21AIE311 – Reinforcement Learning
Lab Assignment - 1

Submitted by :

Muppavarapu Sri Harshini : BL.EN.U4AIE21083

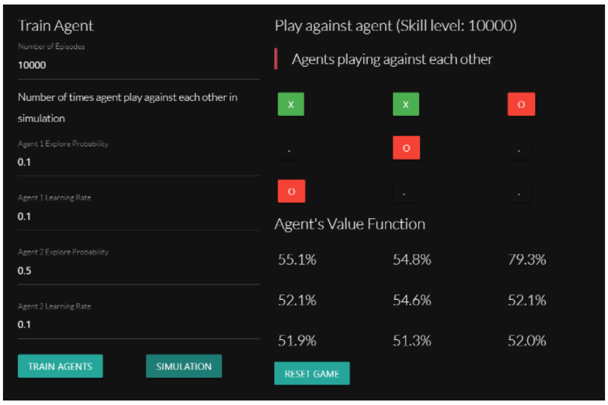
Rachuri Tarun : BL.EN.U4AIE21109

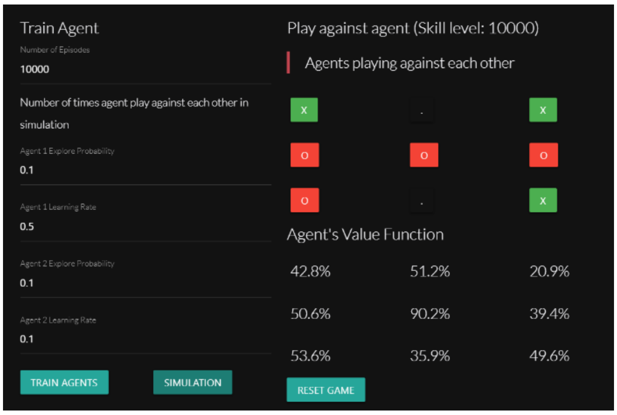
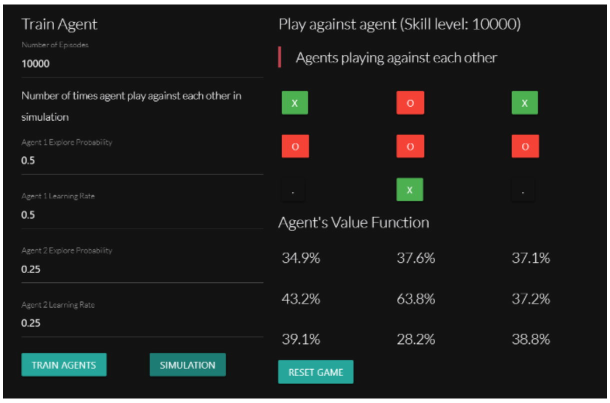
Pranav H : BL.EN.U4AIE21105

Question 1. Explore the demo of the Temporal Difference based RL learning agent to play Tic-Tac-Toe game in the below link and answer the following questions briefly to the point.

a) Agent Vs Agent - Change the hyper-parameters as below and train the agents for 10000 episodes. After each case training, make both the agents play against each other five time in simulate mode. Which player won majority of times and why?

Answer :

Case	Hyper Parameters	Agent - 1 O	Agent – 2 X	Winner	
Case 1	Exploration Probability	0.5	0.1	O	 Train Agent Number of Episodes 10000 Number of times agent play against each other in simulation Agent 1 Explore Probability 0.1 Agent 1 Learning Rate 0.1 Agent 2 Explore Probability 0.5 Agent 2 Learning Rate 0.1 TRAIN AGENTS SIMULATION Play against agent (Skill level: 10000) Agents playing against each other X X O . O . O . . Agent's Value Function 55.1% 54.8% 79.3% 52.1% 54.6% 52.1% 51.9% 51.3% 52.0% RESET GAME
	Learning Rate	0.1	0.1		
Case 2	Exploration Probability	0.1	0.1	O	

	Learning Rate	0.5	0.1		
Case 3	Exploration Probability	0.5	0.25	O	
	Learning Rate	0.5	0.25		

Train Agent

Number of Episodes
10000

Number of times agent play against each other in simulation

Agent 1 Explore Probability
0.1

Agent 1 Learning Rate
0.5

Agent 2 Explore Probability
0.1

Agent 2 Learning Rate
0.1

Play against agent (Skill level: 10000)

Agents playing against each other

X	.	X
O	O	O
O	.	X

Agent's Value Function

42.8%	51.2%	20.9%
50.6%	90.2%	39.4%
53.6%	35.9%	49.6%

Train Agent

Number of Episodes
10000

Number of times agent play against each other in simulation

Agent 1 Explore Probability
0.5

Agent 1 Learning Rate
0.5

Agent 2 Explore Probability
0.25

Agent 2 Learning Rate
0.25

Play against agent (Skill level: 10000)

Agents playing against each other

X	O	X
O	O	O
.	X	.

