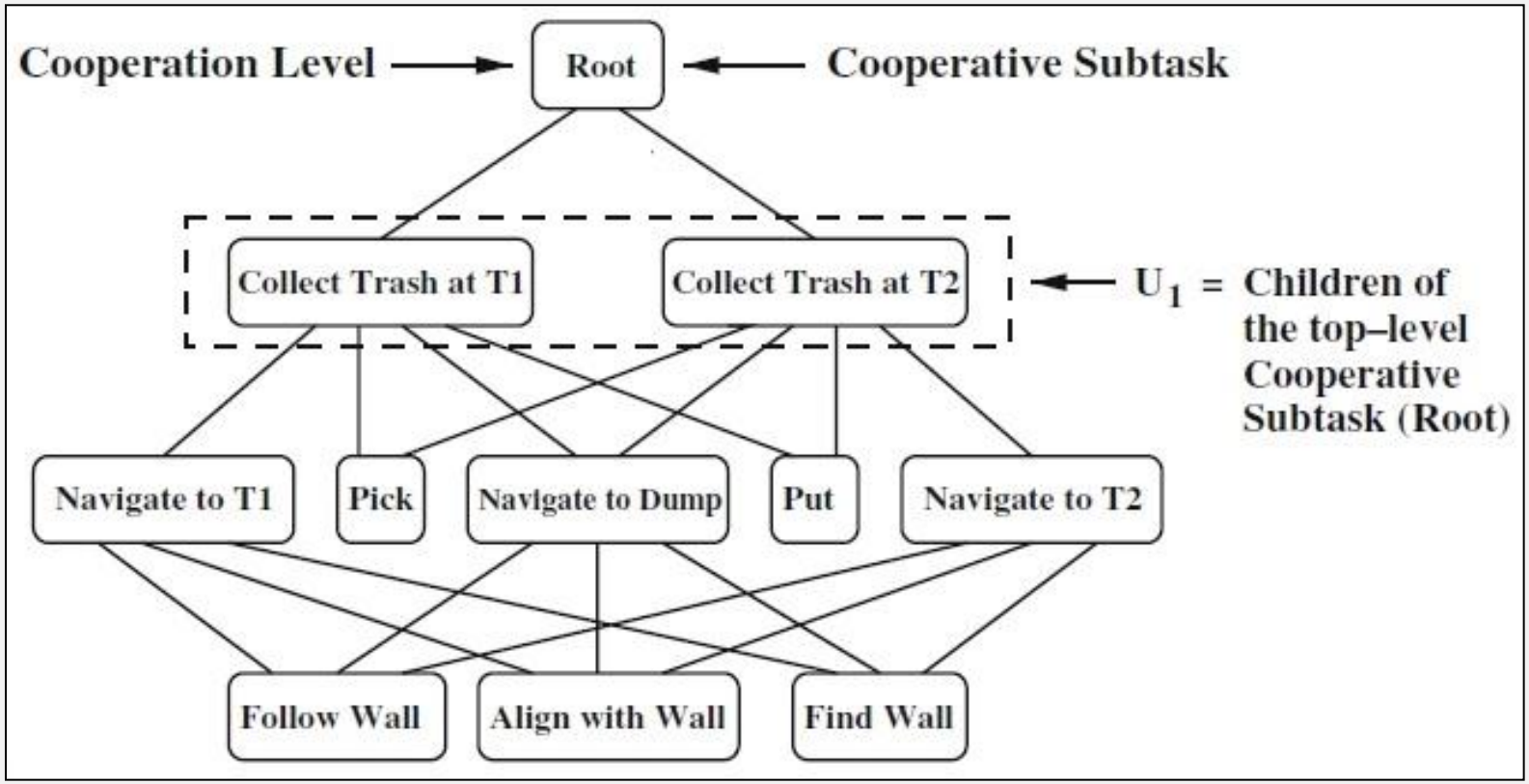


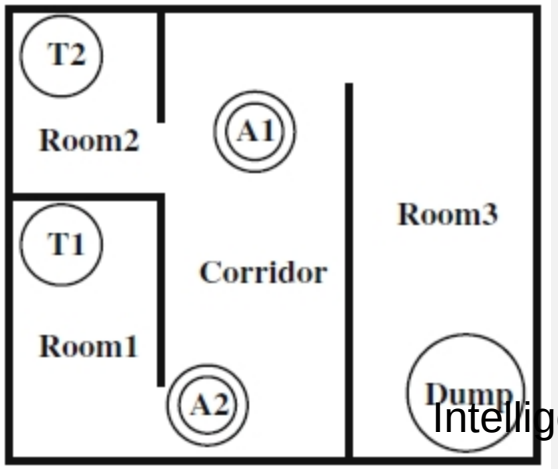
Többágenses mély megerősítéses tanulás

Tanulás: egyszerre 1 ágens, a többinek stratégiája lefixált,
a megtanult stratégia kiosztása a saját tipusú ágensekre.

