PASSIVE REINFORCEMENT LEARNING

For this simulation, discount factor \( \gamma = 1 \) and the reward for non-terminal transitions is \(-0.04\).

Let’s do passive reinforcement learning, using this formula:

\[
U^\pi(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right]
\]

Carry out the series of state transitions and back-propagate the resulting cell values.

\[
(1, 1) \rightarrow_{-0.04} (1, 2) \rightarrow_{-0.04} (1, 3) \rightarrow_{-0.04} (1, 2) \rightarrow_{-0.04} (1, 3) \rightarrow_{-0.04} (2, 3) \rightarrow_{-0.04} (3, 3) \rightarrow_{-0.04} (4, 3) +1
\]

\[
(1, 1) \rightarrow_{-0.04} (1, 2) \rightarrow_{-0.04} (1, 3) \rightarrow_{-0.04} (2, 3) \rightarrow_{-0.04} (3, 3) \rightarrow_{-0.04} (3, 2) \rightarrow_{-0.04} (3, 3) \rightarrow_{-0.04} (4, 3) +1
\]
Based on these three trials, average the values discovered for each state and fill in below:

The policy used and the actual utilities of the states are: