Úkol 7: Chceme použít MDP, chceme vytvořit policy.

1) V init se jednou vypočítá policy, pak už se znovu nepočítá.
2) Chceš to počítat přes $U(s)$ „utility" a optimální policy.

Policy iteration je těžší implementovat, ale počítá rychleji.

C > čas je důležitý v ReCodEx

Value iteration mi na čas už v některých nepujde.

Ve skutečnosti stačí implementovat precompute_probability_policy_trivial

$$\pi(s) = \underset{a \in A}{\arg\max} \left[P(s'/s,a) \cdot U(s') \right]$$

$$U(s) = R(s) + \max_{a \in A} P(s'/s,a) \cdot U(s')$$

V prezentaci!

Value iteration

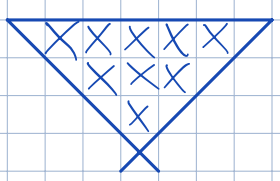
inicializace $U(s) = R(s)$

Bellman update: → discount factor

$$U_{i+1}(s) = R(s) + \gamma \max_{a \in A} P(s'/s,a) \cdot U_i(s')$$

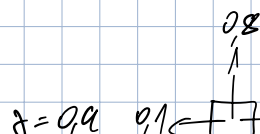
Co vlastně chci počítat? → úplně nechci zapomínat vedlejší stavy. Tedy $\gamma = 1$

$$R(s) = M[s] + \gamma \min_{a \in \{L,R,V,D\}} P(s'/s,a) \cdot U(s')$$



Rewards

0 0 0 +1
0 ■ 0 -1
0 0 0 0



$$U_1(0,2) = 0 + 0{,}9 \cdot \max\left(\leftarrow 0,\ \uparrow 0{,}1 \cdot 1,\ \rightarrow 0{,}8 \cdot 1,\ \downarrow 0{,}1 \cdot 1 \right) = 0{,}9 \cdot 0{,}8$$

$$U_2(0,2) = 0 + 0{,}9 \cdot \max\left(0,\ 0{,}1 \cdot 0{,}72,\ 0{,}8 \cdot 0{,}72,\ 0{,}1 \cdot 0{,}72 \right) = \ldots$$

$$U_2(1,2) = 0 + 0{,}9 \cdot \max\left(0{,}1 \cdot 0{,}72,\ 0{,}8 \cdot 0{,}72 + 0{,}1 \cdot (-1),\ \ldots\right)$$

Jdu přes všechny možný dohodnocení

Jak získat správnou akci? $\pi(s) = \underset{a \in A}{\arg\max} \left[\sum_{s'} P(s'/s,a) \cdot U(s') \right]$

Policy iteration

Inicializuju náhodnou policy $\pi_0(s)$

do

evaluate policy

$$U(s) = R(s) + \gamma \cdot \sum_{s'} P(s'/s, \pi(s)) \cdot U(s)$$

if $\max_{a \in A} \sum_{s'} P(s'/s,a) \cdot U(s) > \sum_{s'} P(s'/s, \pi(a)) \cdot U(s)$

pokud existuje lepší akce než v policy, update policy