EE 399 Final Report

Author: Ewan Lister

In training the agent, from previous work on the task, I established that an epsilon close to 0.04 and gamma between 7 and 5 are optimal for training the snake quickly. For rewards, the following paramaters were experimentally determined: [-100, -32, 1, 10]. With -100 as the reward for death, -32 the reward for moving inefficiently, 1 for moving efficiently, and 10 for eating an apple. Now, the goal is to determine which parameter has less tolerance for error. In this report I will discuss a possible methodology for isolating the more important parameter between gamma and epsilon, apply the methodology, and interpret the results.

One quick method to establish this for many trials in a small amount of time is to train the model only over around 50,000 episodes, and then to take the average score for 100 trials as we modulate one parameter and hold the other fixed. For such a small number of episodes, the agent may not win every time, but the average provides a good metric for the agents potential performance. Shown below is a table containing the average score for the agent over 100 trials, versus the gamma and epsilon values, where epsilon and gamma are altered by 0.5 of their significant digit from their experimentally ideal values.

Modulation with respect to decimal place of significant digit:	-2.0	-1.5	-1.0	-0.5	0.0	+0.5	+1.0	+1.5	+2.0
Epsilon changes, gamma = 0.6	6.7	6.862	6.961	6.714	6.480	6.4	6.60	6.455	6.260
Gamma changes, epsilon = 0.04	6.550	6.359	6.537	6.125	6.480	6.839	1.10	0.310	0.160

It should be noted that only with more trials is the full state and action space of the game available to the agent, so training with more episodes could lead to gamma being more important, for instance, in the agents efforts to avoid its tale. Additionally, in typical reinforcement learning, the epsilon is manually decreased over time as the agent matures. With that being said, the table of results shows that there is greater variance as gamma is varied. Thus the discount factor for future reward is a more important value to consider. If we take the ranges of both parameters, Δreward_gamma = 6.64, while Δreward_epsilon = 0.602, meaning that for their respective orders of magnitude, manipulating gamma resulted in a larger variety of training results, which one could link to greater significance for this parameter.

This conclusion that gamma is more significant makes sense given the game requires predictive behavior to avoid walls and the agents tail in the future. As a side note, around 5 minutes of training time caused the model to win consistently.