Week 10: Monte-Carlo Tree Search

COMP90054 – AI Planning for Autonomy

Key concepts

- Monte Carlo Tree Search (MCTS)
- Compare and contrast online and offline learning

Online learning vs Offline learning

Online learning	Offline learning
An action is selected online at each state. Once an action is executed, we start planning again from the new state E.g. MCTS	We solve the problem offline for all possible states, and then use the solution (a policy) online to act E.g. Value Iteration
We need access to a *simulator* that approximates the transitions function P_a(s'|s) and reward function r of our MDP. The simulator allows us to run repeated simulations of possible futures to gain an idea of what moves are likely to be good moves compared to others.	

Workshop Problem

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).

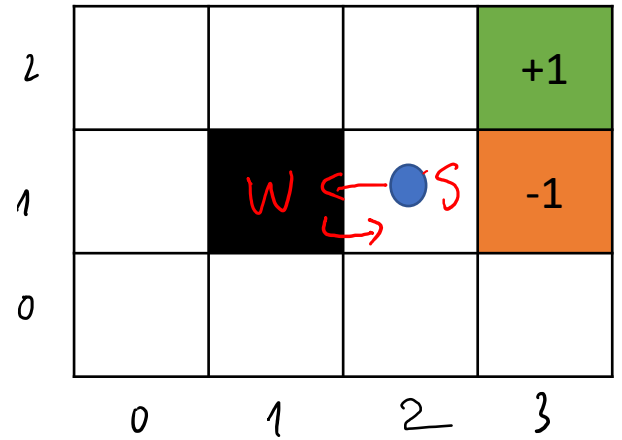
Assume that only action $Down$ has already been expanded from the root node, and $Q((2,1), Down) = -1$ and $N((2,1), Down) = 1$.

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.

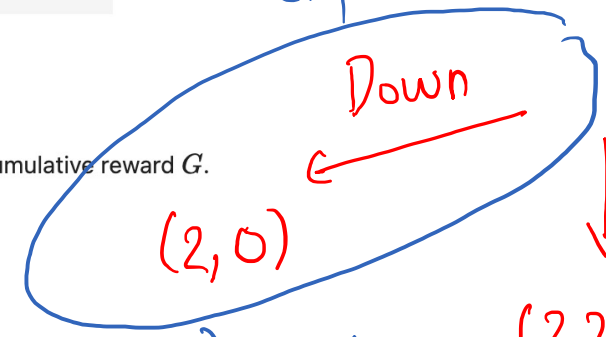


Deterministic action

Q = estimate reward

N = no of visits

expanded



(2,1) Root node

Up → (2,2) Q = ?

Right → (3,1) Q = ?

Left → (2,1) Q = ?

Start at (2,1), what is the best action?

