

## 2 Multi-Armed Bandits

### 2.1

*In  $\varepsilon$ -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  $\varepsilon$ , 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 with  $k = 4$  actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using  $\varepsilon$ -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 these time steps the  $\varepsilon$  case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?*

- On timestep 2 and 5, the greedy actions were 1 and 2 respectively, but different actions were chosen. Here, the  $\varepsilon$  case definitely occurred.
- The  $\varepsilon$  case may have occurred on all other timesteps. Since the  $\varepsilon$  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  $\varepsilon$  is superior in the long run. However, it takes much longer to converge.