

# Üzleti Intelligencia

## 3. Előadás: Markov döntési folyamatok megoldása

Kuknyó Dániel  
Budapesti Gazdasági Egyetem

2023/24  
1.félév

## 1 Bevezetés

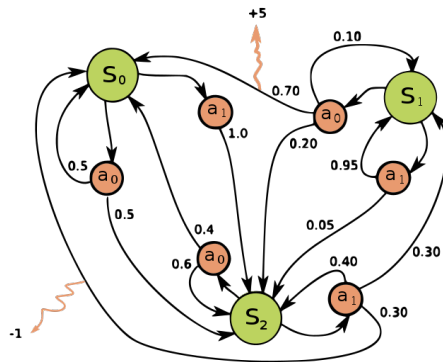
## 1 Bevezetés

# Az RL modellje

## Markov döntési folyamat

$$(S, A, P, R, s_0, \gamma)$$

- $S$ : állapotok halmaza
- $A$ : cselekvések halmaza
- $P : S \times A \times S \rightarrow [0, 1]$ :  
állapotátmeneti valószínűségek
- $R : S \times A \rightarrow \mathbb{R}$ : azonnali jutalmak halmaza
- $s_0$ : kezdőállapot
- $\gamma$ : diszkont faktor



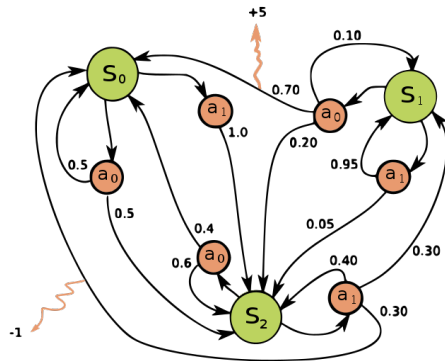
# Az RL modellje

Az MDP folyamata:

- 1 Az ügynök  $s_0$  állapotból indul
- 2 Az ügynök  $\pi$  politika szerint cselekszik:  
 $a_t \sim \pi(s_t)$
- 3 A környezet reagál a cselekvésre, és visszaadja az ügynöknek  $r_{t+1}$  jutalmat és  $s_{t+1}$  következő állapotot
- 4 Ez ismétlődik amíg a kilépési kritérium be nem teljesül

Cél: Az optimális politika megtalálása. A politika optimális, ha a hozamának várható értéke maximális:

$$E_{\pi} (r_1 + \gamma r_2 + \gamma^2 r_3 + \dots) \rightarrow \max$$



# A mohó ügynök

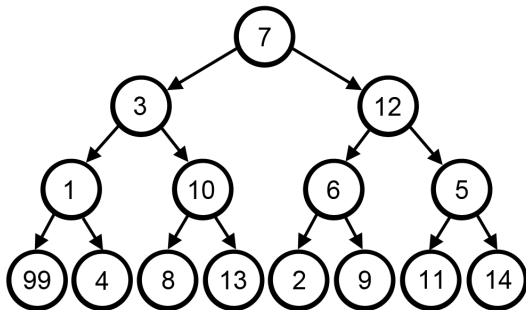
A legegyszerűbb cselekvés kiválasztási szabály, ha az ügynök mindig azt a cselekvést választja, ami számára a lehető legnagyobb várható hozammal rendelkezik.

## Mohó cselekvés választás

Mohó politika mindig azt a cselekvést fogja választani, amelyik - egy lépéses távlatban - a lehető legnagyobb várható jutalommal fog járni az ügynök számára  $v_\pi$  szerint.

$$A_t = \underset{a}{\operatorname{argmax}} Q_t(a)$$

- Mi lenne a mohó politika ebben az esetben?
- Mindig ez a legjobb megoldás?
- A legjobb megoldás mindig mohó?



# A mohó ügynök

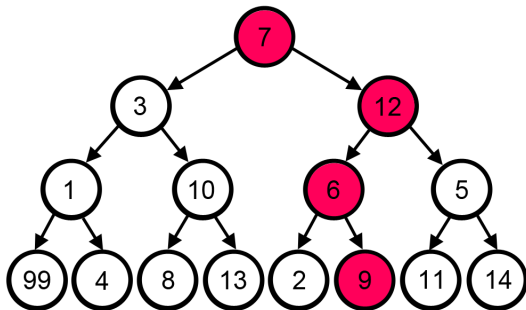
A legegyszerűbb cselekvés kiválasztási szabály, ha az ügynök mindig azt a cselekvést választja, ami számára a lehető legnagyobb várható hozammal rendelkezik.