→ Run some simulations to estimate the Q-value for all actions

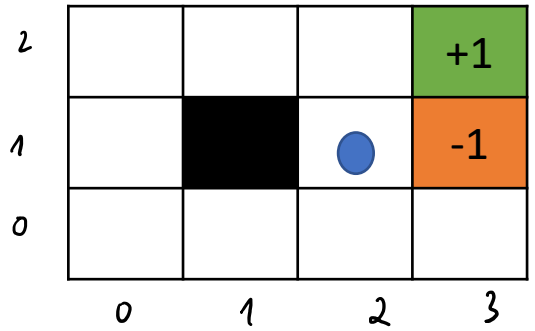
→ Choose the action with max Q

$Q((2,1), Down) = -1$

$N((2,1), Down) = 1$

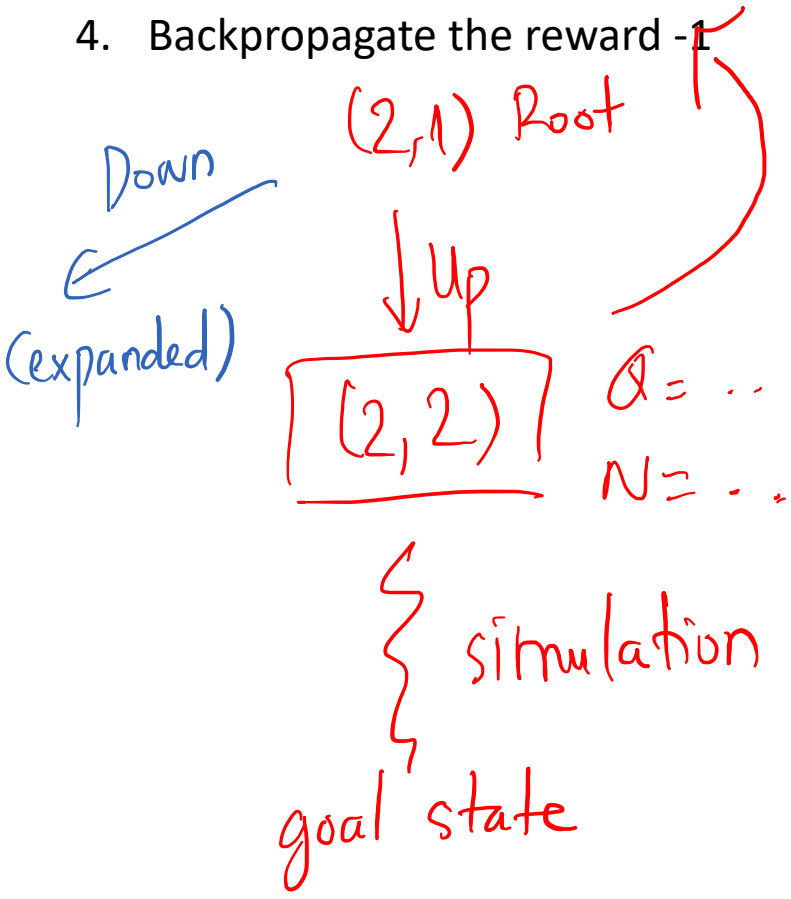
Problem 1

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 1

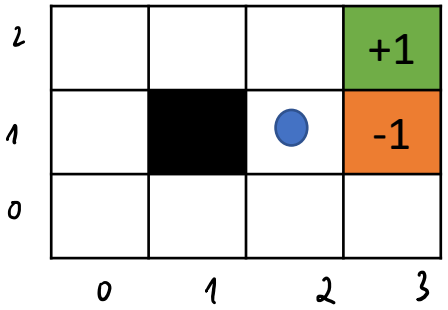
1. Select (2, 1)
2. Expand Up
3. Do simulation from (2, 2)
4. Backpropagate the reward -1



1. Selection: 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.
2. Expansion: Expand this node by applying one available action (as defined by the MDP) from the node.
3. 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.
4. Backpropagation: 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.

