Workshop 9

Offline Planning

So far we have seen value iteration and temporal difference learning (e.g. Q-Learning or SARSA). These work well in some situations, however

- **Value iteration** learns a policy, but it requires a model; and also learns a policy for every state → lots of states, X
- **TD learning** learns a policy, but really only works well for repetitive tasks OR requires extensive training (e.g. up to weeks) → cannot react to things that were not explored a lot!

They are both offline methods: policies must be trained in advance

Online Planning

We are still trying to solve the problem to get a solution. However:

- We will get only one action at a time , execute the action,
- And then we generate a new problem
- And solve it , and so on

Solve problem while during calculation

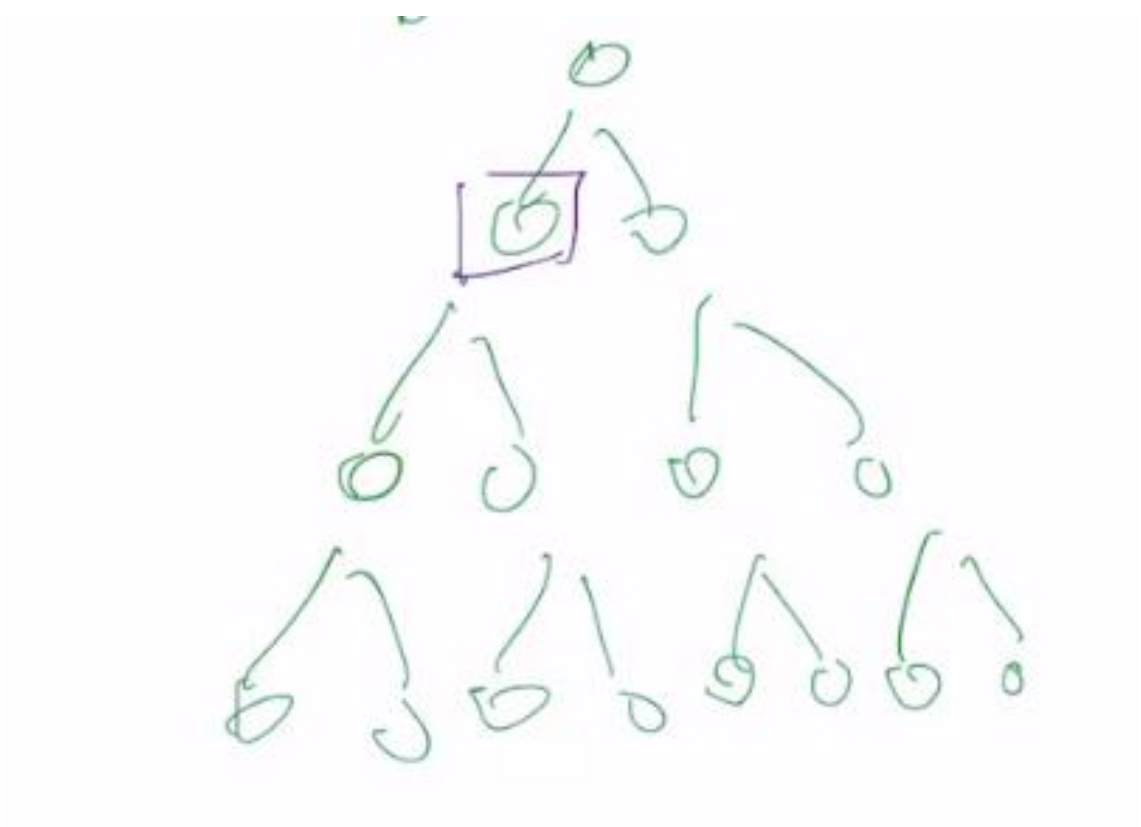
Need to do calculation for each step

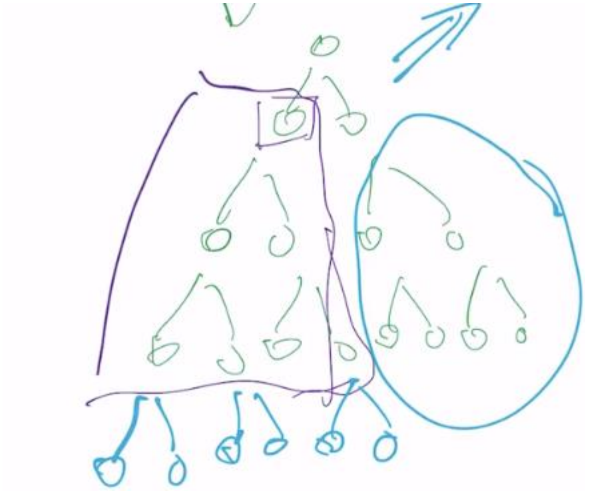
Can we apply online planning approach on offline problem?

MCTS framework

- ***Select***: Select a single node in the tree that is *not fully expanded*. By this, we mean at least one of its children is not yet explored.
- ***Expand***: Expand this node by applying one available action (as defined by the MDP) from the node.
- ***Simulation***: From one of the outcomes of the expanded, perform a complete random simulation (of the MDP) to a terminating state. This therefore assumes that the simulation is finite, but versions of MCTS exist in which we just execute for some time and then estimate the outcome.
- ***Backpropagate***: Finally, the value of the node is *backpropagated* to the root node, updating the value of each ancestor node on the way using expected value.

