ANALYSIS

1. INTRODUCTION:

Flappy Bird is one the most popular mobile games. Q-Learning is form a reinforcement learning that does not require the agents to have prior knowledge of the environment dynamics. This project aims to make the bird learn by making mistakes and score higher by using Q-Learning to learn the best action to be performed at the right position.

2. IMPLEMENTATION STRATEGIES:

Computing the Q value for Q Learning Algorithm is:

$$\underbrace{Q_t(s_t,a_t)}_{\text{old value}} + \underbrace{\alpha_t(s_t,a_t)}_{\text{learning rate}} \times \Big[\overbrace{\underbrace{R_{t+1}}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \underbrace{\max_a Q_t(s_{t+1},a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} - \underbrace{Q_t(s_t,a_t)}_{\text{old value}} \Big]$$

i. State Representation:

The dimension of the State space is considered as the horizontal distance and vertical distance from the next lower pipe i.e. the X distance and Y distance from the next pipe.

ii. Actions:

There are only two actions: jump and stay

iii. Learning Rate:

I tested with three variations of Learning Rate:

a. 1/N+1: where is the number counter at that state. The number of times the bird reaches that set. This approach showed that the bird learnt very slow. Even after training the bird overnight, the performance was not improved to a greater extent.

b. 1/sqrt(N+1): In this approach the bird learnt faster compared to the first approach but after more iterations, the bird's performance tends to drop. As in first on running for 2000 games, this approach towards the end gives a stable output but after running it again. It drops.

c. Constant(0.7) : In this approach the learning rate is set to a constant value of 0.7 and it is observed that the learning is carried out constantly not giving the performance as mentioned in the above two cases.

iv. Reward:

I tried three sets of reward:

a. Reward: 15 and Reward_dead=-1000 : Here the bird learning is comparatively consistent. Based on the paper it is seen that this reward assignment is supposed to have been performed better.
b. Reward:10 and Reward_dead=-1000 : Not much difference. The performance was found to be similar to the above combination.
c. Reward:15 and Reward_dead = -100 : Here the bird crosses the first few pipes quickly but after that it drops.

v. Policy:

The policy factor is used to address the exploration. I have used a factor to address this. Initially while training the bird, I allow the bird to explore at the rate of 0.6. That is for the first 3000 trials the bird has an exploration rate of 0.6. Then after the bird is trained I allow the bird to only use the Q values and not try exploration at all.

vi. Discount Factor:

The change in the discount factor didn't effect the bird's learning much though. But if the discount rate is below 0.7 the performance was not found to be good though. The idle would either be 0.8 or 0.91.

3. TESTING DETAILS:

a. Bird Flying through the pipe of the same height:

The bird takes time to train itself but once the bird is trained then it flies continuously for that height. I left it out for a few hours and it ran till 1300 games and I shut it down. Now on running the training data set. It does fly for the same height. In an hour it ran for approximately 650 pipes.

b. Bird Flying through Random pipes:

So for this set of pipes it took lots of time to train i.e. around 6 hours. After the training, within a span of an hour the bird reaches 212 pipes.