Update Q, N for all parent nodes

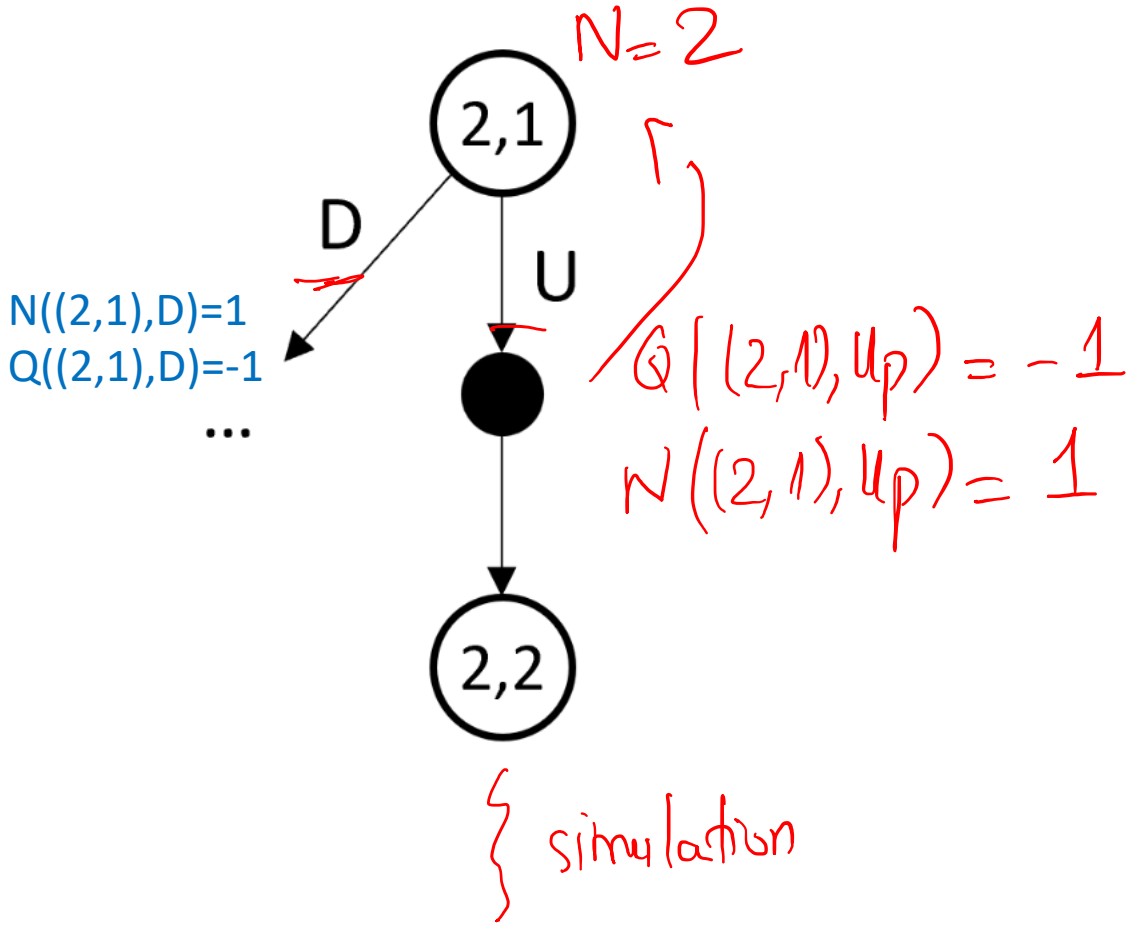
Problem 1



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

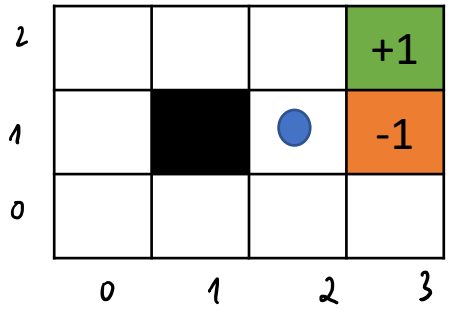
1. Select (2, 1)
2. Expand Up
3. Do simulation from (2, 2)
4. Backpropagate the reward -1 to Q((2,1), Up)