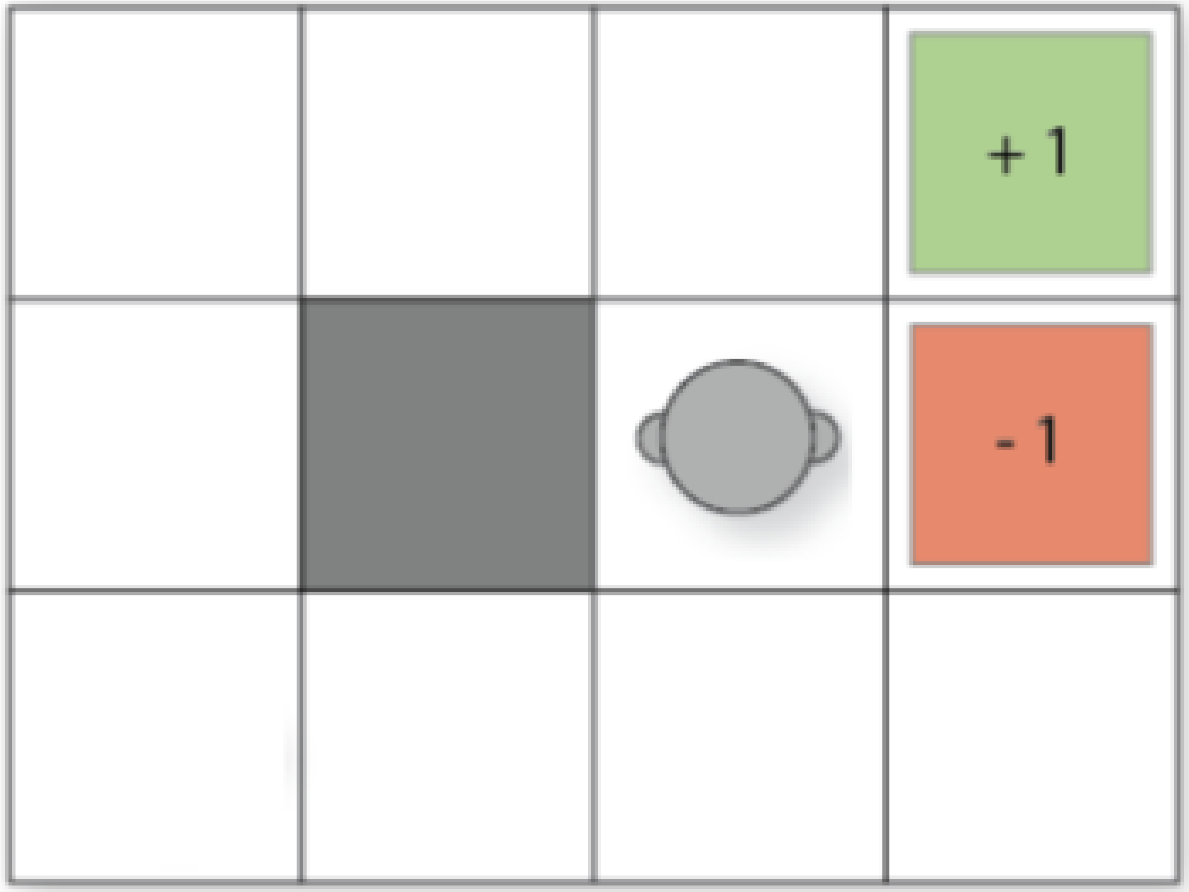
MCTS



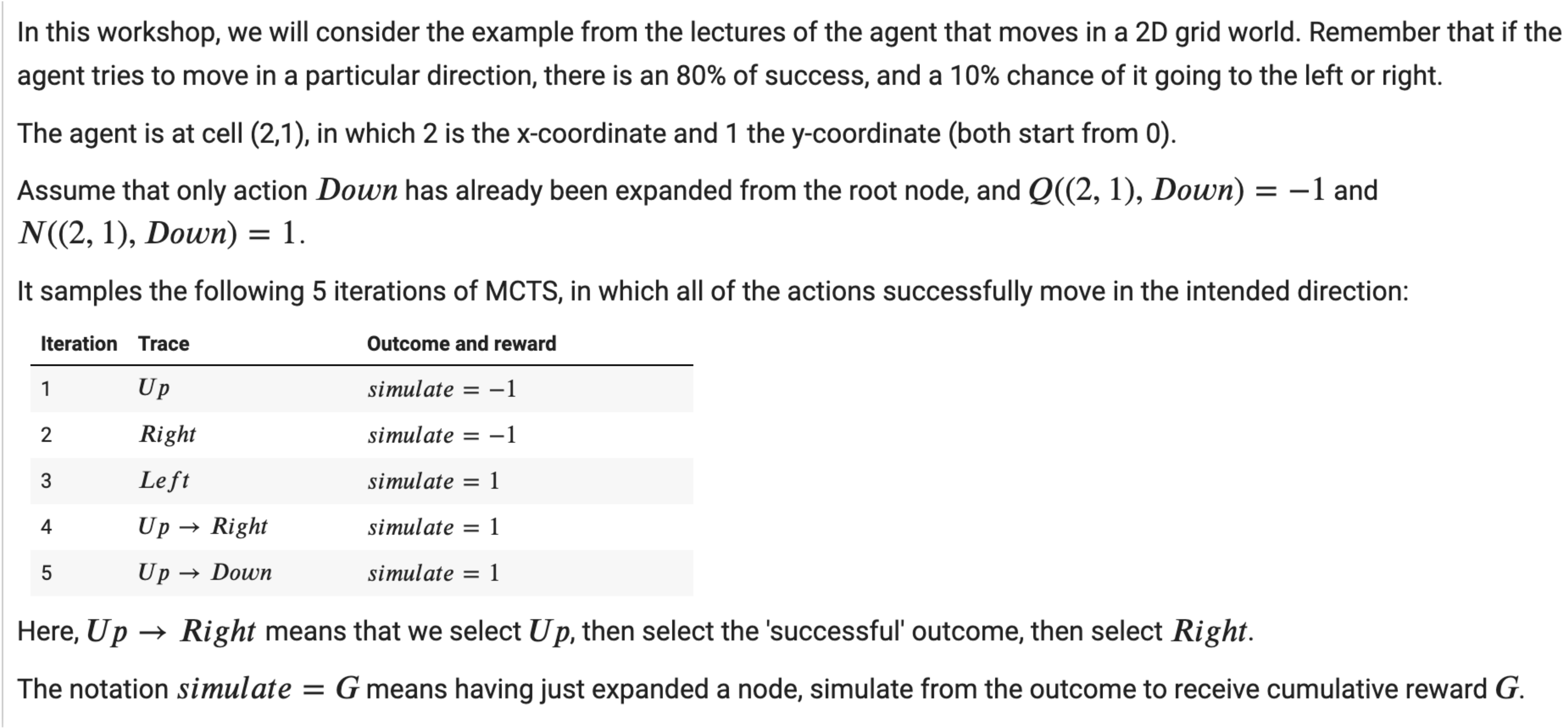


$$Q(s,a) \quad = \quad Q(s,a) + \frac{1}{N(s,a)}\left[r + \gamma G - Q(s,a)\right]$$

Lecture Example



Problem 1



In this workshop, we will consider the example from the lectures of the agent that moves in a 2D grid world. Remember that if the agent tries to move in a particular direction, there is an 80% of success, and a 10% chance of it going to the left or right.

The agent is at cell (2,1), in which 2 is the x-coordinate and 1 the y-coordinate (both start from 0).

