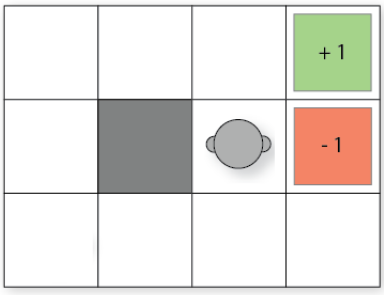
Problem Set VIII: Monte-Carlo Tree Search

Aim The purpose of this workshop is to help you get a better understanding of Monte-Carlo Tree Search for solving MDPs in an online manner.

Tasks

In this workshop, you 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.



1. The agent is at cell (2,1), in which 2 is the x-coordinate and 1 the y-coordinate (both start from 0). It samples the following 5 iterations of MCTS, where E (East) goes right, W (West) goes left, N (North) goes up, and S (South) goes down:

Iter	**Trace**	
1	N	$simulate(succ) = -1$
2	$N \xrightarrow{succ} E$	$simulate(slip(S)) = -1$
3	E	$simulate(succ) = -1$
4	W	$simulate(succ) = 1$
5	$N \xrightarrow{succ} S$	$simulate(slip(W)) = 1$

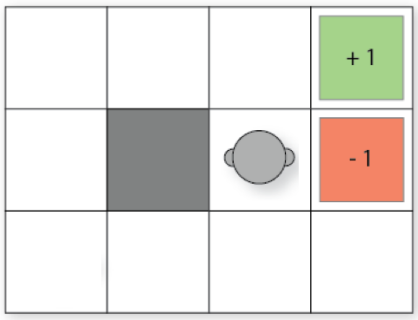
   Here, $N \xrightarrow{succ} E$ means that we select N, then select that the outcome of N was successful, and then select E. The notation $N \xrightarrow{slip(E)} S$ means that we select N and the agent went East instead; then select South.

   The notation $simulate(X) = r$ means having just expanded a node, simulate from the outcome X, which will return reward r.

   Draw the MCTS tree for this, assuming $\lambda = 1.0$. Label the lines on the tree with the actions & outcomes and label the nodes with the value $V(s)$ for each state and N (the number of times this has been visited).

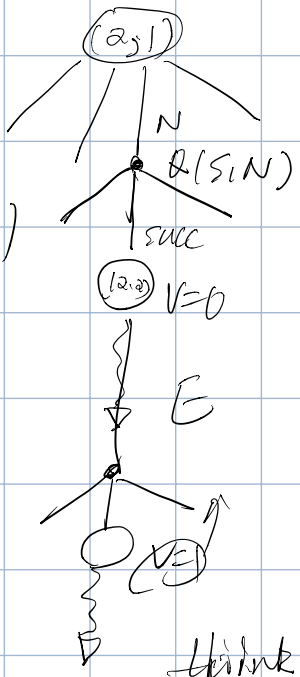
2. Based on your tree, calculate the action with the highest expected return.

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 $C_p = \frac{1}{2}$.



$$\pi(s) = \underset{a \in A(s)}{\operatorname{argmax}}\, Q(s,a)$$

$$Q(s,a) = \sum_{s'} p(r \& V(s'))$$



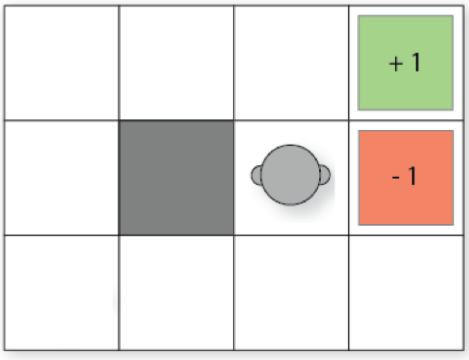
1 – select. based on policy or a set of rule.

