Homework 3 Solutions

1. There are several combinations of the parameters that are acceptable. $\alpha = 0.4$, $\gamma = 0.95$, $\epsilon = 0.04$ worked pretty well but since it depends a lot on implementation, we will accept numbers if the justification looks right. For the number of bounces, students must ensure they get at least 9 bounces averaged over a certain number of games that they choose themselves.

2. Derived from the question above

3. Training time should increase. Since the process will start from various random positions in the start state, the agent will explore more despite following an epsilon greedy approach. For example, if the ball starts near to the right wall, the game will end pretty much immediately if the paddle is already not there. Such sequences of games would lead to a longer training time and the agent would be wasting time looking at a lot of the ‘bad’ places in the grid.

4a. The finite difference Pong environment to be controlled by Q Learning contains far more information than is captured in the discretized Q Learning states. This gives rise to systematic behavior that cannot be correctly modeled by an MDP. The formal properties of Q Learning are based upon the assumption that there is an (unknown) MDP underlying the environment. These properties of Q Learning, including polynomial time convergence to near optimality, are no longer guaranteed. Optimal play for our Pong environment achieves an unlimited number of bounces. This level of play is not observed even after much training.

4b. If the environment were perfectly captured by an MDP, we would expect Q Learning to continue improving, converging to optimal play.

5. There is a fundamental tension in reinforcement learning between learning about the environment (exploration) and employing that knowledge to collect higher rewards faster (exploitation). Exploration tries to discover new states and to improve its estimation of what happens when actions are executed in them. Exploitation revisits known states and performs the (currently believed) greedy action. A higher epsilon favors exploration over exploitation; a lower epsilon favors exploitation over exploration. By concentrating only on exploration, Q1 learns the fastest about the environment. When Q1 is evaluated (with its epsilon set to zero) it will likely have the most accurate knowledge of the environment to exploit. Conversely, Q4 spends most of its time trying to exploit knowledge, even at the very beginning when it has no real knowledge at all. Since Q4 explores little, it learns about its environment much more slowly. In evaluation, it will likely achieve the level of Q1 only much later. Q2 and Q3 are between these extremes, Q2 closer to Q1 and Q3 closer to Q4. The correct answer is (d).