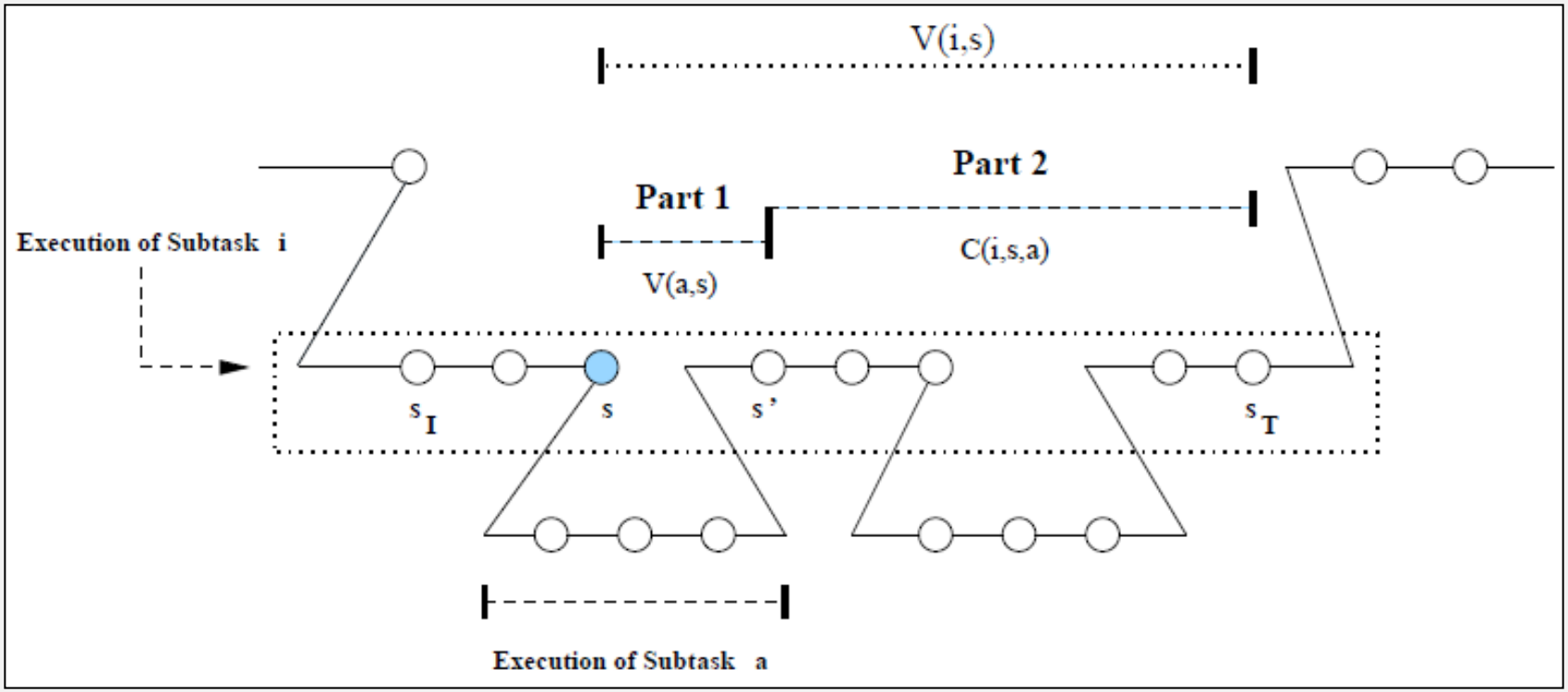


Hierarchikus többágenses megerősítéses tanulás





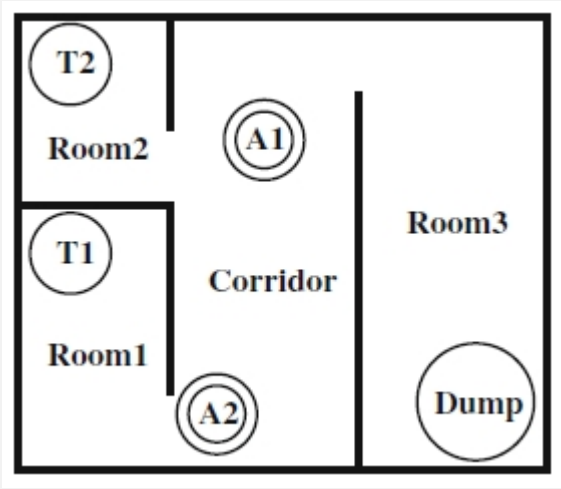
Hierarchikus többágenses megerősítéses tanulás



$$Q^{\pi}(i,s,a) = V^{\pi}(a,s) + C^{\pi}(i,s,a)$$

$$V^{\pi}(i,s) = \begin{cases} Q^{\pi}(i,s,\pi_i(s)) \\ \sum_{s' \in S_i} P(s'|s,i) R(s'|s,i) \end{cases}$$