Assume that only action $Down$ has already been expanded from the root node, and $Q((2, 1), Down) = -1$ and $N((2, 1), Down) = 1$.

It samples the following 5 iterations of MCTS, in which all of the actions successfully move in the intended direction:

Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Here, $Up \rightarrow Right$ means that we select Up, then select the 'successful' outcome, then select $Right$.

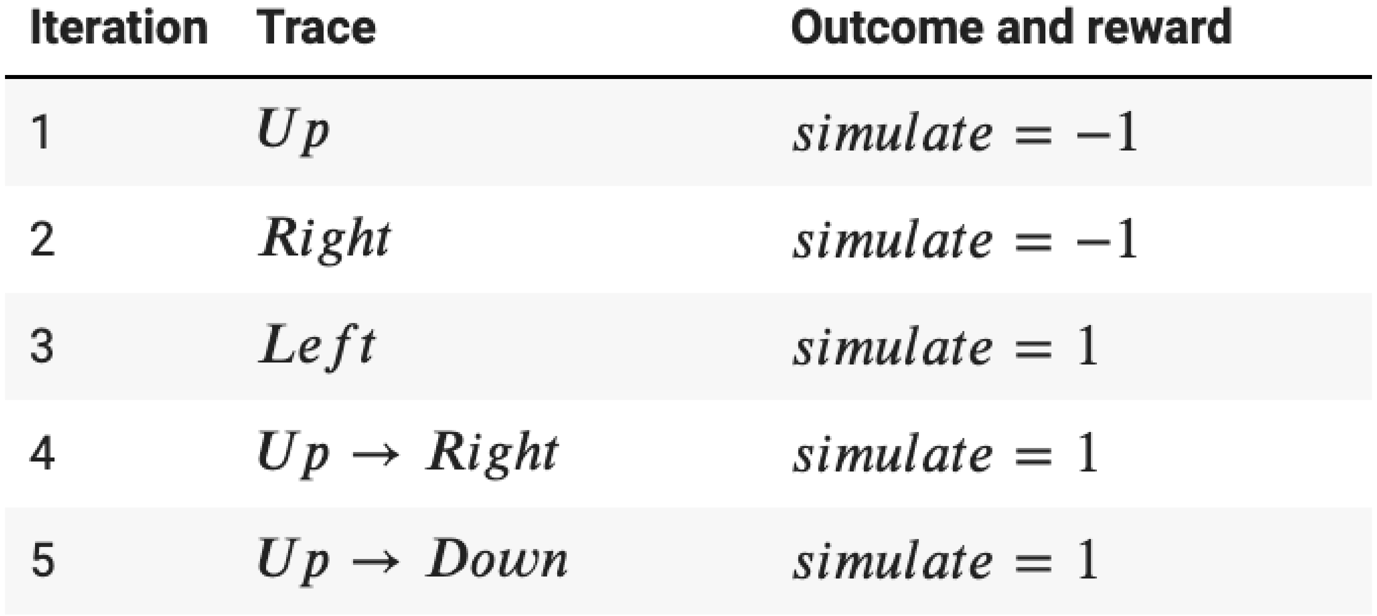
The notation $simulate = G$ means having just expanded a node, simulate from the outcome to receive cumulative reward G.