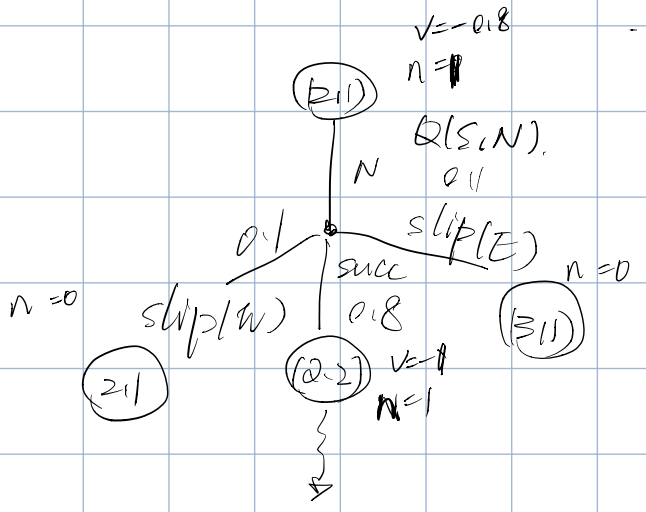
2 – expansion

3 – simulation

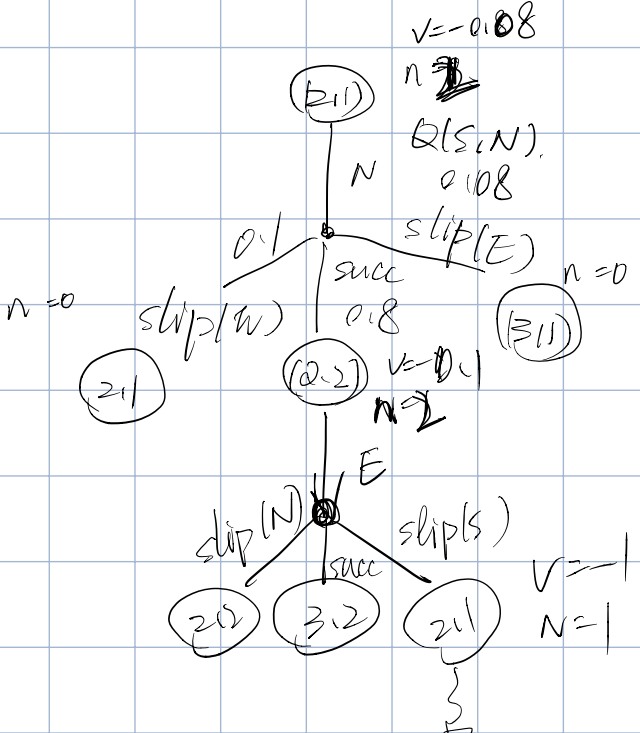
4 – backpropagation

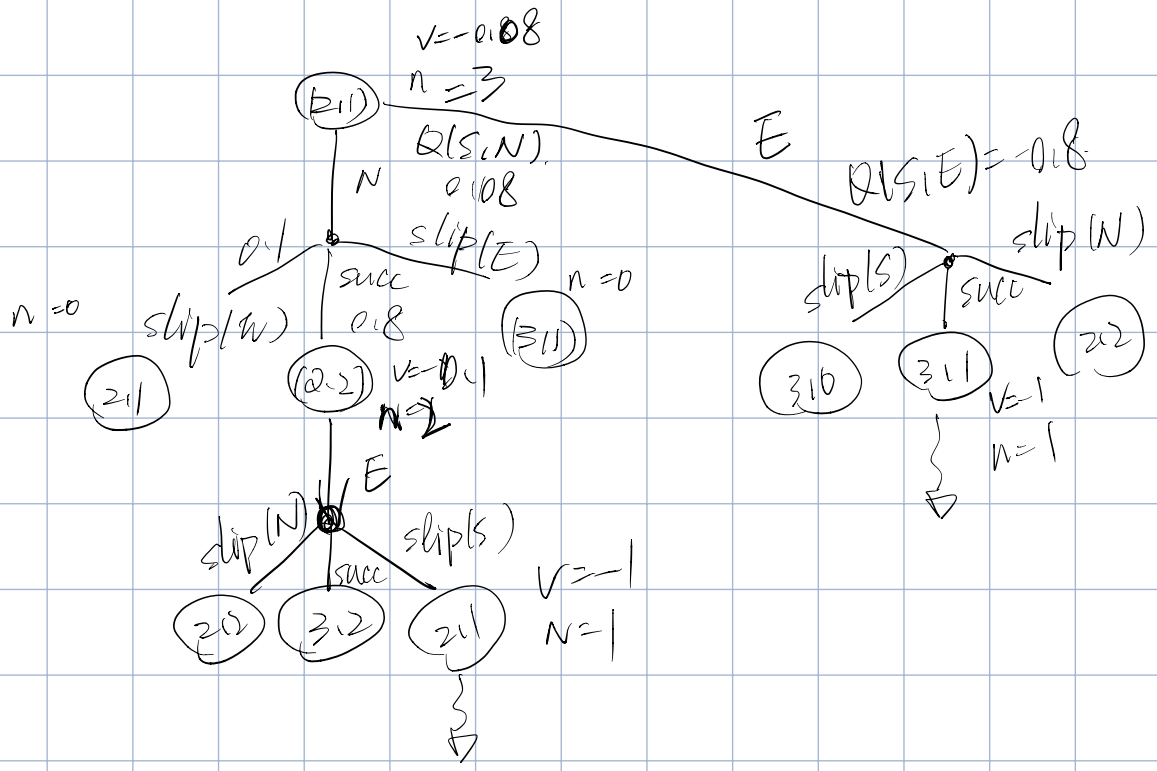
think in mind

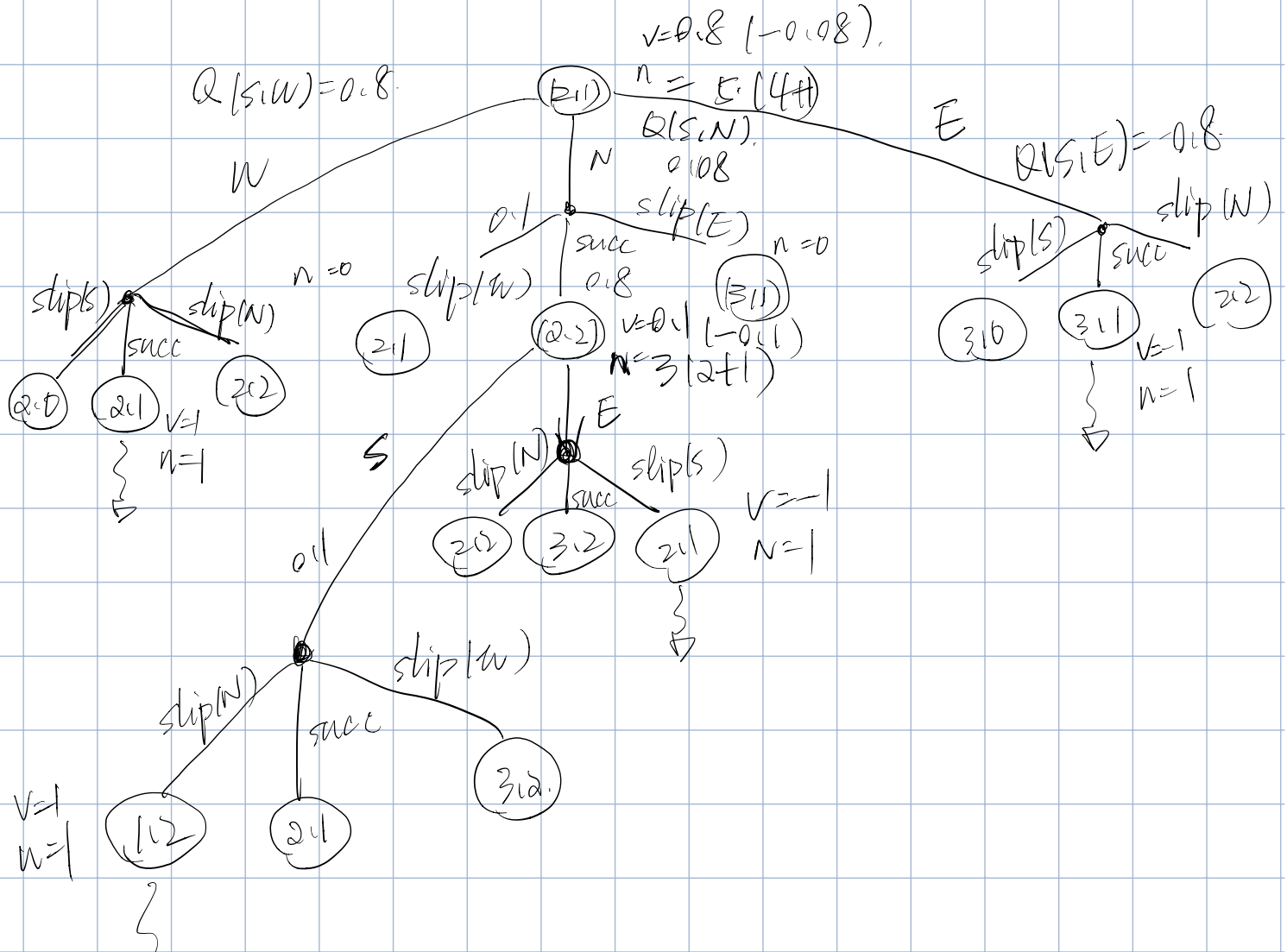




Iter	Trace	
1	N	$simulate(succ) = -1$
2	$N \xrightarrow{succ} E$	$simulate(slip(S)) = -1$
3	E	$simulate(succ) = -1$
4	W	$simulate(succ) = 1$
5	$N \xrightarrow{succ} S$	$simulate(slip(W)) = 1$




V=-0.08
n=3
Q(S,N)
0.08
N
E
0.1
slip(E)
succ
0.8
slip(W)
slip(N)
slip(S)
n=0
n=0
(2,1)
(2,2)
(3,1)
(3,1)
V=0.1
n=2
V=-1
n=1
