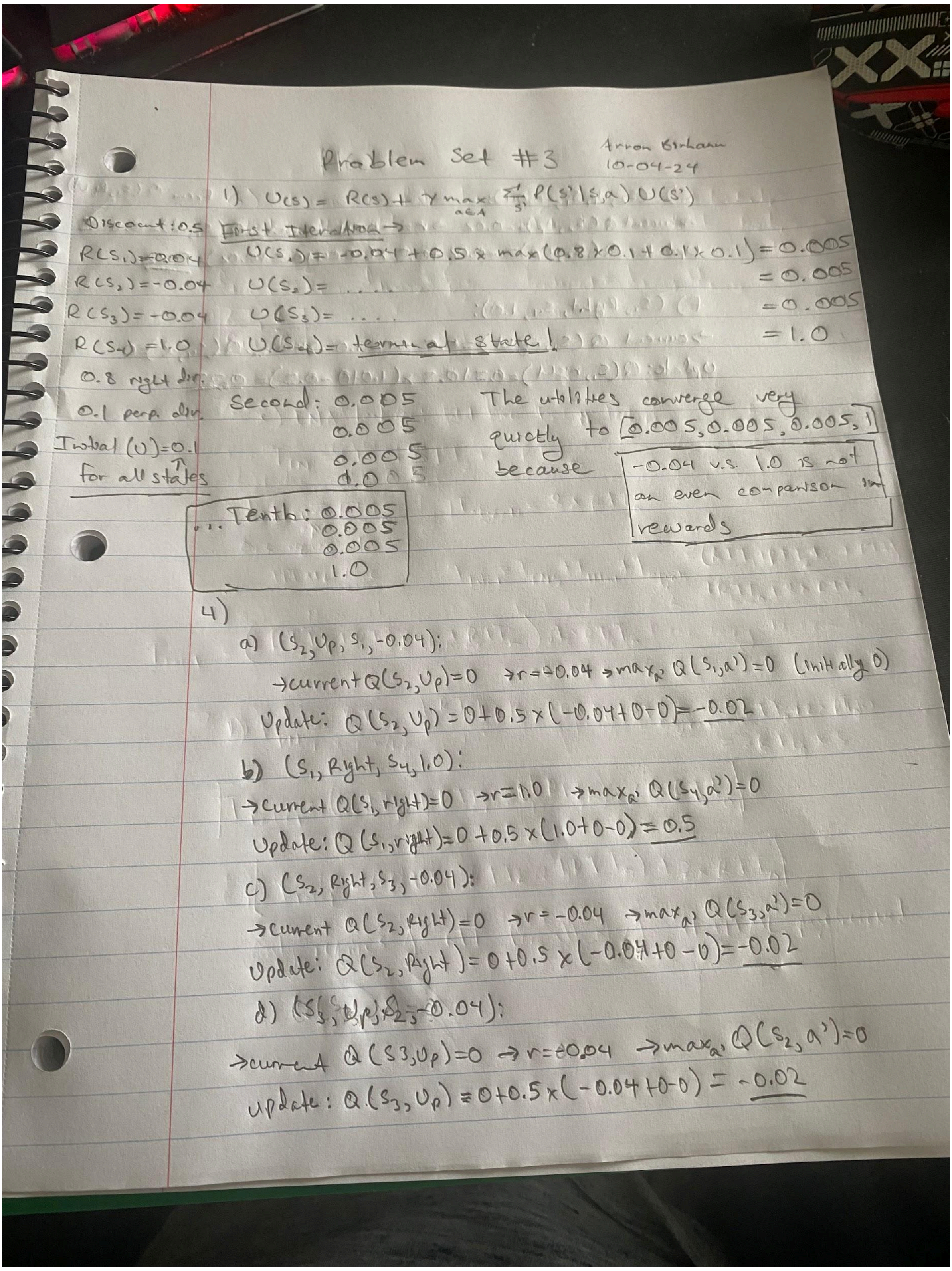
Question 2:
https://drive.google.com/file/d/14Ef8iLV-Te7U8sx_8tPWtPRQjLCb1AYg/view?usp=sharing

Question 3:
https://drive.google.com/file/d/1SLG-GVXTrMF3K6l3ZsJrGz8py_XxjlWn/view?usp=sharing

