Consider the following maze:

A screenshot of a test

AI-generated content may be incorrect.

You want the computer to find a feasible route from A to G (Goal), by making steps of 1 in a

certain direction (North, East, South, West) using Q-learning. You cannot go through blocked

boxes (B). The boxes marked g are suboptimal boxes. The rewards obtained in a step are the

following: step: -1, Block (B): -5, suboptimal goal (g): 50, Goal (G): 200. Use a discount factor

α = 0.98 and try out two learning rates γ = 0.5 and 0.9. Maximum number of steps = 3000 in

one run. If you reach the Goal G in a run, you may restart in A in the same run and continue

with the Q matrix obtained so far to get a better approximation. Actions are N, E,S, W.

Transition probabilities are: 0.7 in the chosen direction and 0.1 in any of the 3 others

directions. Trying to walk outside the maze or into a block keeps you in the state you are in.

The challenge for you is to determine a learning rate, a balance between optimizing and

exploration and a policy for determining which direction to go. You may use the epsilon

greedy policy, but you may also define another policy. You are evaluated on the total

objective function obtained and the number of steps needed to reach the Goal. As

benchmark you should use the policy which either randomly chooses a direction or uses an

optimal action given the value vector so far, each with probability 0.5. Repeat each

experiment with fixed settings 10 times (starting with a reset Q matrix) to get more reliable

results. In total you do not need to do more than 5 different parameter settings, but more is

allowed. Give also a figure with the policy solution per setting (i.e. in the maze indicate the

action per block; if the 10 runs indicate several actions take the action which is advised most

often).