Problem 2

1. higher values of epsilon reach convergence faster, but lower values of epsilon converge on higher optimality over longer timesteps. For epsilon = 0, its asymptote is significantly lower as exploration does not occur, taking the best value at each step with no exploration

2. It appears as observed reward variance increases, the noise from this change weakens convergence. Even if a bandit has an optimal arm that it samples, higher variance reduces the consistency of that arm.

   when changing value variance, it appears to have the same effect of modifying the observed reward variance. More generally, higher variance will result in more noise/fluctuation in the reward, resulting in a weaker convergence on the optimal average reward

3. When considering varying values of # of arms and epsilon, they appear to both affect convergence similarly. For very low values (k=1, or epsilon=0), options are limited, and convergence occurs at sub-optimal values, increasing both will allow for better exploration/exploitation. For k specifically, it appears higher values results in more optimal rewards with similar convergence for all values of k that are higher. This differs from epsilon where lower values (not 0), will converge slower, but on more optimal average rewards over more time steps.

4. If we set our q_init to be "greedy", or of higher initial values, it encourages the exploitation of other options. So, in the case when epsilon=0, setting a value higher than the mean and variance by some significant amount, we can encourage the exploitation of other arms when the true value set for a given arm will likely be lower than the optimistic value initially set. So, if we have:

   q_init = [3,3,3,3,3] with e=0, the bandit will choose the max. Once chosen, the q values will look something like:

   q = [0.5,3,3,3,3].

   So, with no exploration, the bandit will be encouraged to choose another arm, where values are much higher than the observed reward thus far.

Problem 3

1. As epsilon increases, the optimality is reached faster, but perhaps not as optimal with lower, but not e=0, values. Lower values of epsilon approach optimality slower, but at higher values over longer time steps. Convergence to q_*(a) appears to not be possible with e=0, since exploration is not encouraged, meaning only the greedy value will be chosen at each time step.

2. It depends. If the solution requires Q_t(a) to roughly equal Q_*(a), then no. However, after a certain number of timesteps, the convergence of Q_t(a) to Q_*(a) will be met due to law of large numbers for values of e > 0, assuming stationarity. If some threshold of sub-optimality can be determined, exploration should stop once within an acceptance threshold. In experiments conducted, it shows around timestep 400 the gain achieved from higher exploration (where epsilon is higher) tends to have diminishing returns. So, exploration could be decreased in relation to the increase of average reward over time.

3. Exploration would need to be increased to reflect potential changes in the environment. Say an agent found an arm that produced the highest average reward, other arms in a non-stationary environment may become more advantageous to exploit in future timesteps. Understanding these changes would require more exploration of the environment.

4. Sequentiality would require understanding of the relation between the state of the arms and how the actions change that state. Say a certain arm changes the subsequent values to either be better or worse, exploitation would be more difficult as just picking the max value may not suffice. Or in the case a certain combination of actions generates better rewards, exploration-exploitation alone may not be able to take advantage of combinations that converge to Q_*(a)