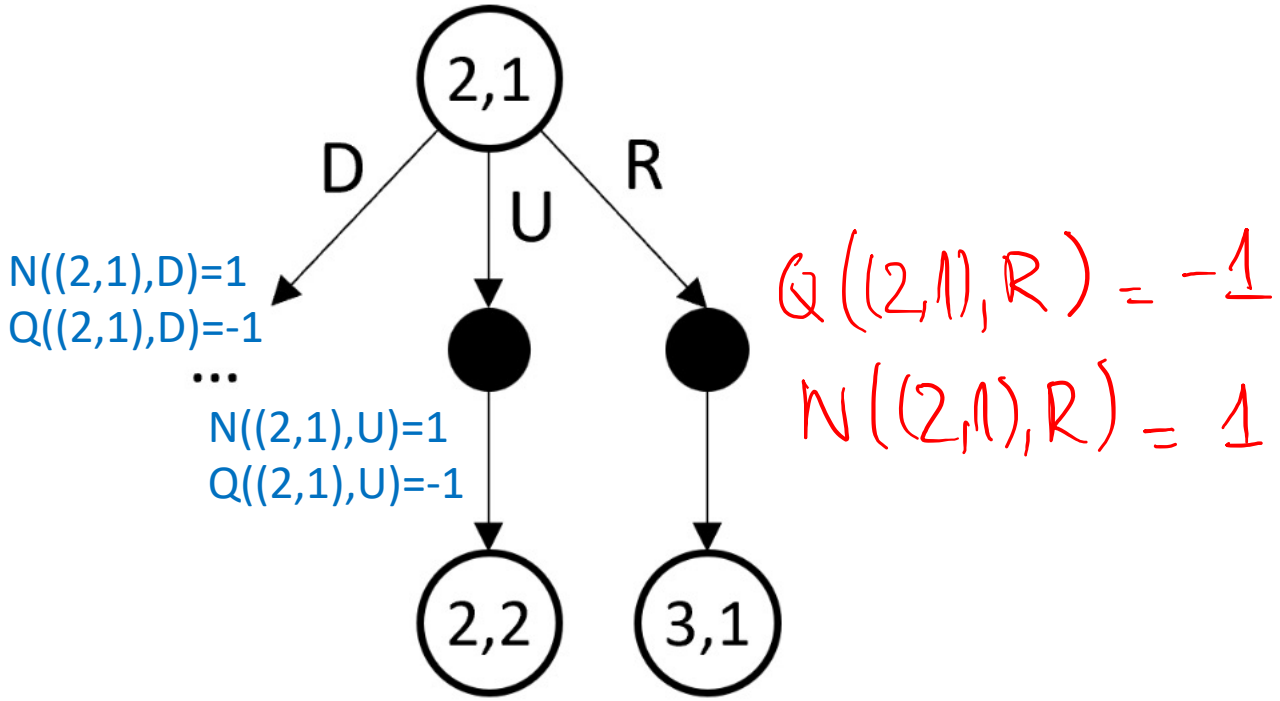
Problem 1

Iteration 2

1. Select (2, 1)
2. Expand Right
3. Do simulation from (3, 1)
4. Backpropagate the reward -1



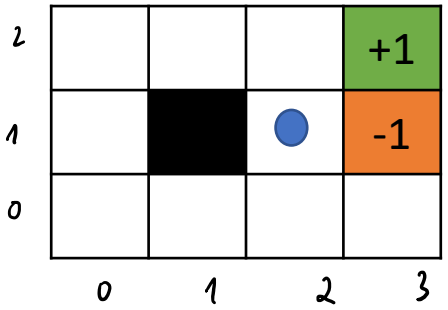
Iteration	Trace	Outcome and reward
1	Up	$simulate = -1$
2	$Right$	$simulate = -1$
3	$Left$	$simulate = 1$
4	$Up \rightarrow Right$	$simulate = 1$
5	$Up \rightarrow Down$	$simulate = 1$



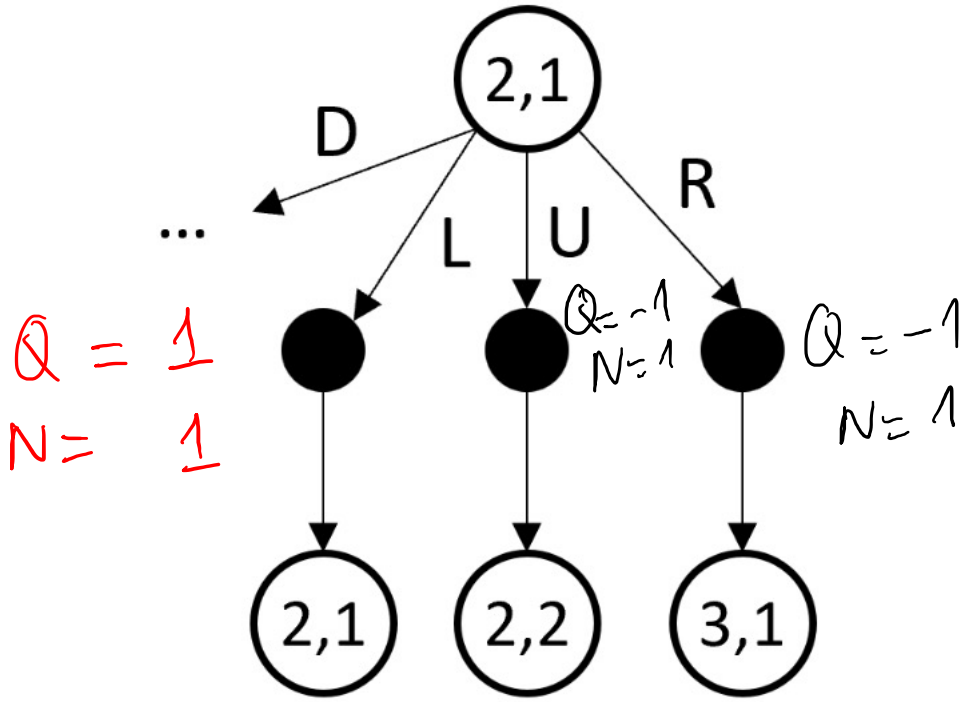
Problem 1

Iteration 3

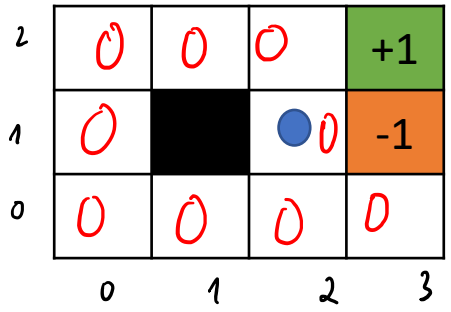
1. Select (2, 1)
2. Expand Left
3. Do simulation from (2, 1)
4. Backpropagate the reward 1



