PROJECT 2 REFLECTION QUESTIONS

Answer 1

- Decreasing the noise parameter results in crossing the bridge. That is because removing noise removes the possibility of taking an unintended action making our actions deterministic and our agent makes deterministic decisions towards positive rewards, and avoids falling into the cliff which is negative reward.

ANSWER 2

3a): A low discount value indicates that the agent values immediate rewards making the agent favor closer exit. Without any noise, the environment is deterministic, so it can confidently risk walking near the cliff to get to the close exit. The negative reward allows the agent to finish the episode earlier and favor a closer exit.

3b) A preference for immediate rewards, driving the agent towards the closer exit so low discount. Noise: there's a 20% chance the agent's actions might not result in the intended outcome. so the agent is more likely to take a safer route to avoid the cliff. A negative reward for the agent to finish the episode sooner than later.

3c:

- The high discount value indicates the agent values future rewards (+10) more than immediate ones.
- Noise (0.0): The deterministic environment means the agent knows exactly what will happen when it takes an action. It's willing to risk the cliff to reach the distant exit with a higher reward.
- Living Reward (-0.2): The agent is motivated to finish the episode through negative living reward

3d:

- The agent still values future rewards significantly, so it prefers the distant exit with the higher reward.
- With high noise that actions might have unintended consequences, the cliff becomes a significant hazard. The agent will be more cautious and likely choose a path that avoids the cliff. A negative living reward allows episode to end sooner than later

3e)

- The agent values future rewards equally to immediate ones, willing to continue the episode indefinitely if there's a possibility of earning more rewards in the future.

- A positive living reward means the agent gains value just by continuing to exist in the episode. It's incentivized to avoid both exits and the cliff, essentially wandering around without ever terminating the episode.

Answer 3:

Changing the epsilon affects the exploration-exploitation trade-off. Meanwhile, adjusting the learning rate determines how quickly our Q-values update:. Given these constraints of the BridgeGrid environment, and the **limited 50 episodes**, ensuring a >99% probability of learning the optimal policy is challenging. It is likely that the agent cannot sufficiently explore and solidify the optimal policy within this timeframe regardless of the epsilon and learning rate values, making the task "NOT POSSIBLE".