Hierarchikus többágenses megerősítéses tanulás



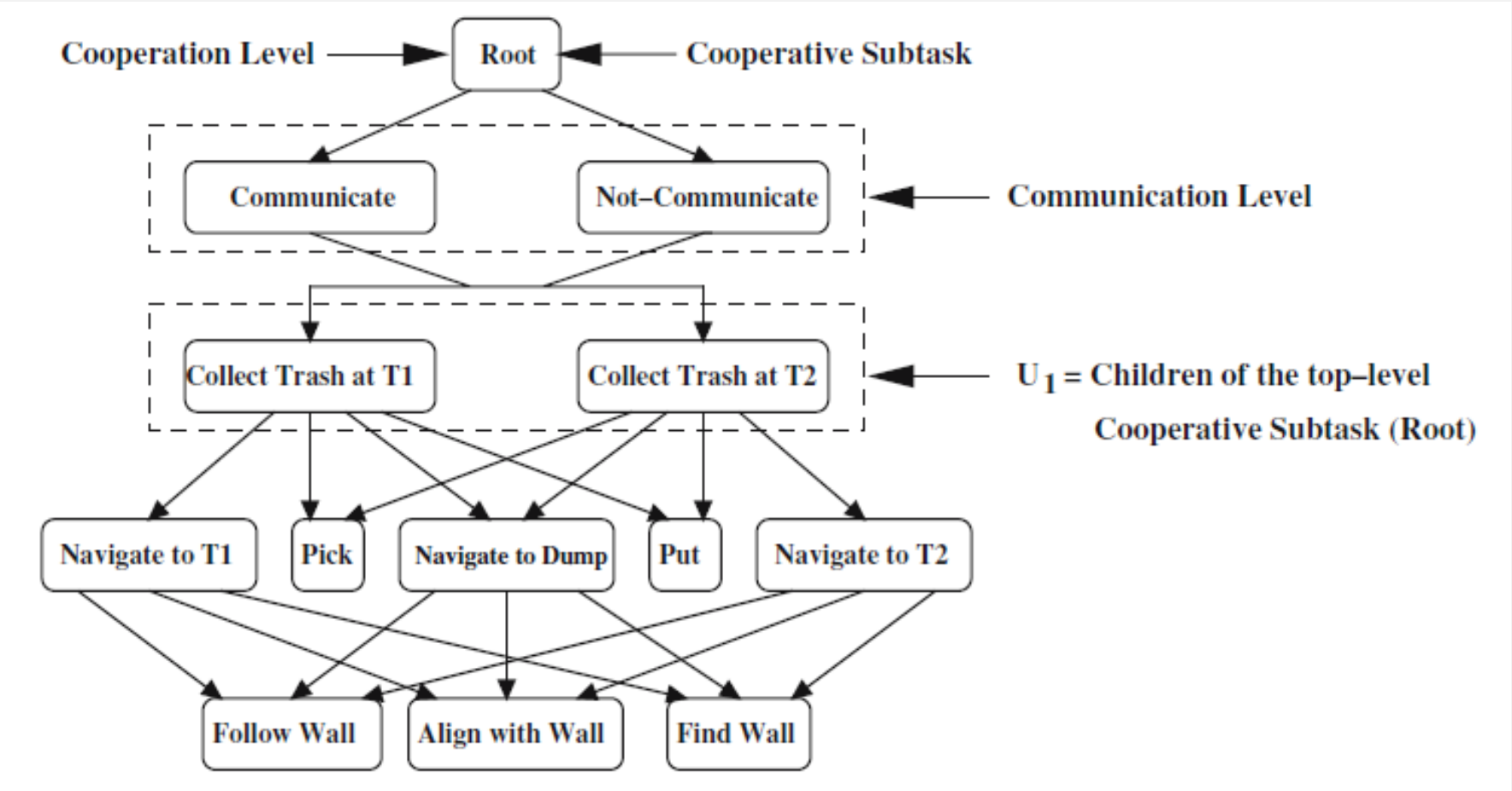
Learned Policy for Agent 1

```
root
  navigate to T1
     go to location of T1 in room 1
  pick trash from T1
  navigate to Dump
     exit room 1
     enter room 3
     go to location of Dump in room 3
  put trash collected from T1 in Dump
end
```