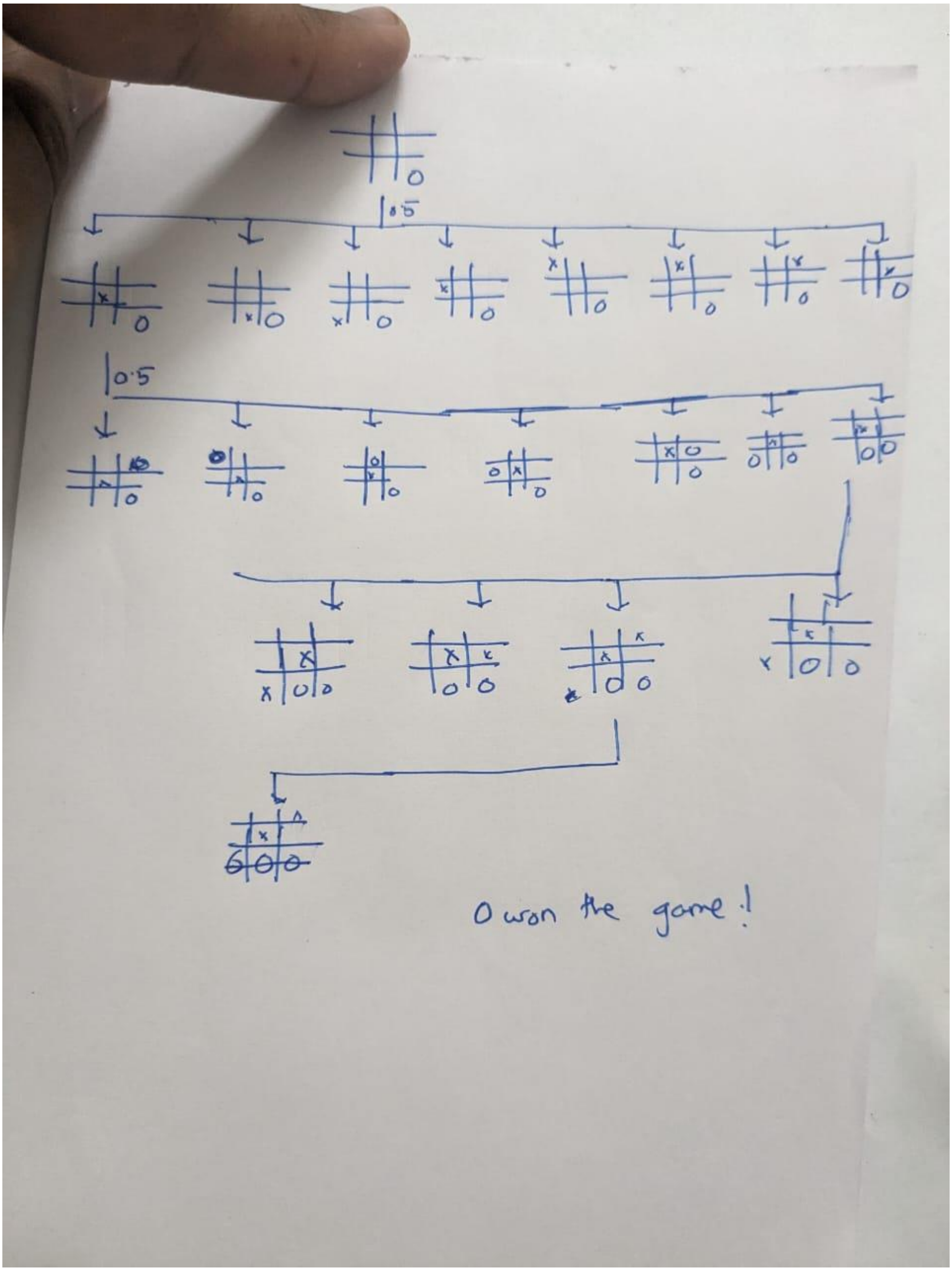
Agent's Value Function

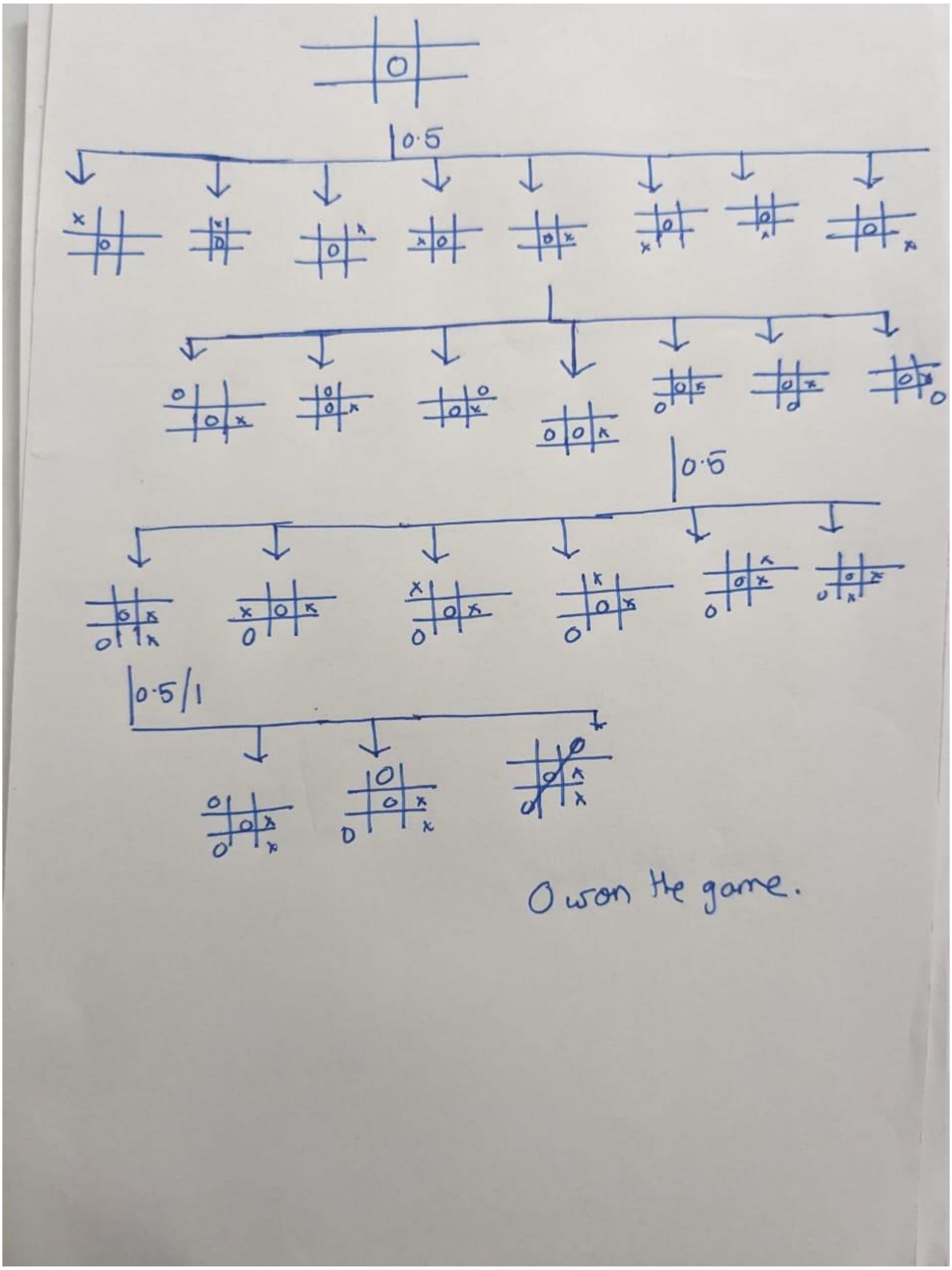
34.9%	37.6%	37.1%
43.2%	63.8%	37.2%
39.1%	28.2%	38.8%