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



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	$Up \rightarrow Down$	$simulate = 1$

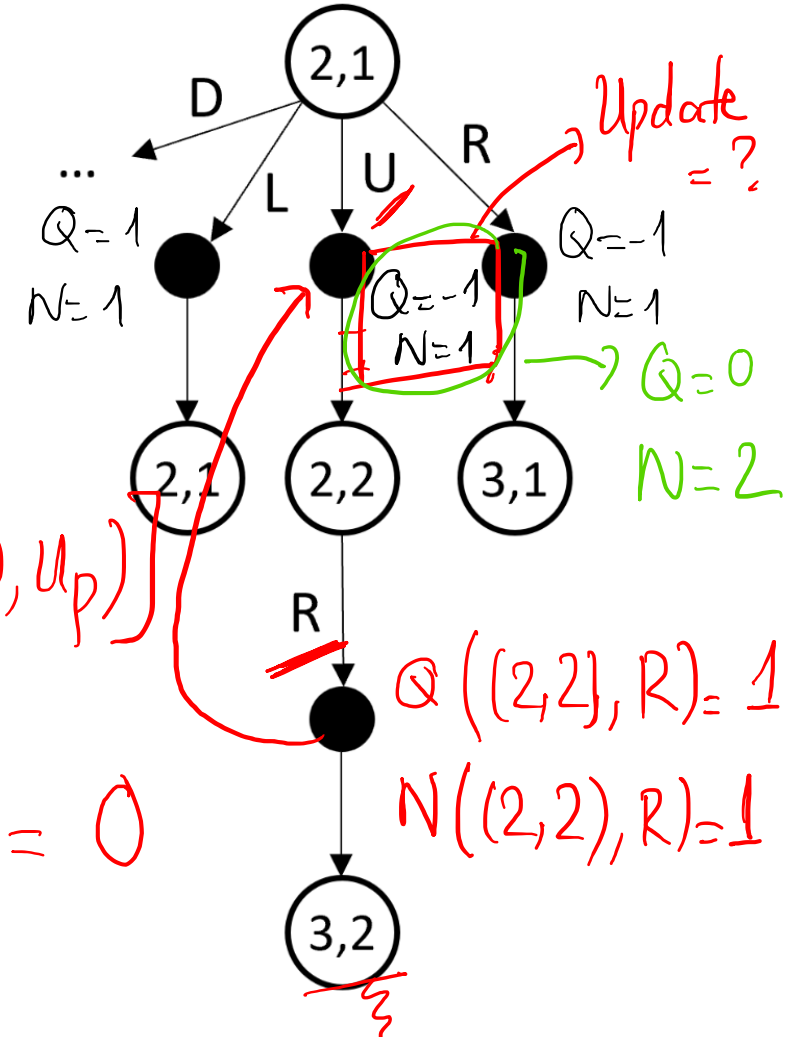
(G)

Iteration 4

1. Select (2, 1): the node is fully expanded, so recursively select Up
2. Expand Right
3. Do simulation from (3, 2)
4. Backpropagate the reward 1

r = 0

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

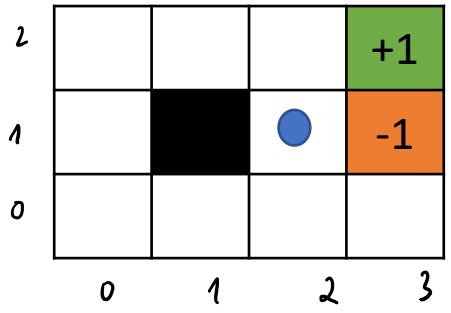


$N((2,1),Up) = 1+1 = 2$

$Q((2,1),Up) = Q((2,1),Up) + \frac{1}{N((2,1),Up)}[r + \gamma G - Q((2,1),Up)]$

$= -1 + \frac{1}{2}[0 + 1 \times (1) - (-1)] = 0$

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

1. Select (2, 1): the node is fully expanded, so recursively select Up
2. Expand Down
3. Do simulation from (2, 1)
4. Backpropagate the reward 1