Learned Policy for Agent 2

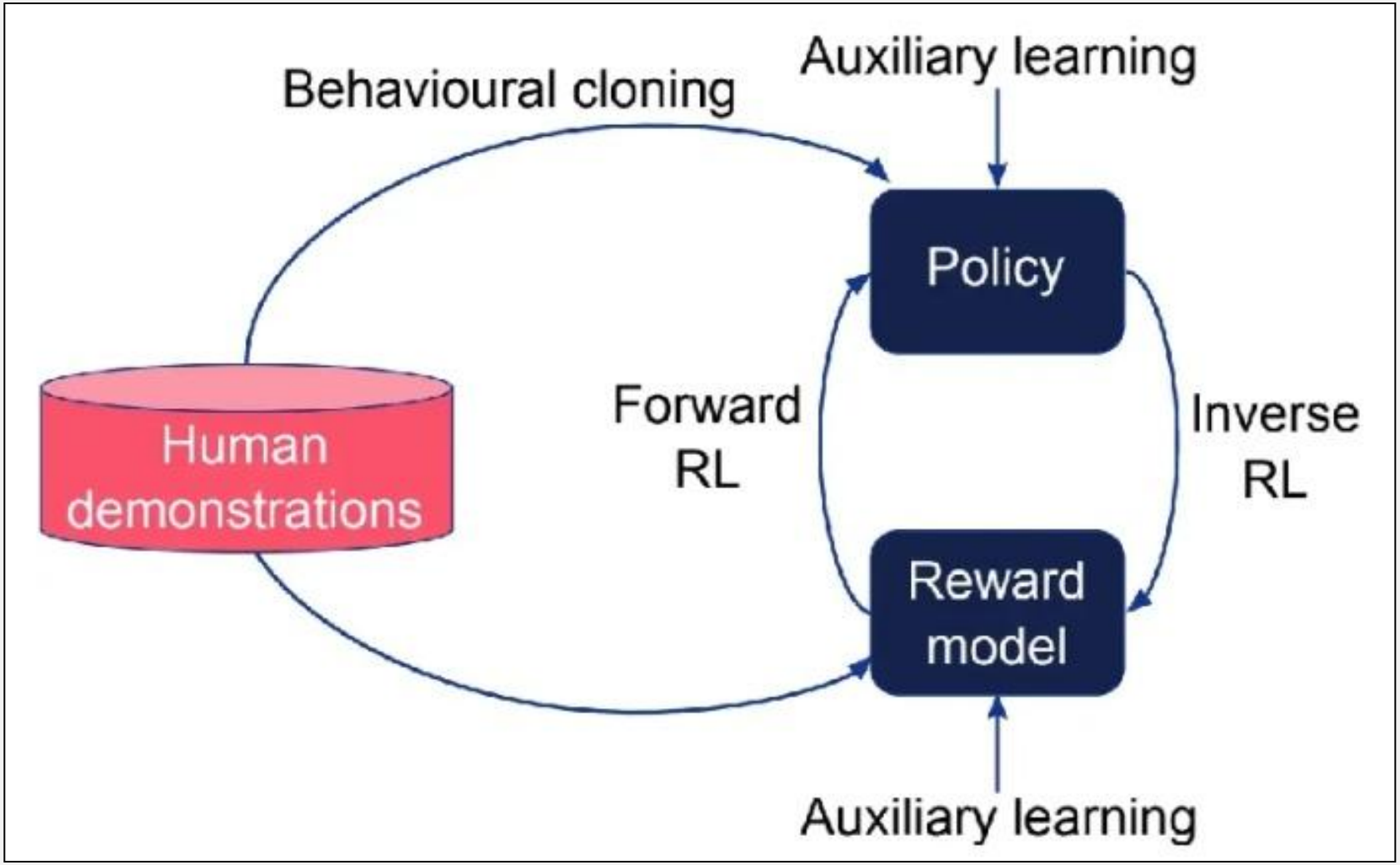
```
root
  navigate to T2
     go to location of T2 in room 2
  pick trash from T2
  navigate to Dump
     exit room 2
     enter room 3
     go to location of Dump in room 3
  put trash collected from T2 in Dump
end
```

