

method uses the step-size parameter  $\frac{1}{n}$ . In this book we denote the step-size parameter by  $\alpha$  or, more generally, by  $\alpha_t(a)$ .

Pseudocode for a complete bandit algorithm using incrementally computed sample averages and  $\varepsilon$ -greedy action selection is shown in the box below. The function *bandit*(*a*) is assumed to take an action and return a corresponding reward.

### A simple bandit algorithm

Initialize, for  $a = 1$  to  $k$ :

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$A \leftarrow \begin{cases} \arg \max_a Q(a) & \text{with probability } 1 - \varepsilon \quad (\text{breaking ties randomly}) \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$

$R \leftarrow \text{bandit}(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)} [R - Q(A)]$

## 2.5 Tracking a Nonstationary Problem

The averaging methods discussed so far are appropriate for stationary bandit problems, that is, for bandit problems in which the reward probabilities do not change over time. As noted earlier, we often encounter reinforcement learning problems that are effectively nonstationary. In such cases it makes sense to give more weight to recent rewards than to long-past rewards. One of the most popular ways of doing this is to use a constant step-size parameter. For example, the incremental update rule (2.3) for updating an average  $Q_n$  of the  $n - 1$  past rewards is modified to be

$$Q_{n+1} \doteq Q_n + \alpha [R_n - Q_n], \quad (2.5)$$

where the step-size parameter  $\alpha \in (0, 1]$  is constant. This results in  $Q_{n+1}$  being a weighted average of past rewards and the initial estimate  $Q_1$ :

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha [R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha) Q_n \\ &= \alpha R_n + (1 - \alpha) [\alpha R_{n-1} + (1 - \alpha) Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \\ &\quad \dots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i. \end{aligned} \quad (2.6)$$