In all the three cases, O wins the game, as shown above. This is the because
In Case – 1 :
The Exploration rate of O is greater than X and hence it wins the game.
In Case – 2 :
The Learning rate of O is greater than X and hence it wins the game.
In Case – 3 :
The Learning & Exploration rate of O is greater than X and hence it wins the game.

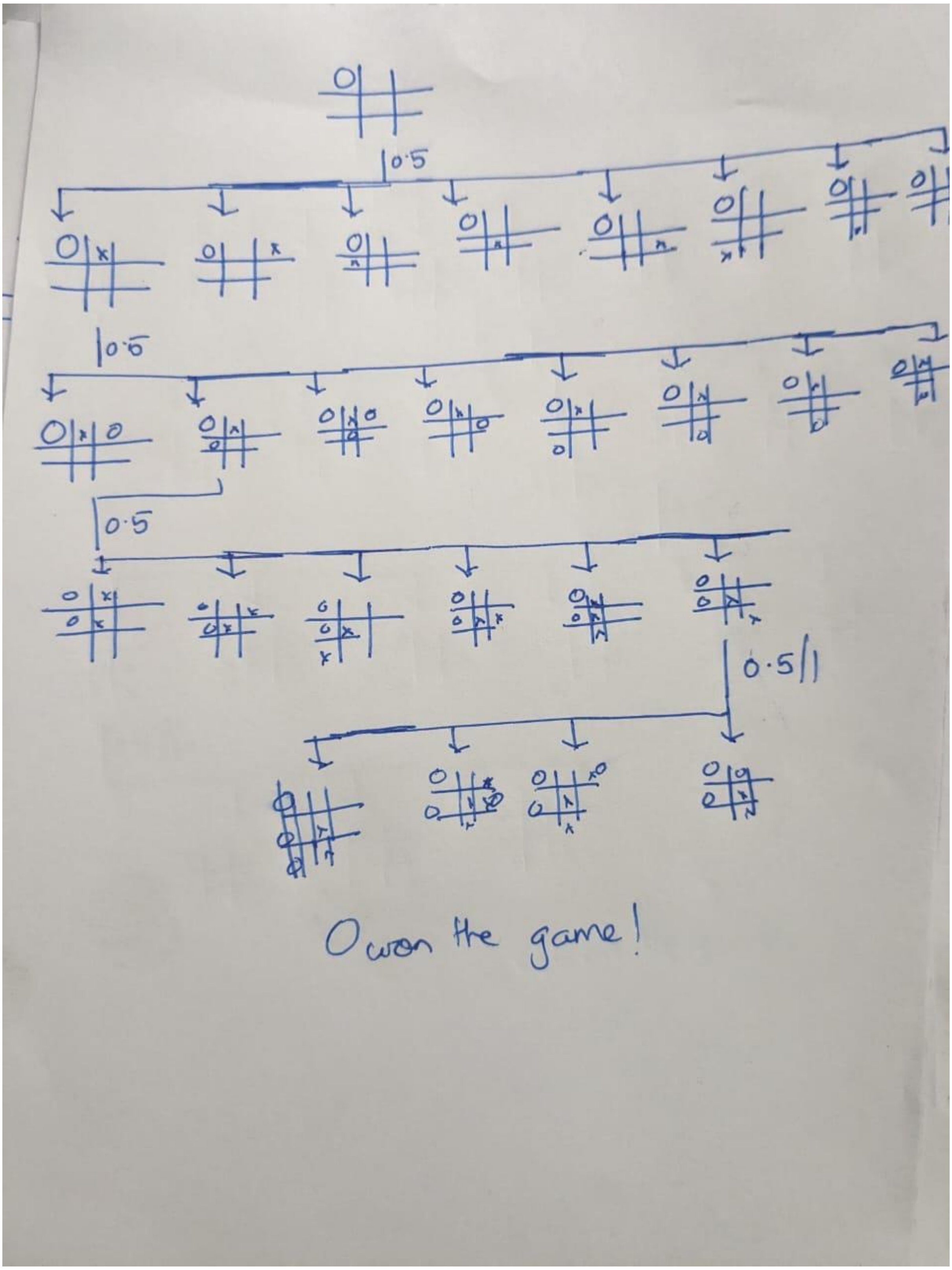
b) Develop a state tree and value table (initialization as discussed in class) for the tic-tac-toe game starting from black state (9 cells empty) and calculate the value for each state as per the next move choice. Show the final state tree and the corresponding value table for 10 episodes of training.