Problem 1

- Draw the MCTS tree for this, assuming γ⍰⍰⍰⍰. Calculate all of the Q-values Q⍰s⍰a⍰ and the number of times that each state-action pair been selected, N⍰s⍰a⍰, after each iteration.

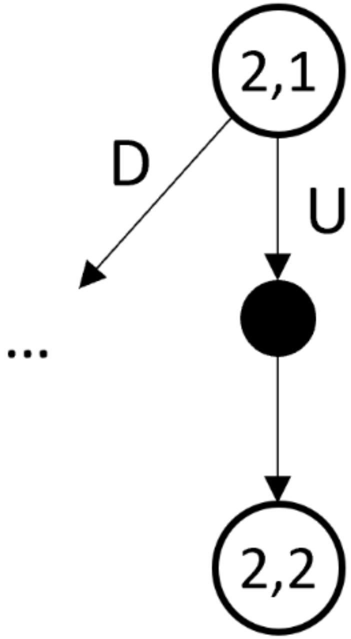
$$Q(s,a) \quad = \quad Q(s,a) + \frac{1}{N(s,a)}\left[r + \gamma G - Q(s,a)\right]$$

Problem 1



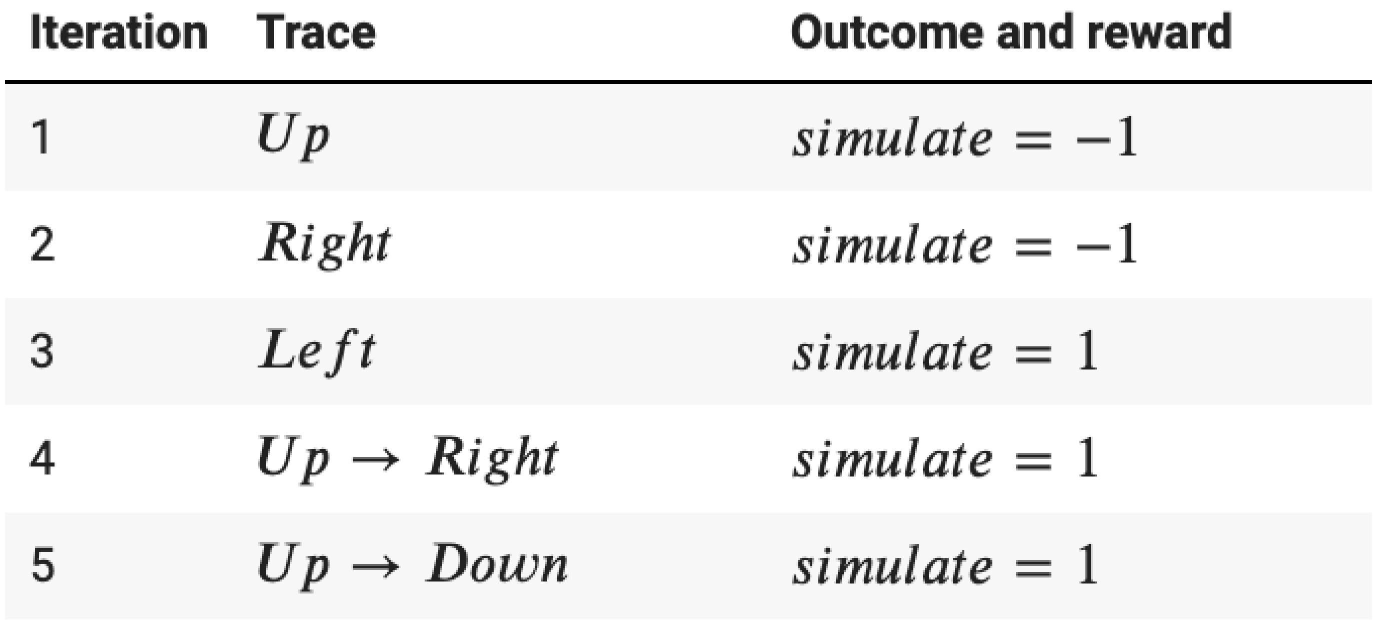
Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Iteration 1



