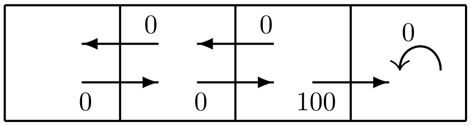
The world

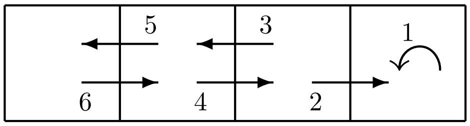
A discount factor: $\gamma = 1/2$. The learning rate: $\alpha = 0.1$.

A	B	C	D

$r(s, a) =$



initial $Q(s, a) =$



$$R = (0, 0, 100, 0, 0, 0)^T$$
$$Q_0 = (6, 4, 2, 1, 3, 5)^T$$
$$P_1 = (5, 3, 1, 1, 5, 6)^T$$

$$Q_1 = Q_0 + 0.1 \cdot (R + 0.5 \cdot P_1 - Q_0) = \begin{pmatrix} 5.65 \\ 3.75 \\ 11.85 \\ 0.95 \\ 2.95 \\ 4.8 \end{pmatrix}$$

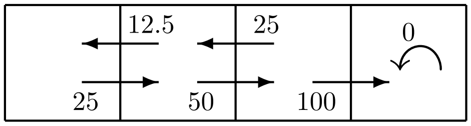
$$P_2 = (4.8, 11.85, 0.95, 0.95, 4.8, 5.65)^T$$

$$Q_2 = Q_1 + 0.1 \cdot (R + 0.5 \cdot P_2 - Q_1) = \begin{pmatrix} 5.325 \\ 3.9675 \\ 20.7125 \\ 0.9025 \\ 2.895 \\ 4.6025 \end{pmatrix}$$

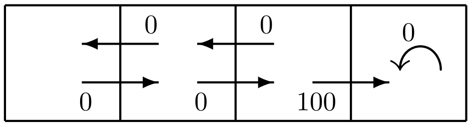
$$P_3 = (4.6025, 20.7175, 0.9025, 0.9025, 4.6025, 5.325)^T$$

$$Q_3 = Q_2 + 0.1 \cdot (R + 0.5 \cdot P_3 - Q_2) = \begin{pmatrix} 5.02263 \\ 4.60663 \\ 28.6864 \\ 0.857375 \\ 2.83563 \\ 4.4085 \end{pmatrix}$$

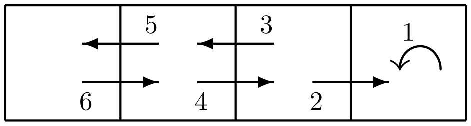
$\overrightarrow{Q}_{\text{true}} =$



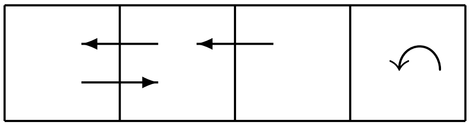
$\overrightarrow{R} =$



$\overrightarrow{Q}_0 =$

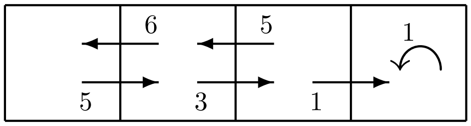


$\hat{\pi} =$

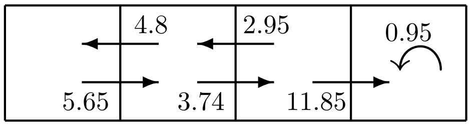


(not part of the algorithm)

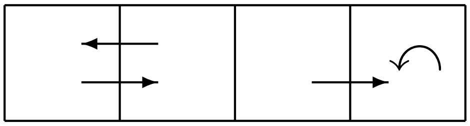
$\overrightarrow{P}_1 =$



$\overrightarrow{Q}_1 =$

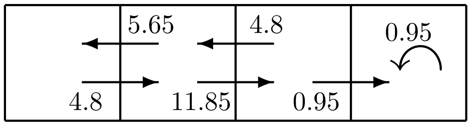


$\hat{\pi} =$

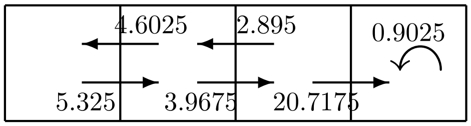


(not part of the algorithm)

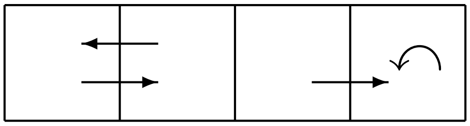
$\overrightarrow{P}_2 =$



$\overrightarrow{Q}_2 =$

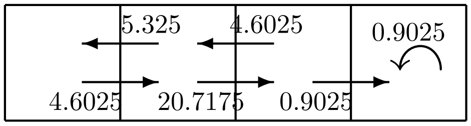


$\hat{\pi} =$

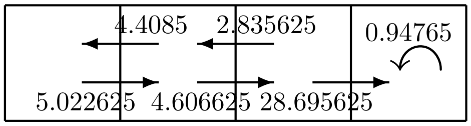


(not part of the algorithm)

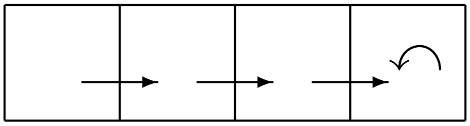
$\overrightarrow{P}_3 =$



$\overrightarrow{Q}_3 =$



$\hat{\pi} =$



(not part of the algorithm)