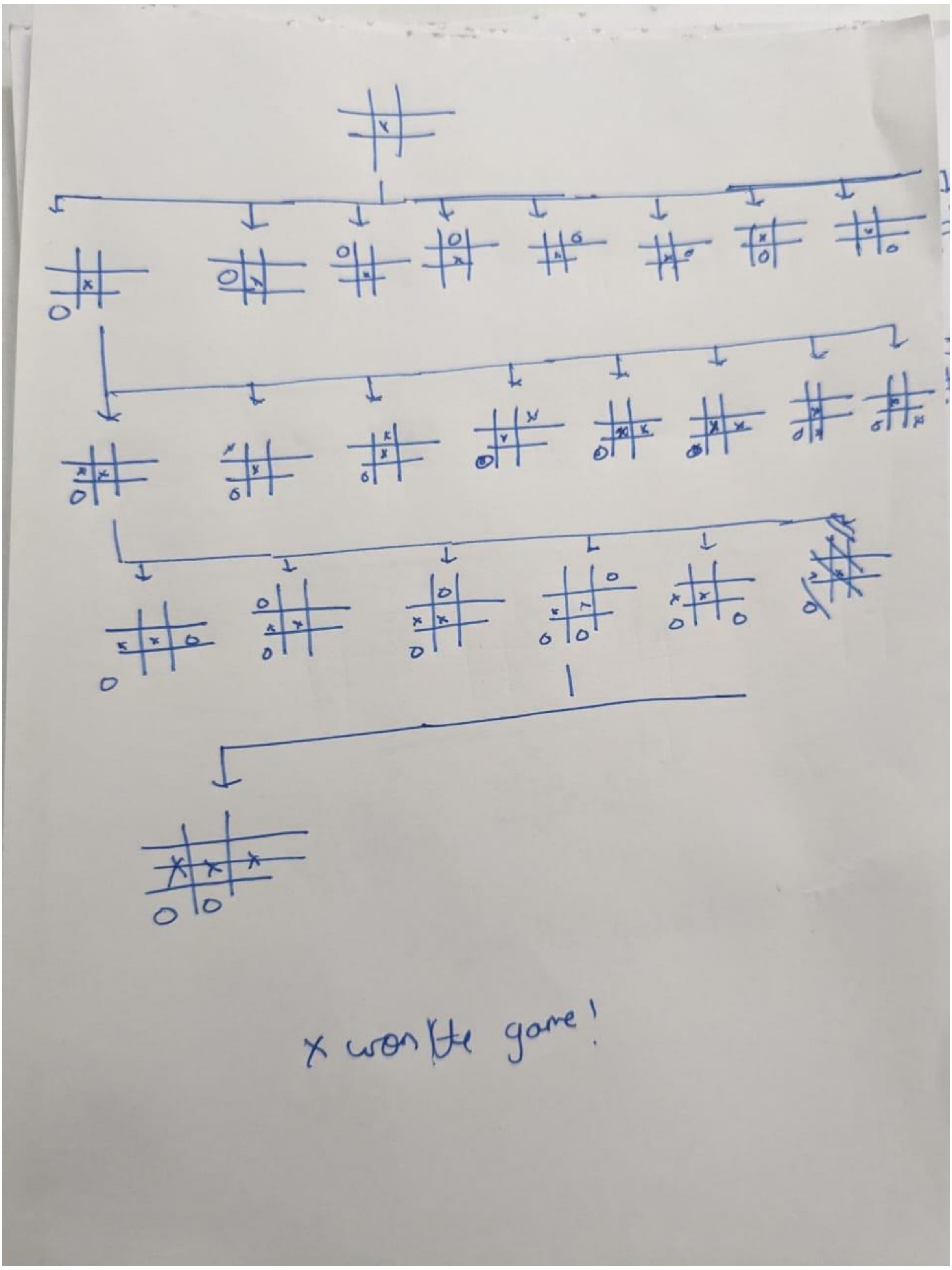
Answer :