(2,1)
(2,0)
(3,2)
(2,1)
(3,0)
(3,1)
(2,2)


V=0.8 (-0.08)
n=5 (4+1)
Q(S,W)=0.8
Q(S,N)
0.08
N
E
Q(S,E)=-0.8
W
S
0.1
slip(E)
succ
0.8
slip(W)
slip(N)
slip(S)
n=0
n=0
(2,1)
(2,2)
(3,1)
(3,1)
V=0.1 (-0.1)
n=3 (2+1)
V=-1
n=1
(2,0)
(2,1)
(2,2)
(3,0)
(3,1)
(2,2)
V=-1
n=1
(1,2)
(2,1)
(3,2)
V=1
n=1

$$UCT = \bar{R}(s,a) + 2c_p\sqrt{\frac{2\ln(N(s))}{N(s,a)}}$$

$$UCT(N) = 0.8 + 2c_p\sqrt{\frac{2\ln(5)}{3}}$$

$$UCT(W) = 0.8 + 2c_p\sqrt{\frac{2\ln(5)}{1}}$$

$$UCT(E) = 0.8 + 2c_p\sqrt{\frac{2\ln(5)}{1}}$$

$$UCT(S) = 0 + 2c_p\sqrt{\frac{2\ln(5)}{0}} = \infty$$