2 Multi-Armed Bandits

2.1

In ε-greedy action selection, for the case of two actions and $\varepsilon = 0.5$, what is the possibility that a non-greedy action is selected?

- With probability $(1 - \varepsilon)$, the greedy action is chosen outright. Then, with probability ε, one of the two actions is chosen at random (with equal chance), so the non-greedy action is chosen with probability $\frac{\varepsilon}{2} = 0.25$

2.2 Bandit Example

Consider a k-armed bandit problem wiht $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$ for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of thee time steps the ε case may have occured, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occured?

- On timestep 2 and 5, the greedy actions were 1 and 2 respectively, but different actions were chosen. Here, the ε case definitely occured.
- The ε case may have occured on all other timesteps. Since the ε case can still select the greedy action, it is not possible to rule out that this happened.

2.3

In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answers quantitatively.

- Both the $\varepsilon = 0.1$ and $\varepsilon = 0.01$ cases will eventually converge to the true probabilities. Let $R*$ be the reward when choosing the optimal action every time, and L the expected loss when choosing a suboptimal reward. Then the average reward will converge to $R * -\varepsilon \cdot L$, meaning there will be a difference of $0.09L$ between the two cases.
- In the long run, both algorithms will be (nearly) certain about what the best reward is. The $\varepsilon = 0.1$ agent will select the optimal action in $1 - \varepsilon = 90\%$ of cases. The $\varepsilon = 0.01$ agent will select the optimal action in $1 - \varepsilon = 99\%$ of cases.
- By both metrics, the lower ε is superior in the long run. However, it takes much longer to converge.

2.4

If the step-size parameters, α_n, are not constant, then the estimate Q_n is a weighted average of previously received rewards with a weighting different from that given by (2.6). What is the weighting on each prior reward for the general case, analogous to (2.6), in terms of the sequence of step-size parameters?

- This is fundamentally the same as in (2.6), only the notation comes out less cleanly because we cannot aggregate the different step size parameters into one neat exponent.

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n + Q_n] && (1)\\
&= \alpha_n R_n + (1-\alpha_n)Q_n && (2)\\
&= \alpha_n R_n + (1-\alpha_n)[\alpha_{n-1}R_{n-1} + (1-\alpha_{n-1}Q_{n-1}] && (3)\\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})Q_{n-1} && (4)\\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1}) && (5)\\
&\quad + ... + (1-\alpha_2)\alpha_1 R_1 && (6)\\
&= \sum_{i=1}^{n}\left[\alpha_i \cdot \left(\prod_{k=i+1}^{n}(1-\alpha_k)\right) R_i\right] && (7)
\end{aligned}
$$

2.5 programming

Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for nostationary problems. Use a modified version of the 10-armed testbed in which all the $q_*(a)$ start out equal and then take independent random walks (say by adding a normally distributed increment wiht mean zero and standard deviation 0.01 to all the $q_*(a)$ on each step). Prepare plots like Figure 2.2 for an action-value method using sample averages, incrementally computed, and another action-value method using a constant step-size parameter, $\alpha = 0.1$. Use $\varepsilon = 0.1$ and longer runs, say of 10.000 steps.

- Answer