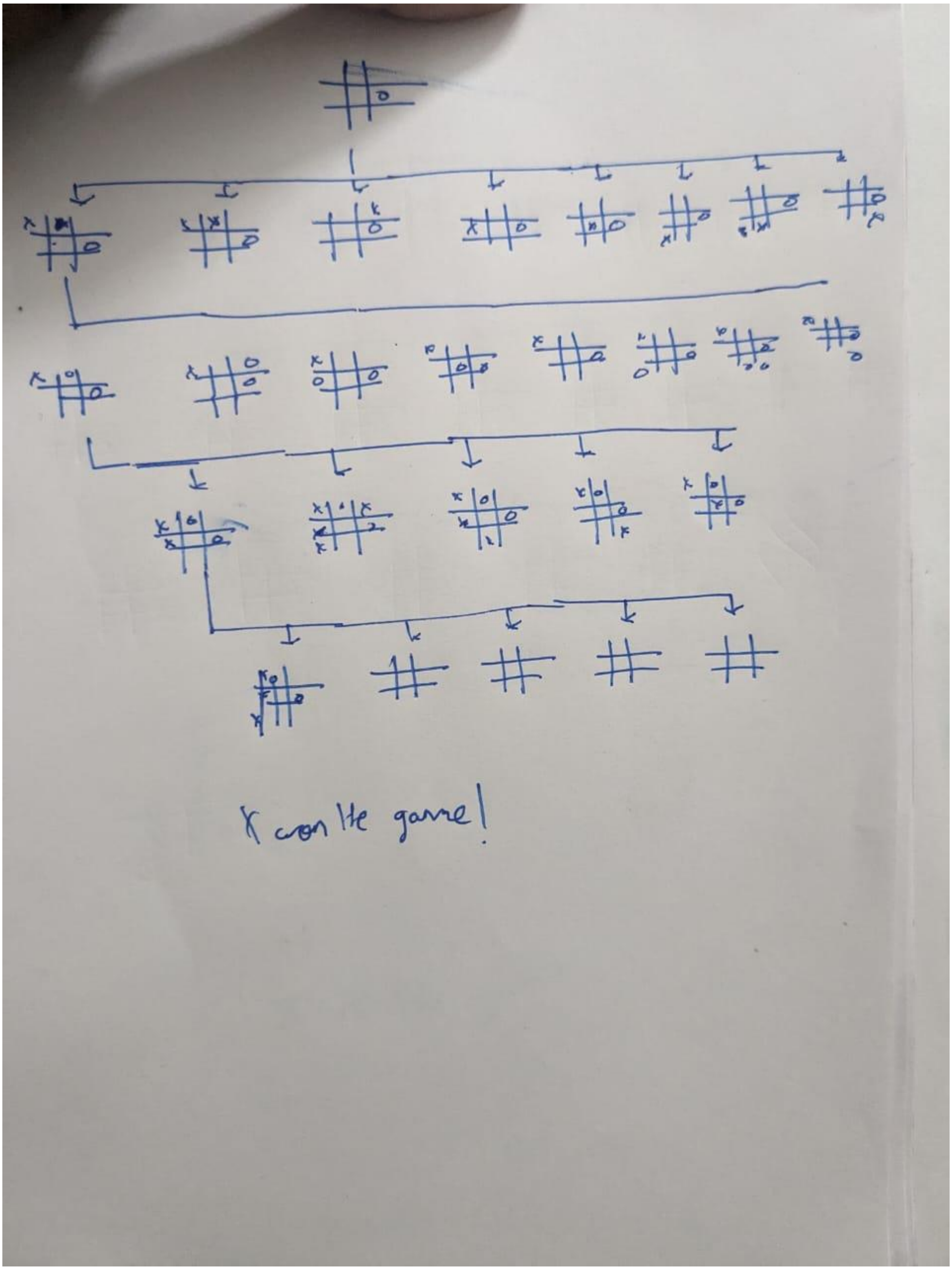


O won the game.