## Mohó cselekvés választás

Mohó politika mindig azt a cselekvést fogja választani, amelyik - egy lépéses távlatban - a lehető legnagyobb várható jutalommal fog járni az ügynök számára  $v_\pi$  szerint.

$$A_t = \underset{a}{\operatorname{argmax}} Q_t(a)$$

- Mi lenne a mohó politika ebben az esetben?
- Mindig ez a legjobb megoldás?
- A legjobb megoldás mindig mohó?



# Az $\varepsilon$ -mohó stratégia

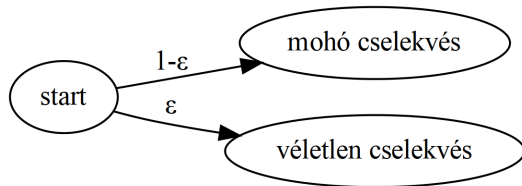
Egy másik lehetőség, ha adott valószínűséggel az ügynök véletlen cselekvést hajt végre remélve, hogy ezzel elér egy olyan állapotba amelyhez nagy jutalom tartozik. A véletlen cselekvés a **felfedezés**, és végrehajtásának valószínűsége  $\varepsilon$ .

## $\varepsilon$ -mohó cselekvés választás

$$A_t \leftarrow \begin{cases} \underset{a \sim A}{\operatorname{argmax}} Q(a) & P=1-\varepsilon \\ a \sim A & P=\varepsilon \end{cases}$$

Az ügynök tehát  $\varepsilon$  valószínűséggel véletlen cselekvést választ az ismeretlen, de nagyobb jutalom reményében. Ez a **felfedezés** művelete.

$\varepsilon$  valószínűséggel pedig a már ismert és a legnagyobb várható jutalommal járó cselekvést hajtja végre. Ez a **kizsákmányolás** művelete.





# Példák

A következő valós példák alkalmasak a felfedezés/kizsákmányolás dilemma bemutatására:

- Étterem választás:
  - **Kizsákmányolás:** elmész a kedvenc éttermedbe.
  - **Felfedezés:** elmész egy új étterembe, hátha találsz egy jobbat mint a kedvenced.
- Online hirdetés:
  - **Kizsákmányolás:** a legjobb reklám megmutatása a felhasználónak.
  - **Felfedezés:** egy új reklám megmutatása a felhasználónak, hátha tetszik neki.
- Olajfúrás:
  - **Kizsákmányolás:** Egy meglévő helyen fúrás az olajért.
  - **Felfedezés:** Egy új helyen fúrás.
- Klinikai kezelés:
  - **Kizsákmányolás:** A bevált kezelés alkalmazása.
  - **Felfedezés:** Új kezelés kipróbálása.

## 1 Bevezetés

# A rabló probléma

A  $k$ -karú rabló problémája egy elméleti megerősítéses tanulás probléma. A játékos egy rablógépen játszik, amelynek  $k$  karja van.

Minden karhúzás után egy állandó eloszlásból választott jutalmat kap az ügynök. Az ügynök célja, hogy olyan politikát válasszon, ami az elvárt hozamot maximalizálja 1000 cselekvés vagy időlépés után.



# A rabló probléma

Az ügynöknek számon kell tartania, mennyi a jutalom várható értéke, ha adott egy  $a$  cselekvés. Ez a  $Q(s, a)$  állapot-cselekvés minőség függvény. A rabló problémában csak egy állapot van, ezért elég csak a cselekvésekhez tartozóan számon tartani:

$$q_*(a) = E[r_t | A_t = a]$$

A jutalom várható értéke:

$$Q_n = \frac{r_1 + r_2 + \dots + r_{n-1}}{n - 1}$$



# Algoritmus a rabló megoldására

---

**Algorithm 1:** Egyszerű rabló algoritmus

---

```
 $Q(a) \leftarrow 0$  for  $a = 1 \rightarrow k$ ;           /* Cselekvés minőségének függvénye */  
 $N(a) \leftarrow 0$  for  $a = 1 \rightarrow k$ ;       /* Kar meghúzásainak a száma */  
for  $t = 1 \rightarrow \max_t$  do  
     $p = \text{random}(0, 1)$ ;                /* Véletlen szám 0 és 1 között */  
    if  $p > \varepsilon$  then  
         $a \leftarrow \underset{a}{\operatorname{argmax}} Q(a)$ ;    /* Legnagyobb ismert jutalom cselekvése */  
    else  
         $a \leftarrow a \sim A$ ;                /* Véletlen cselekvés */  
    end  
     $r \leftarrow \text{env}(a)$ ;                /* Cselekvés végrehajtása a környezetben */  
     $N(a) \leftarrow N(a) + 1$ ;            /* Cselekvés számlálójának növelése */  
     $Q(a) \leftarrow Q(a) + \frac{1}{N(a)} [r - Q(a)]$ ;    /*  $Q$ -érték frissítése */  
end
```

---