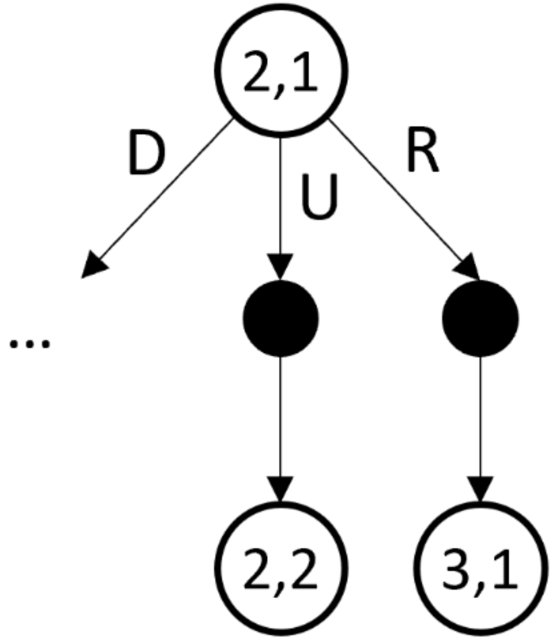
Q(2,1, U) = -1

Problem 1



Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Iteration 2

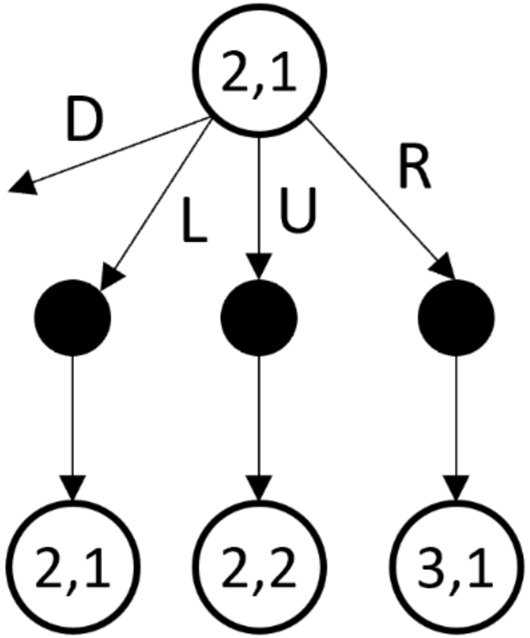


Q(2,1, U) = -1
Q(2,1, R) = -1

Problem 1

Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Iteration 3

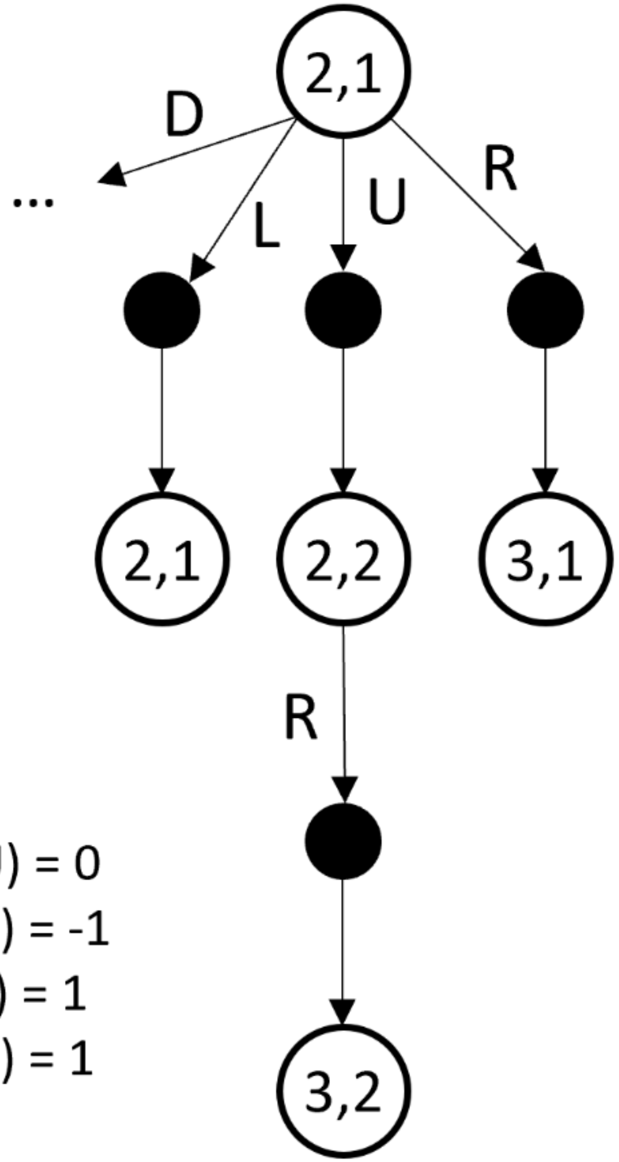


Q(2,1, U) = -1
Q(2,1, R) = -1
Q(2,1, L) = 1

Problem 1

Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Iteration 4

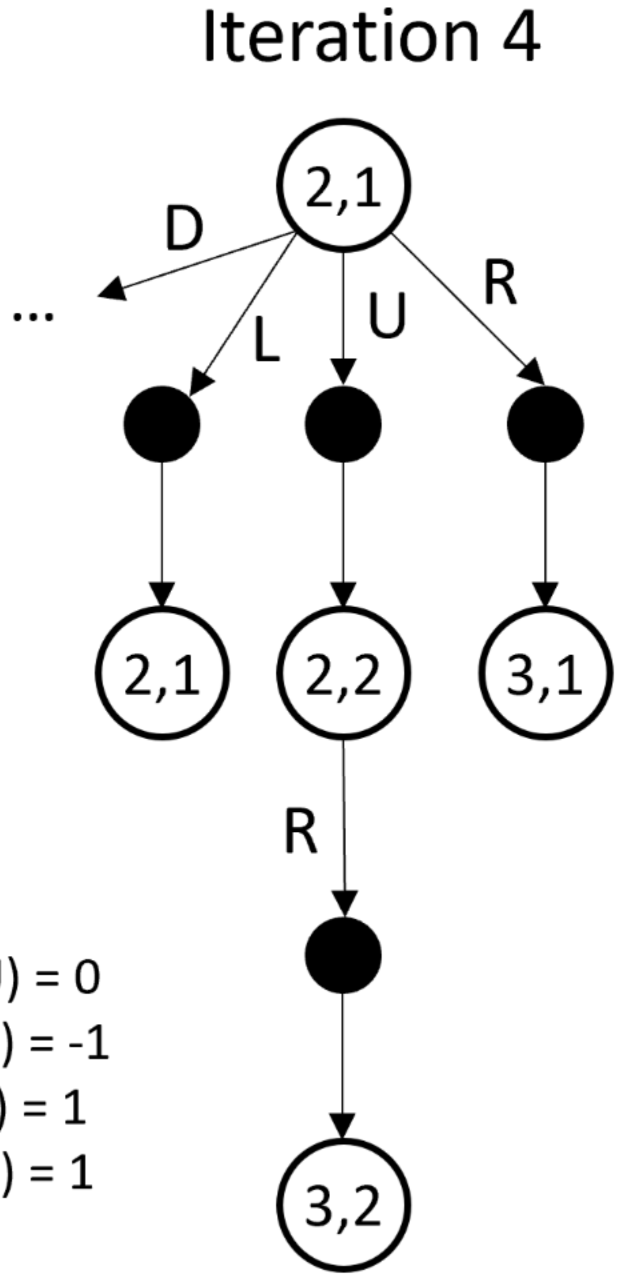


Q(2,1, U) = 0
Q(2,1, R) = -1
Q(2,1, L) = 1
Q(2,2, R) = 1

Problem 1

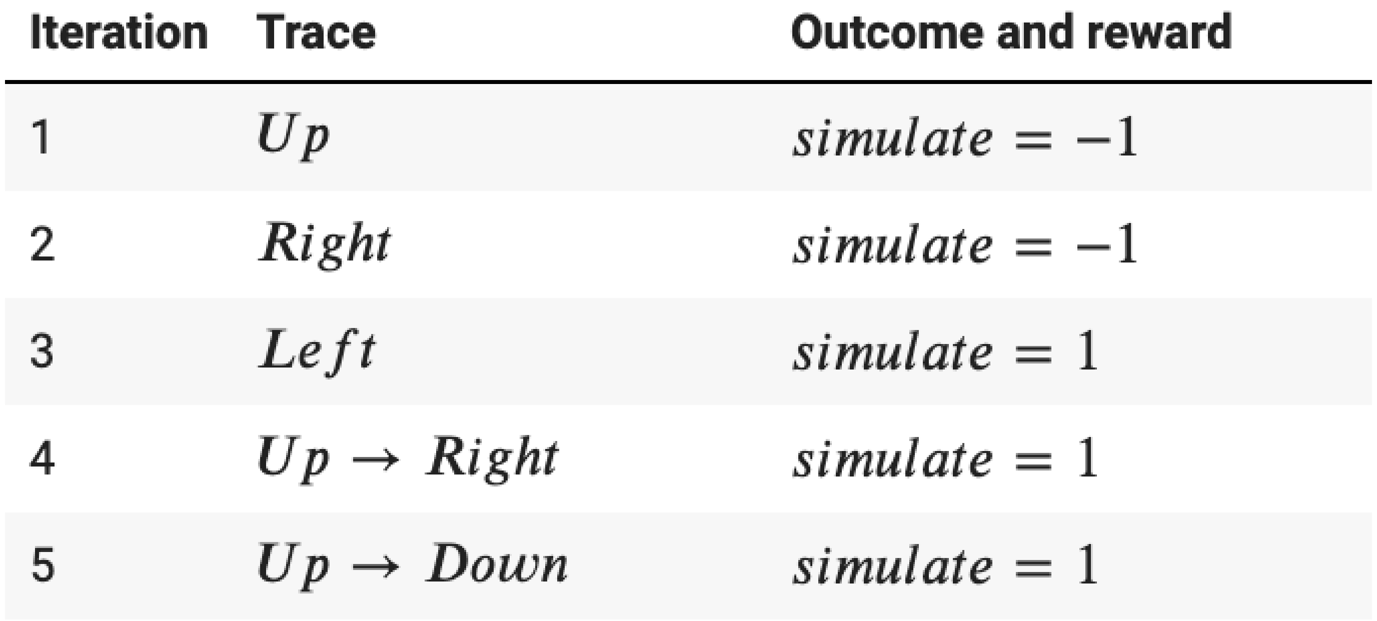
Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

$$\begin{aligned} Q(2,1,Up) &= Q(2,1,Up) + \tfrac{1}{N(2,1,Up)}[r + \gamma G - Q(2,1,Up)] \\ &= -1 + \tfrac{1}{2}[0 + 1 \cdot 1 - (-1)] \\ &= 0 \end{aligned}$$