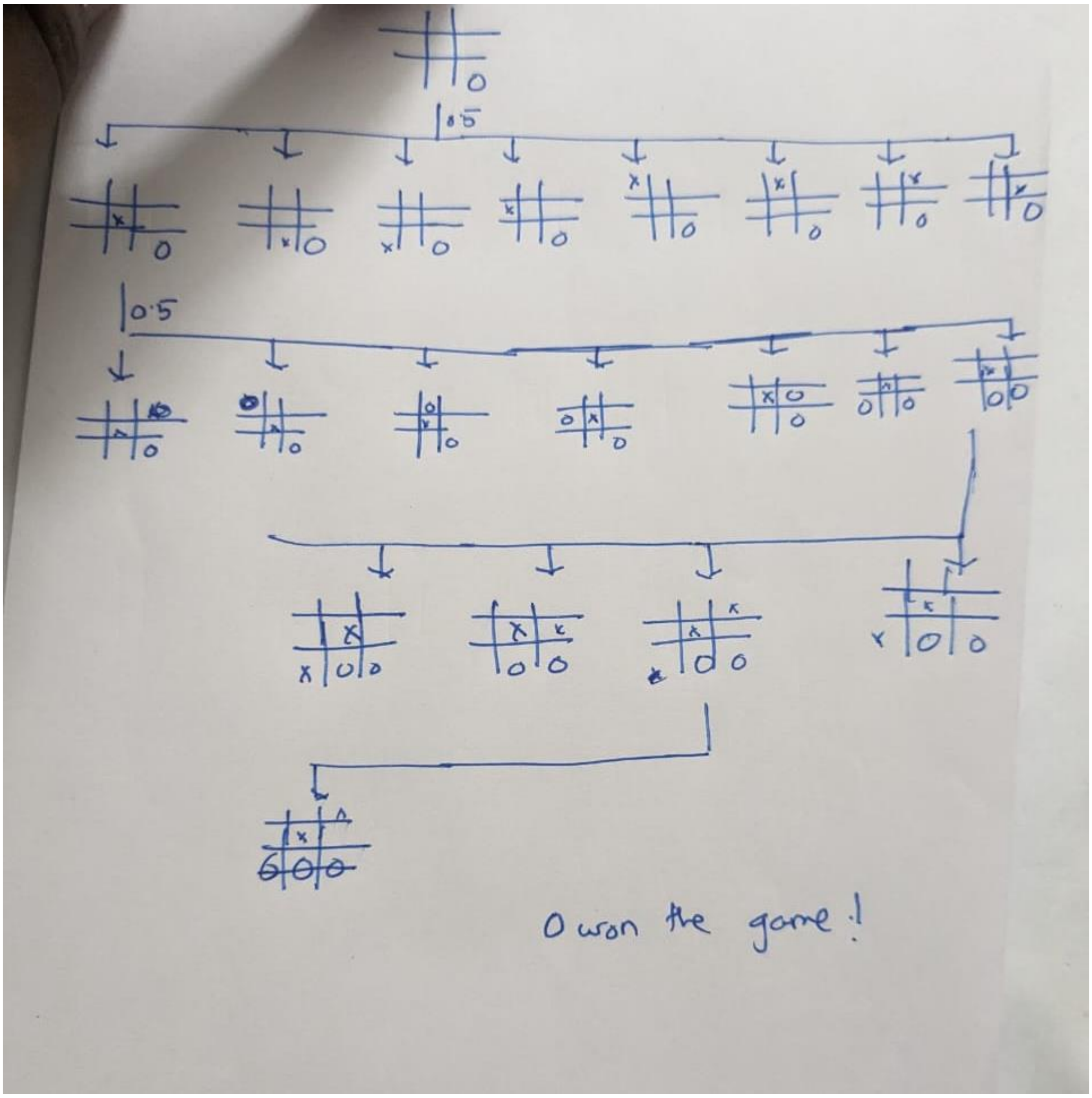

0.5
0.5
0.5
0.5/1
O won the game!

