

Question 3:

All Q-learning runs have been conducted using the ϵ -greedy approach. For all experiments the following parameters are constant:

Learning Rate, $\alpha = 0.1$

Discount Rate, $\gamma = 0.9$

Sampled Episodes = 1,000

Initial experiment has been conducted with $\epsilon = 0$ i.e. the next step in each episode is chosen greedily based on expected return the move yields. $\epsilon = 0$ implies random exploration is done only in case multiple moves yield the same expected return.

The second experiment has been conducted with $\epsilon = 0.1$. This implies that 10% of the time the greedy decision based on expected return of a move is not followed and a move other than the greedy decision is randomly selected. This noise introduced in the greedy approach induces random exploration of the neighboring cells in the grid.

The learning curve for both experiments suggests that although both approaches converge the range of convergence is broader in case of the experiment with induced noise in the decision making. The optimal solution for a 15x15 grid is 28 steps.

In case of the pure greedy approach the learning algorithm converges in the range of 28-50`ish never settling on 28 since random stepping kicks in every time the Q-values for multiple moves converges on the optimal and mutually equal values. Steadily the new Q-values backup the grid and converge to the optimal relative intervals yielding the optimal value yet again. This behavior is nearly periodic but does not exhibit symmetry due to random stepping.

In case of the noise-induced greedy approach the algorithm does converge but within a slightly broader range of 28-70`ish. This is due to noise which induces random movement on the grid irrespective of the optimal move indicated by the corresponding Q-values. Also we notice that this randomness causes the converging learning curve to spend lesser time around the optimal "28 steps" value and explore more than the vanilla greedy approach. This increased randomness causes a larger time-span to pass before the optimal relative intervals amongst Q-values are established.