Question 1 & 4:



Problem Set #3

Arron Birhanu
10-04-24

Discount: 0.5
$R(S_1) = -0.04$
$R(S_2) = -0.04$
$R(S_3) = -0.04$
$R(S_4) = 1.0$
0.8 right dir.
0.1 perp. dir.
Initial (U) = 0.1
for all states

1) $U(s) = R(s) + \gamma \max_{a \in A} \sum_{s'} P(s'|s,a)\, U(s')$

First Iteration →

$U(S_1) = -0.04 + 0.5 \times \max(0.8 \times 0.1 + 0.1 \times 0.1) = 0.005$

$U(S_2) = \ldots = 0.005$

$U(S_3) = \ldots = 0.005$

$U(S_4) =$ terminal state! $= 1.0$

Second: 0.005
0.005
0.005
0.005

... Tenth: 0.005
0.005
0.005
1.0

The utilities converge very quickly to [0.005, 0.005, 0.005, 1] because

-0.04 v.s. 1.0 is not an even comparison in rewards

4)

a) $(S_2, Up, S_1, -0.04)$:

→ current $Q(S_2, Up) = 0$ → $r = -0.04$ → $\max_{a'} Q(S_1, a') = 0$ (initially 0)

Update: $Q(S_2, Up) = 0 + 0.5 \times (-0.04 + 0 - 0) = \underline{-0.02}$

b) $(S_1, Right, S_4, 1.0)$:

→ current $Q(S_1, right) = 0$ → $r = 1.0$ → $\max_{a'} Q(S_4, a') = 0$

Update: $Q(S_1, right) = 0 + 0.5 \times (1.0 + 0 - 0) = \underline{0.5}$

c) $(S_2, Right, S_3, -0.04)$:

→ current $Q(S_2, Right) = 0$ → $r = -0.04$ → $\max_{a'} Q(S_3, a') = 0$

Update: $Q(S_2, Right) = 0 + 0.5 \times (-0.04 + 0 - 0) = \underline{-0.02}$

d) $(S_3, Up, S_2, -0.04)$:

→ current $Q(S3, Up) = 0$ → $r = -0.04$ → $\max_{a'} Q(S_2, a') = 0$

update: $Q(S_3, Up) = 0 + 0.5 \times (-0.04 + 0 - 0) = \underline{-0.02}$

From $Q(S_1, \text{Right})$

e) $(S_2, \text{Up}, S_1, -0.04)$:

$\rightarrow$ current $Q(S_2, \text{Up}) = -0.02 \rightarrow r = -0.04 \rightarrow \max_{a'} Q(S_1, a') = 0.5$

Update: $Q(S_2, \text{Up}) = -0.02 + 0.5 \times (-0.04 + 0.5 + 0.02) = \underline{0.22}$

f) $(S_1, \text{Right}, S_4, 1.0)$:

$\rightarrow$ current $Q(S_1, \text{right}) = 0.5 \rightarrow r = 1.0 \rightarrow \max_{a'} Q(S_4, a') = 0$

Update: $Q(S_1, \text{right}) = 0.5 + 0.5 \times (1.0 + 0 - 0.5) = \underline{0.75}$

Question 5:
The experience traces suggest that the agent is acting greedily. In each state, it consistently selects the action with the highest Q-value, favoring immediate rewards over exploring other potential options. This behavior matches that of a greedy agent, which always exploits its current knowledge to choose the action with the best immediate outcome.

If the world resets after visiting s4, the agent starts again in s2 and likely repeats this behavior. Given that the agent repeatedly chooses actions that maximize its immediate Q-values (e.g., "Up" in s2 and "Right" in s1), this indicates that the agent is likely following a greedy policy.

Question 6:
If a greedy agent is used to generate experience traces for Q-learning, **we are not guaranteed** to visit every state in the environment (in the limit) since it always chooses the highest Q-value in each state, based on its current knowledge.
The single aspect of the environment that could be changed to guarantee visiting every state is introducing exploration into the agent's behavior. This could be done by changing the agent from purely greedy to using an **e-greedy policy**, where with a small probability e, the agent chooses a random action instead of the greedy one. This encourages exploration, ensuring that the agent has a chance to visit all states over time.