1. Reflection

a. We're finally getting to the most interesting portion of this course (or at least in my opinion). By that I mean we're finally getting to the machine learning aspect of this course where we start to understand how machines learn how to make decisions on their own as opposed to hard coding it ourselves or through algorithms (as we have been doing in the past. The reinforcement learning - introduction was honestly really refreshing to get into. I enjoyed the other methods of learning, such as super vised learning where we feed the agent some labeled data, and it tries to draw the dots, or unsupervised learning when the agent is fed unlabeled data and tries to form clusters (or groups) amongst the data. The idea that separates reinforcement learning from supervised and unsupervised learning is that the agent is constantly interacting with the environment, as opposed to just being fed data by some external source (such as ourselves). In reinforcement learning, the agent is typically trying to maximize some reward, or utility, that is being given to it by the environment every time the agent interacts with it. The very specifics as to how this is achievable were very interesting to me, since I like to learn a lot about machine learning and data science.

b. Expanding this concept of machine learning into how it's applied in the multi-armed bandit made learning a lot better, in my opinion. It's always nice to see how the theory we learn in class can be applied to more practical situations, such as knowing which restaurants to go to or who to hire. It's really what makes learning computer science actually worthwhile. In the multi-armed bandit problem, we have multiple possible actions to take. The goal is to maximize some reward given to us whenever we choose which action to take. However, we don't know beforehand which actions will lead us to the maximum possible reward (on average). As a result, we have to do some exploration in the beginning to better understand our environment. Once we think we are satisfied and have gathered enough information, then we can exploit the information we have collected. In this stage, we're taking (what we believe) is the optimal decision every single time because we believe we have learned enough about the environment. The interesting part here is knowing how to strike a balance between exploration and exploitation. It'd be nice to exploit every single time, but possible that we might get stuck with a non-optimal choice in the beginning. On the contrary, exploration is important to learning about our environment, but at some point we have to exploit more as to make use of our knowledge.

c. The second part of the multi-armed bandit expands more into the mathematical aspects of how the problem works. So far, we've only gone over a more efficient way to calculate the estimated payoff of a given action. I really like the method because not only was it simple, but it was also very familiar since we've done multiple moving averages before.

2. Action-Value Estimation

a. We define a way to simply calculate the average of a set as $Q_n = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i$. All this equation really says is sum up all reward values of R_i , and divide it by $\frac{1}{n-1}$, which should be the number of entries of R_i . However, as n grows, calculating the average will take more time. We can derive a quicker method through the following derivation. Start with the equation $Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i$ (all we did was increment n by 1). Then, we can simplify the equation as such…

   $Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}(R_n + \sum_{i=1}^{n-1} R_i) = \frac{1}{n}(R_n + (n - 1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i)$

   $= \frac{1}{n}(R_n + (n - 1)Q_n) = \frac{1}{n}(R_n + nQ_n - Q_n) = Q_n + \frac{1}{n}(R_n - Q_n)$

   In the end, we get $Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$. All this really says is that the new average is equal to the old average plus some bias $\frac{1}{n}(R_n - Q_n)$. This equation should be familiar, since this is just the equation for the moving average.

b. All that $\alpha_n = \alpha \forall n$ is saying is that the weights will remain constant for all values of α_n . Substituting into our equation, we get $Q_{n+1} = Q_n + \alpha(R_n - Q_n)$. Interpreting this equation, we get that the new average is the old average plus the same bias where this time, the bias is weighted by α . However, since α is consistent for all values of n, we can safely say that the equation is just the weighted average of $R_1, \ldots , R_n$ and Q_1 since the weights for any reward value should be the same as Q_1 . We know that at Q_1 , we are given an optimistic value that we are sure that each arm won't go above. This will encourage exploration and, thus, make learning more efficient for the agent.