Lab2 Report (2024)

Priyansh Gupta (prigu857)

December 14, 2024

Trade-off Between Exploration and Exploitation

1. Evaluate and discuss the effects of ϵ on performance. What strategy for updating ϵ did you use?

ϵ controls the exploration and exploitation tradeoff: A high value of ϵ favors exploration, where the agent tries random action to learn the environment. Whereas, low value of ϵ favors the exploitation, where the agent used learned Q values to choose the best action possible.

In learning the frozen-lake experiment, Initially, we use the high ϵ of 0.95, which allows the agent to explore a wide variety of state-action pair. Thus, reducing the risk of getting stuck at low maxima early-on. The strategy of linearly decay is used to decay ϵ at a fixed rate of 0.001 per episode until it reaches the minimum value of 0.001. This will gradually shift the agent towards exploitation, ensuring the agent capitalizes on learned knowledge. The plot of accumulated reward shows that the agent consistently improves and learn to achieve higher rewards over time.

3. What are the major difficulties for learning in this environment? Include a discussion.

Major difficulties in learning the Frozen-lake environments are:

1. **Stochastic transitions:** In this environment, transition between states are probabilistic rather than deterministic. The agent might end up in unintended states due to slippery ice, even if it chooses the optimal actions. This makes the agent's learning process harder in order to associate the actions with rewards accurately.

2. **Lack of intermediate rewards:** Rewards are only given when the agent reaches the goal. The lack of intermediate rewards makes it difficult for the agent to evaluates the states-action pairs and adjust its behavior. A potential solution can be giving the agent intermediate rewards for avoiding the holes and decreasing the distance from goal.

3. **Exploration vs. Exploitation tradeoff:** Too much exploration leads to wasting time on suboptimal actions. However, too little exploration risks the agent being unable to find the optimal path. Techniques like Linear decay, exponential decay or upper confidence bound(UCB) can be used in order to balance exploration and exploitation.

While Q-learning works well in our case of small-action space. However, to address these issues more advance algorithms, such as advance Q-learning with UCB will be required, especially in larger boards where high number of holes increases the problem's complexity.

2. Include a plot of your accumulated reward for your best result.

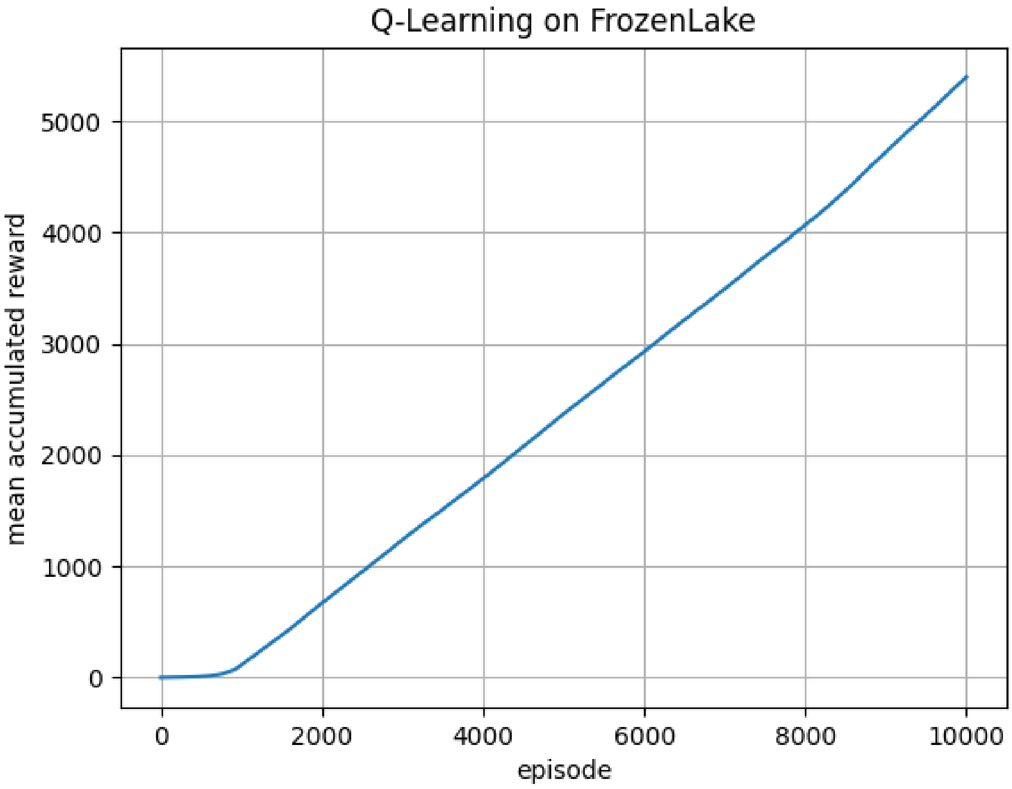

Figure 1: Mean accumulated reward vs. Episods

Competitive Multi-Agent Deep Reinforcement Learning

1. Describe the reward system you designed.

For an agent and adversary, reward systems looks as follows:

- Reward between 0 and 1.0 is given based on the agent's distance from the puck.
- Reward of 0.3 is given if the agent's distance from the puck is less than that of the opponent (puck possession). However, a penalty of 0.2 is also given if it's the opposite case (loosing puck possession).
- A penalty of 1.0 is given if the agent goes out of playing area.
- A reward of 10 is given for scoring a goal.
- A penalty of 5.0 is given for conceding a goal.
- A Reward is given if the agent reduces the puck's distance from goal, based on the difference of the previous and current distance.

2. How well do the agents perform after training? Discuss your results and relate them to your reward system design.

Based on my observation, the agents perform well, They try to be as close possible to the puck and try to minimize the distance of the puck from the respective goal post. This behavior is clearly seen after training as both the agent try to defend and at the same try to push the puck towards their goal. This is related to the reward system design as well. As both the agent and adversary will get a reward of 0.3 for having the possession of puck. on the other side, a penalty of 0.2 is given, if they lose the possession of the puck. The behavior of decreasing the distance of puck from goal induces a competitive nature, where both agents try to push the puck towards their respective goals.

However, sometimes the agent gets out of bound, For example, they might leave the boundary to get the possession of the puck. We will give them a penalty of 1.0, But as the penalty is not that big, agents try to have the possession of the puck. One more behavior, which seems unwanted in the context of the game, is agents trying to fight for the puck even if a goal is scored. That might be because the penalty/reward is based on the puck position, the agent will continuously get a penalty if he concedes a goal. so it tries to take puck possession and take the puck to score a goal. This can be seen in the attached demo.

3. Are the agents equally good? Can you see any reason/explanation why they would not be?

Yes, the agents are equally good. Since the same reward function is used for both the agent, they both try to take the possession of the puck, which induce the competitiveness between them.

4. How do you think the length of episodes and size of the hockey rink would affect learning for your choice of reward system?

If the size of the hockey rink is large, then the agents will require more time/episodes to learn the best moves for each state, leading to a need for larger episode length to ensure convergence.

On the contrary, if the size of the rink is smaller than the current one, agents can explore the states/action pair much more quickly, leading to earlier convergence. Thus, fewer episodes would ve needed in such cases.

Link to demo of agents playing against each other: Simple Hockey Model