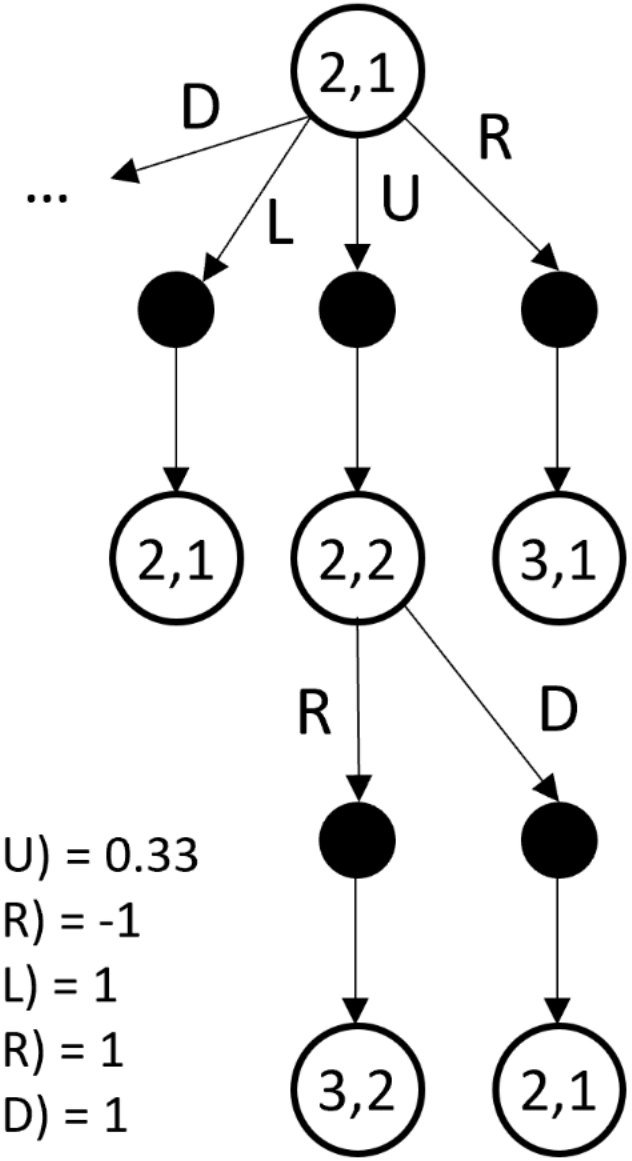
Q(2,1, U) = 0
Q(2,1, R) = -1
Q(2,1, L) = 1
Q(2,2, R) = 1

Problem 1



Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

Iteration 5



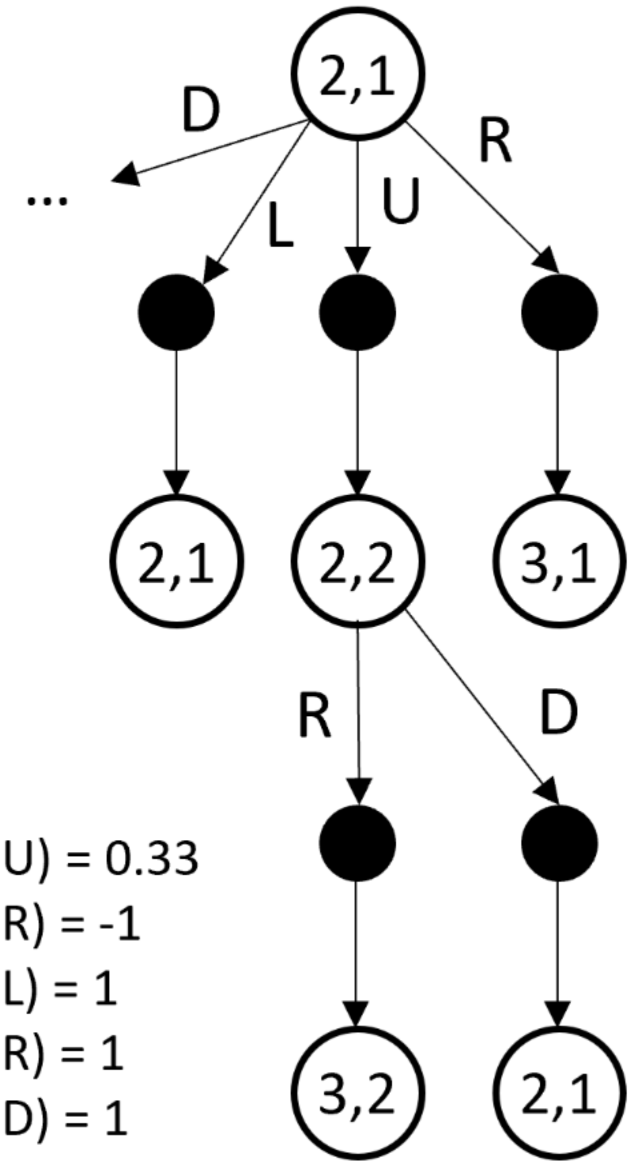
Q(2,1, U) = 0.33
Q(2,1, R) = -1
Q(2,1, L) = 1
Q(2,2, R) = 1
Q(2,2, D) = 1

Problem 1

Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$

$$\begin{aligned} Q(2,1,Up) &= Q(2,1,Up) + \frac{1}{N(2,1,Up)}[r + \gamma G - Q(2,1,Up)] \\ &= 0 + \frac{1}{3}[0 + 1 \cdot 1 - 0] \\ &= 0.33 \end{aligned}$$