Hierarchikus többágenses megerősítéses tanulás



Taszkgráf kommunikációs cselekvésekkel

Inverz megerősítéses tanulás



Inverz megerősítéses tanulás

Ha adott a π^*, visszaállítható-e az R?
Ha adottak a végrehajtási pályák, visszaállítható-e az R?

Bemenet:
- Állapottér
- Cselekvéshalmaz
- Állapotátmeneti modell
- **R megerősítés nincs!**
- Tanári bemutató van!
  - s0,a0,s1,a1,s2,a2,…
  - (a tanári π^* eljárásmód nyoma)

Kimenet: R?

Tudományosan érdekes
Inas tanulás/ Imitációs tanulás
Kooperatív és versengő ágensek modellezése
- Járművezetés, városi navigálás
- Légi navigálás
- Parkolás
- Emberi útvonaltervezés
- **Emberi célok azonosítása**
- Négylábú robotok mozgása

…

Azt az **R*** megerősítésfüggvényt találjuk meg, ami a **szakértő viselkedését megmagyarázza**!

Inverz megerősítéses tanulás

$$\mathrm{E}[\textstyle\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi^*] \geq \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi] \quad \forall \pi$$

(Számítási problémák)

Jellegvektor alapú megerősítés (regressiós számítás)

$$R(s) = w^\top \phi(s)$$

$$\mathrm{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi] = \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t w^\top \phi(s_t)|\pi] = w^\top \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)|\pi]$$

$$= \quad w^\top \mu(\pi)$$

Várható kummulált diszkont jellegértékek

$$\mathrm{E}[\textstyle\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi^*] \geq \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi] \quad \forall \pi$$

w ! → $w^{*\top} \mu(\pi^*) \geq w^{*\top} \mu(\pi) \quad \forall \pi$

Inverz megerősítéses tanulás

Jellegvektor alapú megerősítés (regressiós számítás)

Egyértelmű megoldás érdekében

$$w^{*\top}\mu(\pi^*) \geq w^{*\top}\mu(\pi) \quad \forall\pi$$

Standard max tartalék

$$\min_w \|w\|_2^2$$
$$\text{s.t.} \quad w^\top\mu(\pi^*) \geq w^\top\mu(\pi) + 1 \quad \forall\pi$$

„Struktúrált predikciós” max tartalék

$$\min_w \|w\|_2^2$$
$$\text{s.t.} \quad w^\top\mu(\pi^*) \geq w^\top\mu(\pi) + m(\pi^*,\pi) \quad \forall\pi$$

„Struktúrált predikciós” max tartalék gyengítő változókkal, stb.

$$\min_{w,\xi} \|w\|_2^2 + C\xi$$
$$\text{s.t.} \quad w^\top\mu(\pi^*) \geq w^\top\mu(\pi) + m(\pi^*,\pi) - \xi \quad \forall\pi$$

Többágenses megerősítéses tanulás - Szimulációk

http://busoniu.net/files/repository/2010-08-04_marl-1.3.zip