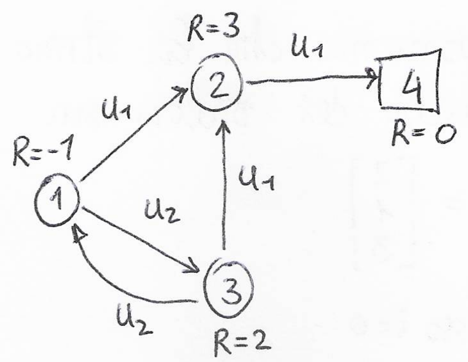
## POLICY IMPROVEMENT

$\gamma = 0.5$




$\pi_0 = \begin{bmatrix} u_1 \\ u_1 \\ u_1 \\ - \end{bmatrix}$ , $X = \{1, 2, 3, 4\}$

$n = 4$

passo 1: calcolare il valore di $\pi_0$: $\quad (v_\pi(x) = R_{k+1} + \gamma v_\pi(x'), \ \forall x \in X)$

$$\begin{cases} v_{\pi_0}(1) = -1 + \frac{1}{2} v_{\pi_0}(2) \\ v_{\pi_0}(2) = 3 + \frac{1}{2} v_{\pi_0}(4) \\ v_{\pi_0}(3) = 2 + \frac{1}{2} v_{\pi_0}(2) \\ v_{\pi_0}(4) = 0 \end{cases} \Rightarrow$$

da $v_{\pi_0}(4) = 0$, abbiamo $v_{\pi_0}(2) = 3$

dunque $v_{\pi_0}(1) = -1 + \frac{1}{2} \cdot 3 = 0.5$

$v_{\pi_0}(3) = 2 + \frac{1}{2} \cdot 3 = \frac{7}{2} = 3.5$

$$v_{\pi_0} = \begin{bmatrix} 0.5 \\ 3 \\ 3.5 \\ 0 \end{bmatrix}$$

passo 2: eseguire policy improvement, calcolando la funzione $q_{\pi_0}(x,u)$

$$q_{\pi_0}(x,u) = R_{k+1} + \gamma \underbrace{v_{\pi_0}(x')}_{\text{calcolato in precedenza}}$$

x', funzione della scelta di u

in questo caso:

$q_{\pi_0}(x, u_1) = v_{\pi_0}(x)$ (perché $\pi_0$ sceglie sempre $u_1$)

calcolare dunque $q_{\pi_0}(x, u_2)$

$q_{\pi_0}(1, u_2) = -1 + \frac{1}{2} v_{\pi_0}(3) = -1 + 1.75 = 0.75$

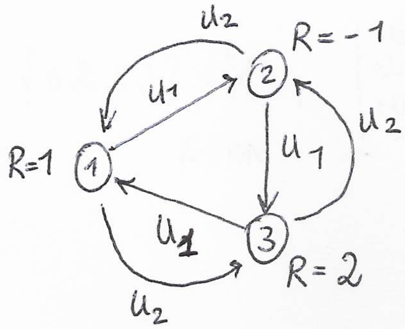
$q_{\pi_0}(2, u_2) = *$ (non si può scegliere $u_2$ nello stato 2)

$q_{\pi_0}(3, u_2) = 2 + \gamma v_{\pi_0}(1) = 2 + 0.25 = 2.25$

$q_{\pi_0}(4, u_2) = *$

quindi $\Rightarrow$ $q_{\pi_0}(x,u) = \begin{array}{c} \begin{array}{cc} u_1 & u_2 \end{array} \\ \begin{bmatrix} 0.5 & 0.75 \\ 3 & * \\ 3.5 & 2.25 \\ * & * \end{bmatrix} \end{array}$ $\Rightarrow q_{\pi_0}(1, u_2) > v_{\pi_0}(1) \Rightarrow \pi_1 = \begin{bmatrix} u_2 \\ u_1 \\ u_1 \\ * \end{bmatrix}$

# VALUE ITERATION




$\gamma = 1$

supponiamo che la stima iniziale dei valori sia

$$v_0 = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

stima al passo $i=0$

eseguire un passo di "value iteration" $v_{i+1}(x) = \max_u \{R_{k+1} + \gamma v_i(x')\} \quad \forall x$

$$v_1(1) = \max_u \{\underset{R}{1} + \underset{\gamma}{1} \cdot v_0(x')\} = \max\{1+1,\ 1+3\} = 4$$

stima aggiornata al passo $i=1$

con $u_1$, lo stato successivo è $x'=2$ e $v_0(2)=1$

con $u_2$, $x'=3$ e $v_0(3)=3$

$$v_1(2) = \max\{-1+3,\ -1+2\} = 2$$

$$v_1(3) = \max\{2+2,\ 2+1\} = 4$$

Quindi

$$v_1 = \begin{bmatrix} 4 \\ 2 \\ 4 \end{bmatrix}$$