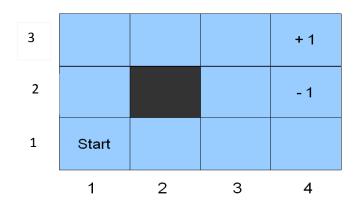
Tutorial 6 (week 7)

Temporal difference learning

Given the following situation:



Task: Use the temporal difference learning algorithm to <u>compute the utilities</u> when using the following trial:

Trial n=1: (1,1) -> (1,2) -> (1,3) -> (1,3) -> (2,3) -> (3,3) -> (4,3)

Rewards: R(s) = -0.04 for all states s which are not a goal state.

Learning rate: $\alpha(n)=60/(59+n)$, where n is a counter which counts trials (here we have the first trial. Hence: n=1).

Discount factor: Assume that $\gamma = 1$

The temporal difference algorithm is given as follows:

```
function Passive-TD-Agent (percept) returns an action inputs: percept, a percept indicating the current state s' and reward signal r' variable: \pi, a fixed policy U, a table of utilities, initially empty Ns, a table of frequencies for states, initially zero s, a, r, the previous state, action, and reward, initially null if s' is new then U[s'] \leftarrow r' if s is not null then increment Ns[s] U[s] \leftarrow U[s] + \alpha(Ns[s]) (r + \gamma U [s'] - U [s]) if Terminal[s'] then s, a, r \leftarrow null else s, a, r \leftarrow s', \pi[s'], r' return a end
```

Homework:

- Repeat the algorithm for 2 more iterations by using trial 1.
- Continue the above exercise by using a second trial: (1,1) -> (1,2) -> (1,3) -> (2,3) -> (2,3) -> (3,3) -> (4,3)
- Continue the above exercise by using a third trial:

$$(1,1) \rightarrow (2,1) \rightarrow (3,1) \rightarrow (3,2) \rightarrow (4,2)$$