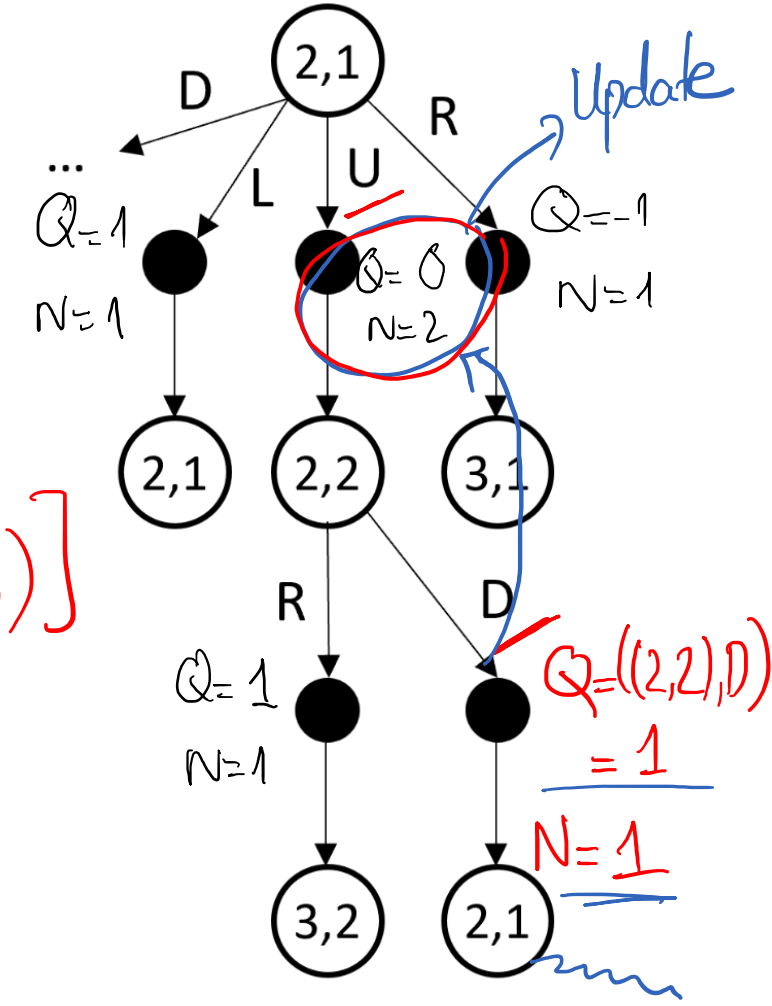
$$Q(s,a) = Q(s,a) + \frac{1}{N(s,a)}[r + \gamma G - Q(s,a)]$$