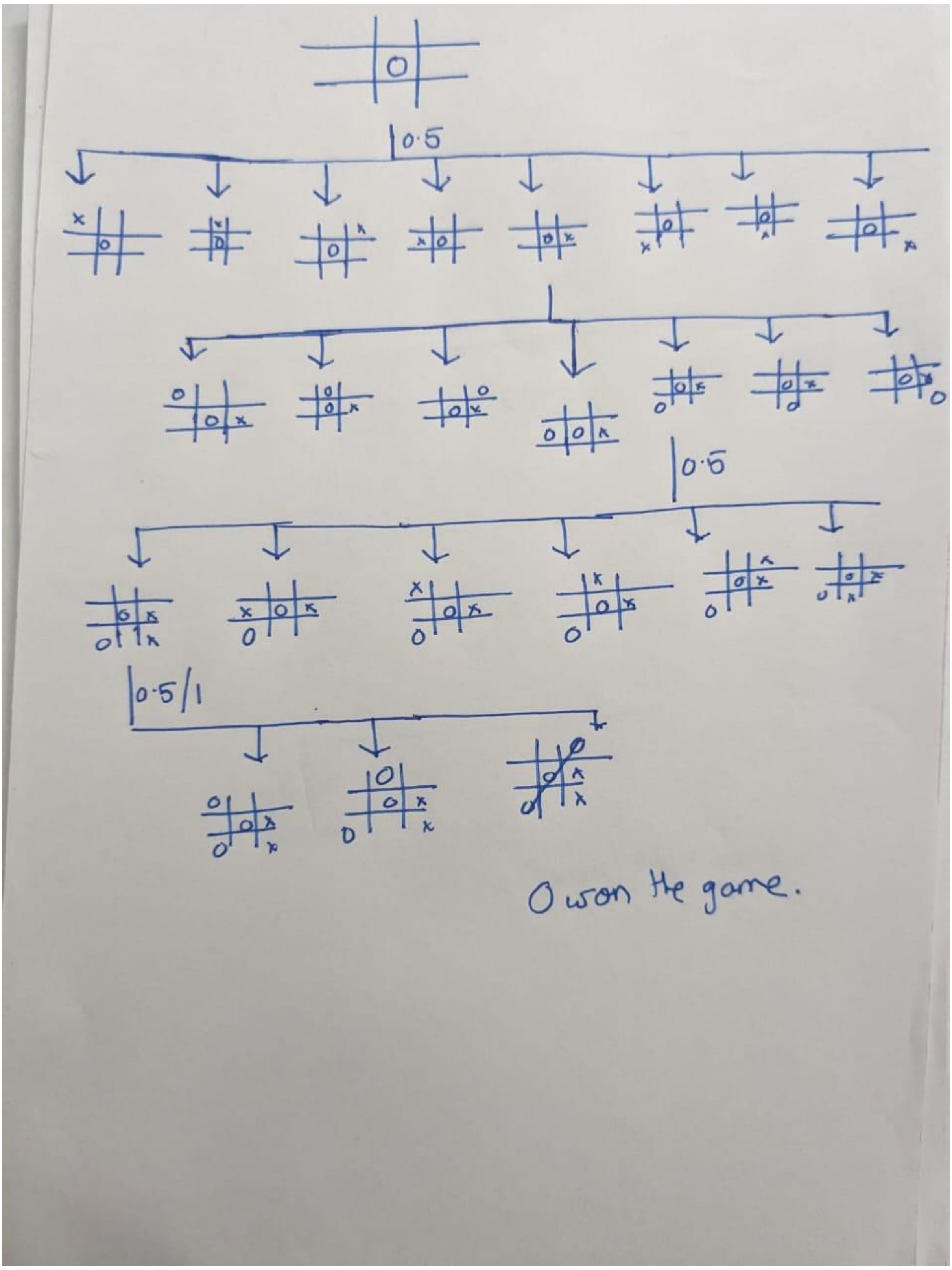


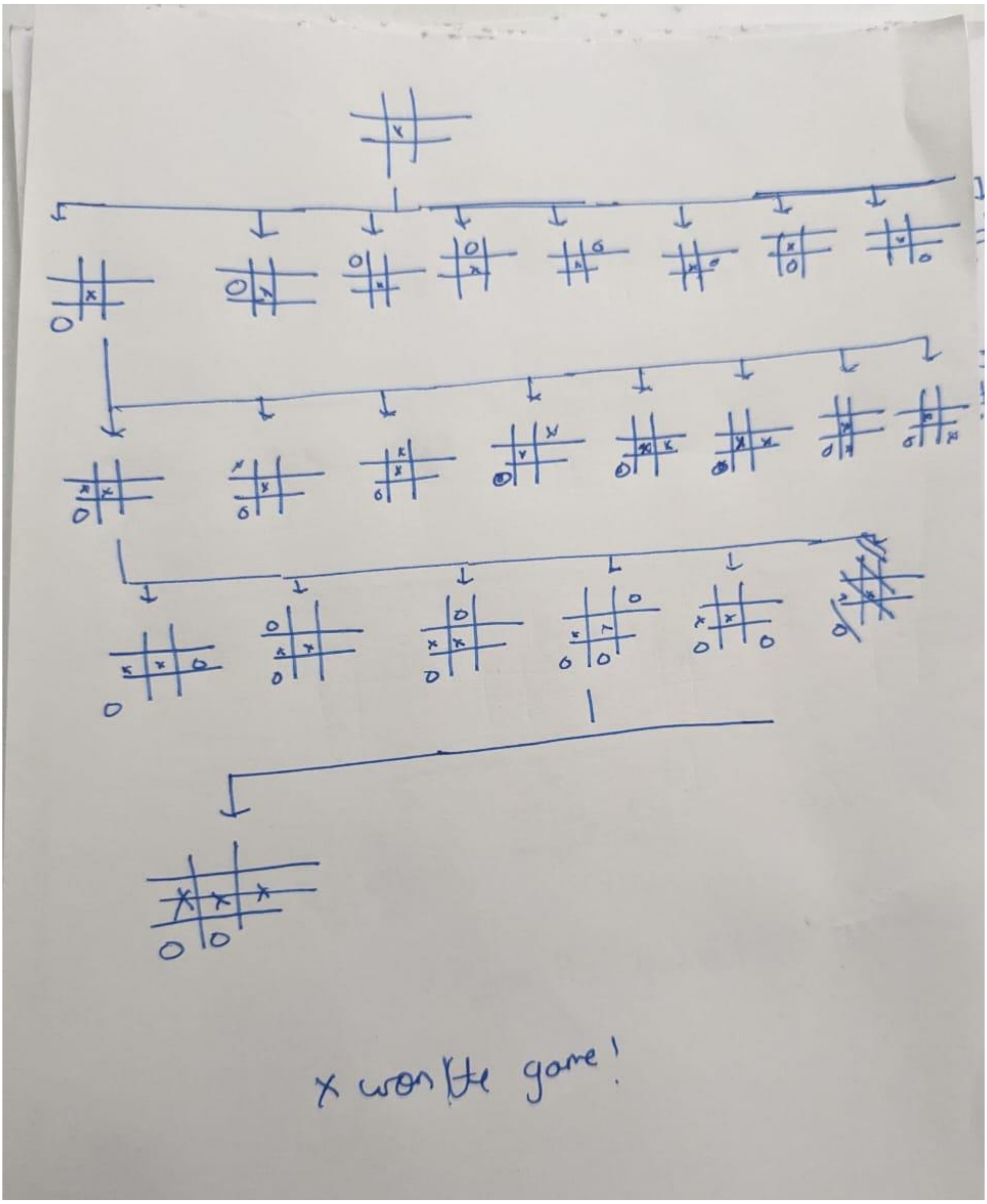
x won the game!