Iteration 5



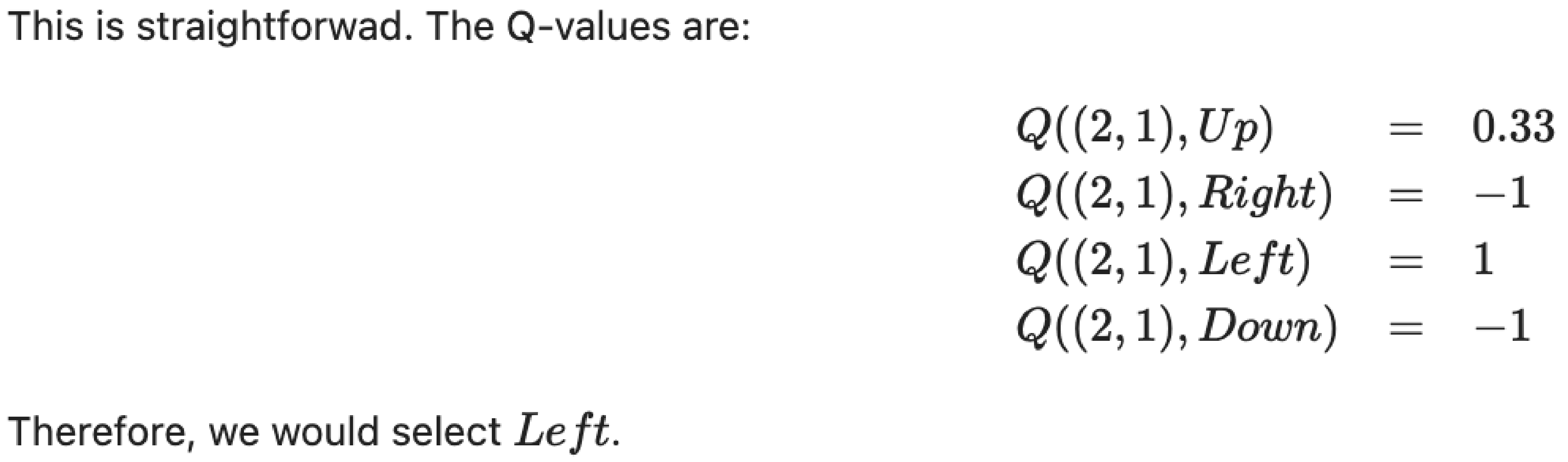
Q(2,1, U) = 0.33
Q(2,1, R) = -1
Q(2,1, L) = 1
Q(2,2, R) = 1
Q(2,2, D) = 1

Problem 2

- Based on your tree, what is the action with the highest expected return?

Problem 2

- Based on your tree, what is the action with the highest expected return?



This is straightforwad. The Q-values are:

$$\begin{aligned} Q((2,1),Up) &= 0.33 \\ Q((2,1),Right) &= -1 \\ Q((2,1),Left) &= 1 \\ Q((2,1),Down) &= -1 \end{aligned}$$

Therefore, we would select $Left$.

Problem 3

- Based on your tree, which of action, North, South, East, or West, would be more likely to be chosen if we use UCT to probabilistically select the next action? Show your working. Assume that Cp⍰⍰⍰⍰.
- Recall that there have been six iterations: the first iteration that choose $Down$ and the five iterations in the table above.

$$\text{argmax}_{a\in A(s)} Q(s,a) + 2C_p\sqrt{\frac{2\ln N(s)}{N(s,a)}}$$

Problem 3

$$\mathrm{argmax}_{a\in A(s)} Q(s,a) + 2C_p\sqrt{\frac{2\ln N(s)}{N(s,a)}}$$

$$\pi(s) \;=\; argmax_{a\in A(s)} \begin{pmatrix} Up & : & 0.33 + \sqrt{\frac{2\ln 6}{3}} \\ Right & : & -1 + \sqrt{\frac{2\ln 6}{1}} \\ Left & : & 1 + \sqrt{\frac{2\ln 6}{1}} \\ Down & : & -1 + \sqrt{\frac{2\ln 6}{1}} \end{pmatrix}$$

Problem 3

$$\mathrm{argmax}_{a\in A(s)} Q(s,a) + 2C_p\sqrt{\frac{2\ln N(s)}{N(s,a)}}$$

$$\pi(s) \;=\; argmax_{a\in A(s)} \begin{pmatrix} Up & : & 0.33+\sqrt{\frac{2\ln 6}{3}} \\ Right & : & -1+\sqrt{\frac{2\ln 6}{1}} \\ Left & : & 1+\sqrt{\frac{2\ln 6}{1}} \\ Down & : & -1+\sqrt{\frac{2\ln 6}{1}} \end{pmatrix}$$

Problem 3

$$\mathrm{argmax}_{a \in A(s)} Q(s,a) + 2C_p \sqrt{\frac{2 \ln N(s)}{N(s,a)}}$$

$$\pi(s) \;=\; argmax_{a \in A(s)} \begin{pmatrix} Up & : & 0.33 + 1.09 = 1.42 \\ Right & : & -1 + 1.89 = 0.89 \\ Left & : & 1 + 1.89 = 2.89 \\ Down & : & -1 + 1.89 = 0.89 \end{pmatrix}$$