$N((2,1), Up) = 2 + 1 = 3$

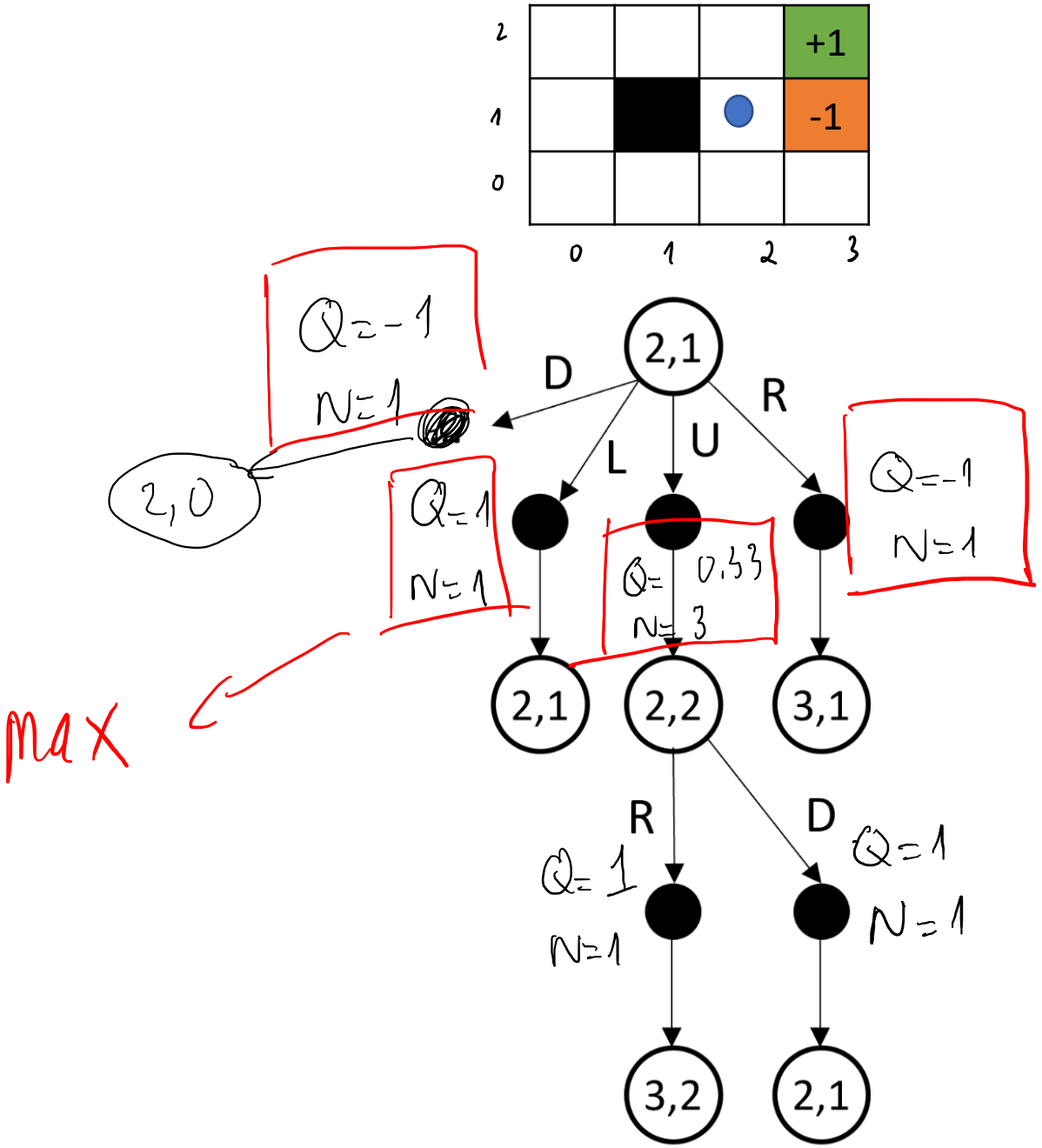
$Q((2,1), Up) = Q((2,1), Up) + \frac{1}{N((2,1), Up)}[r + \gamma(G) - Q((2,1), Up)]$

$= 0 + \frac{1}{3}[0 + 1 \times (1) - 0]$

$= 0.33$



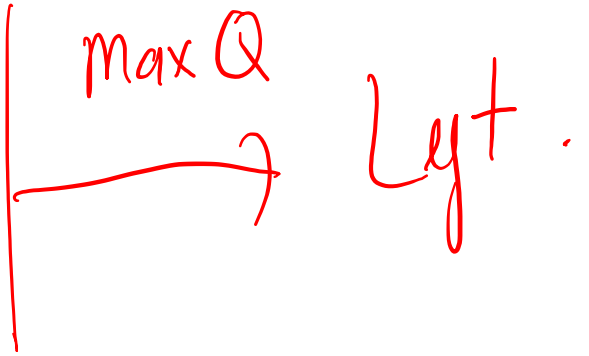
Problem 1



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

Problem 2: Action selection

Q((2, 1), Up) = 0.33
Q((2, 1), Right) = -1
Q((2, 1), Left) = 1
Q((2, 1), Down) = -1



Problem 3: Upper Confidence Trees (UCT) = MCTS + UCB1

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 work. Assume that $C_p = \frac{1}{2}$

Recall that there have been six iterations: the first iteration chooses *Down* and the five iterations in the table above.

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

Exploitation

Exploration

Exploration constant

- Exploitation: Q(s,a) is high for an action that has a high reward
- Exploration: High for actions that have been explored less
- Exploration Constant: Increase to encourage more exploration, decrease to encourage less exploration

Problem 3: Upper Confidence Trees (UCT)

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 work. Assume that $C_p = \frac{1}{2}$

Recall that there have been six iterations: the first iteration chooses *Down* and the five iterations in the table above.