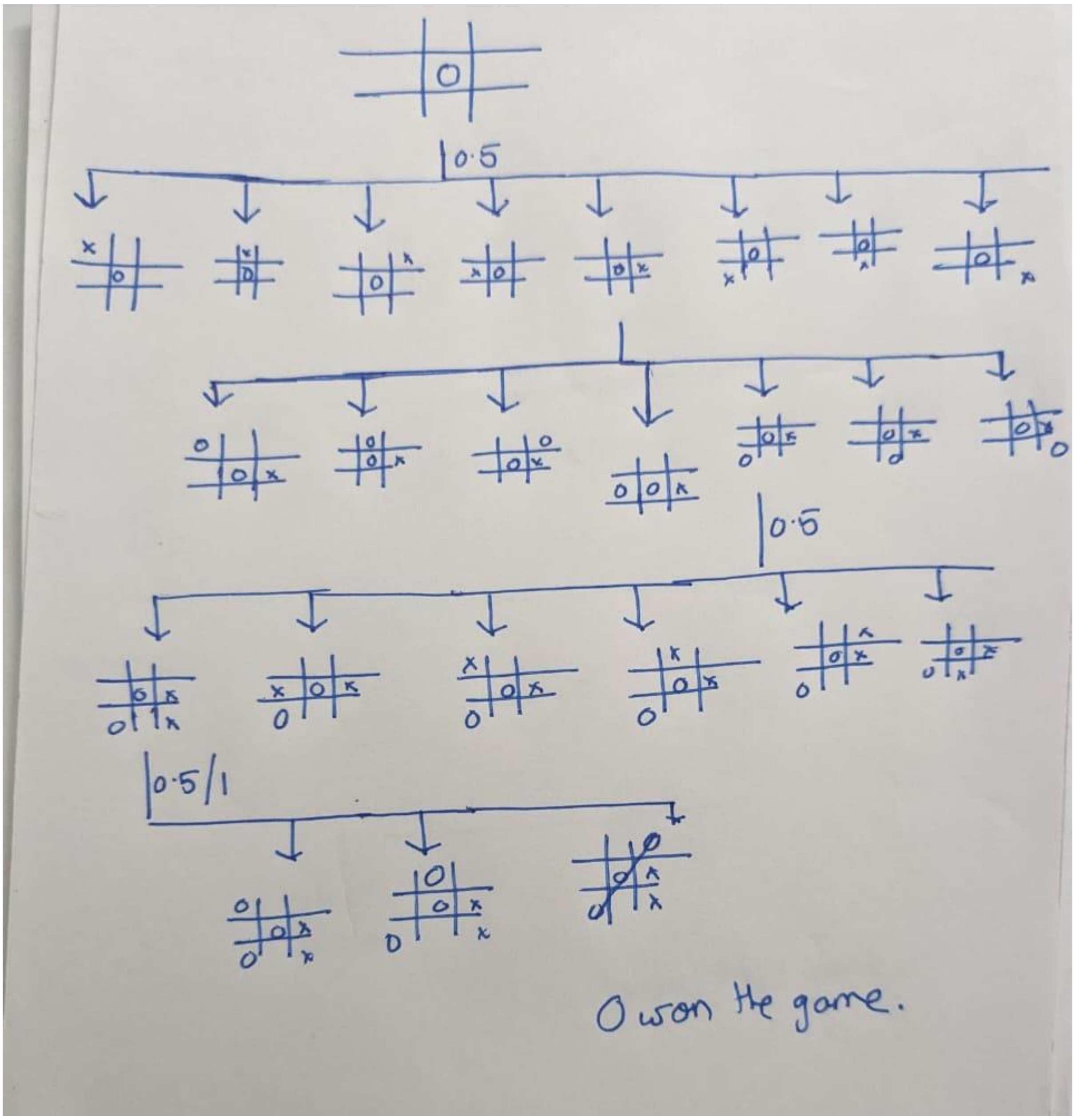


X won the game!


0.5
0.5
O won the game !


0.5
0.5
0.5/1
O won the game.


x won the game!


0·5
0·5
0·5/1
O won the game.

Question 2. Visit the Gymnasium webpage given in the link below. Gymnasium is an API standard for reinforcement learning with a diverse collection of reference environments developed by OpenAI. Answer the following questions.

a) Name all the classic environment available in Gymnasium.

Answer : The classic control environments in Gymnasium are :

1. Acrobot - A two-link pendulum, where the goal is to swing the free end of the pendulum above a certain height by applying torques on the actuated joint.
2. CartPole - A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The goal is to balance the pole by applying forces in the left and right direction on the cart.
3. Mountain Car - A car is placed at the bottom of a valley and the goal is to reach the top of a hill on the right by applying accelerations to the car.
4. Continuous Mountain Car - Similar to the Mountain Car but with a continuous range of possible accelerations that can be applied to the car.
5. Pendulum - An inverted pendulum that starts in a random position, and the goal is to swing it into an upright position by applying torques.

b) Describe the action space, rewards, and terminal condition for every episode for the following environments

i. Cartpole

Answer :

- Action Space: The action is a ndarray with shape (1,) which can take values {0, 1} indicating the direction of the fixed force the cart is pushed with. 0: Push cart to the left, 1: Push cart to the right.
- Rewards: A reward of +1 for every step taken, including the termination step, is allotted. The threshold for rewards is 500 for v1 and 200 for v0.
- Terminal Condition: The episode ends if any one of the following occurs: Pole Angle is greater than ±12°, Cart Position is greater than ±2.4 (centre of the cart reaches the edge of the display), Episode length is greater than 500 (200 for v0).

ii. Mountain Car Continuous

Answer :

- Action Space: The action is a ndarray with shape (1,), representing the directional force applied on the car. The action is clipped in the range [-1,1] and multiplied by a power of 0.0015.
- Rewards: A negative reward of -0.1 * action^2 is received at each timestep to penalise for taking actions of large magnitude. If the mountain car reaches the goal, then a positive reward of +100 is added to the negative reward for that timestep.
- Terminal Condition: The episode ends if either of the following happens: The position of the car is greater than or equal to 0.45 (the goal position on top of the right hill), The length of the episode is 999.

c) In all the environment give in Gymnasium there is an Observation Space. What is significance of it?

Answer : The Observation Space in Gymnasium environments is crucial as it describes the format of valid observations. It clearly defines what observations will look like and how to interact with environments3. The observation space represents the information that the agent observes from the environment at each step, which is used to make decisions about the next action. This could include things like the current state of the game, the position of the agent, the status of certain variables, etc. The observation space can vary greatly depending on the specific environment and task.

Question 3. Identify the problem domain and a brief problem statement that you wish to implement as your course project. Will Gymnasium environments can be used for that? If yes, name the environment.

Project Title : Supply Chain Management

Problem Statement : The project is designed to manage a sophisticated setting that includes a factory and several warehouses. Its goal is to maximize profits by addressing various factors such as geographical distribution, transportation challenges, fluctuations in demand throughout the year, and production expenses.

We can use the GYM frameworks to create a suitable environment, as far as explored no exisiting GYM environments will be suitable.