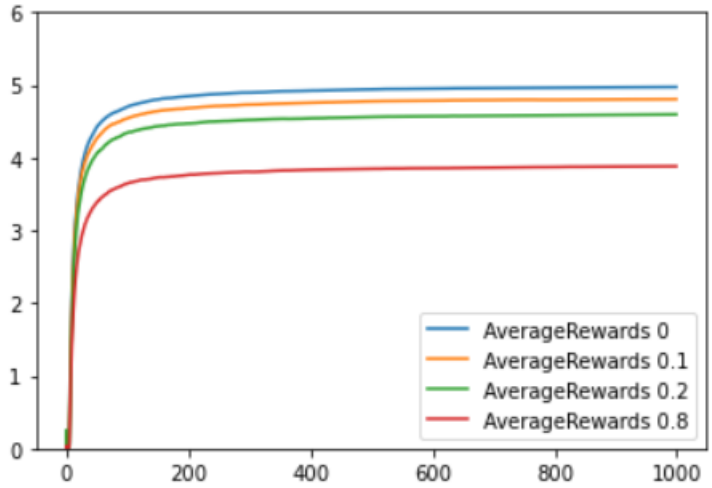
108061217 鍾永桓

Q1:因為本題 action 的 reward 間差距較大，因此基本上只要每個 action 都嘗試後就找到了最佳的 action，幾乎不用探索，若設為更大的 epsilon 0.2 或 0.8，反而花太多次於探索上，選最佳 action 的次數很少，導致最終的 aceragereward 減少。



Q2:如果要在 Epsilon 為 0 時可以得到最佳答案，也就表示過程中完全不會進行 exploration，必須要能確保所選的 action 是最佳的，所以可行的方法應該是在開始選擇前要先純粹進行 exploration，也就是不斷隨機選擇，評估每個 actionvalue，等到從每個 action 取得足夠的資訊能確保其 actionvalue 正確後，再開始進行選擇，此時每次都會選擇 actionvalue 最大者，所以不用在 exploration，epsilon 設為 0 即可，不過這個方法主要就是前面預先花時間 exploration，開始選擇後便不再 exploration。

Q3:因為本題當中的所有 action 的 reward 皆為 normal distribution，可以預期只要執行次數夠多，actionvalue 最終將會非常接近於它們各自的 mean，而隨機選取的機率為 epsilon， 最終應會收斂至 max(mean)*(1-epsilon)+(mean 的總和/action 的數量)*epsilon，因為有 epsilon 的機率會選 mean 最大者，而 1-epsilon 則隨機選取。