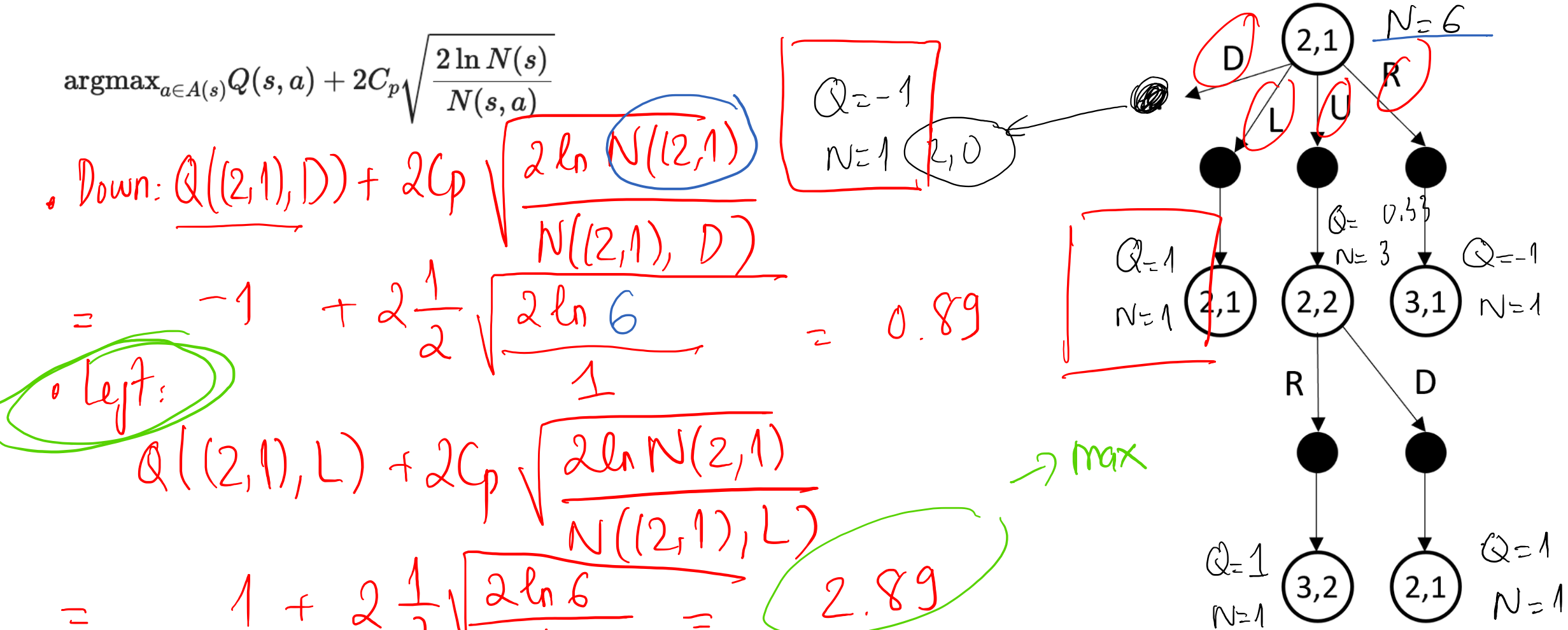
$$\text{argmax}_{a \in A(s)} Q(s,a) + 2C_p \sqrt{\frac{2 \ln N(s)}{N(s,a)}}$$

- Down: $Q((2,1),D) + 2C_p \sqrt{\frac{2 \ln N(2,1)}{N((2,1),D)}}$

$= -1 + 2\frac{1}{2}\sqrt{\frac{2 \ln 6}{1}} = 0.89$

- Left: $Q((2,1),L) + 2C_p \sqrt{\frac{2 \ln N(2,1)}{N((2,1),L)}}$

$= 1 + 2\frac{1}{2}\sqrt{\frac{2 \ln 6}{1}} = 2.89$ → max



Problem 3: Upper Confidence Trees (UCT)

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 work. Assume that $C_p = \frac{1}{2}$

Recall that there have been six iterations: the first iteration chooses *Down* and the five iterations in the table above.

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

- Up: $Q((2,1), Up) + 2C_p \sqrt{\frac{2 \ln N(2,1)}{N((2,1), Up)}}$

$= 0.33 + 2\frac{1}{2}\sqrt{\frac{2 \ln 6}{3}} = 1.42$

- Right: $Q((2,1), R) + 2C_p \sqrt{\frac{2 \ln N(2,1)}{N((2,1), R)}}$

$= -1 + 2\frac{1}{2}\sqrt{\frac{2 \ln 6}{1}} = 0.89$