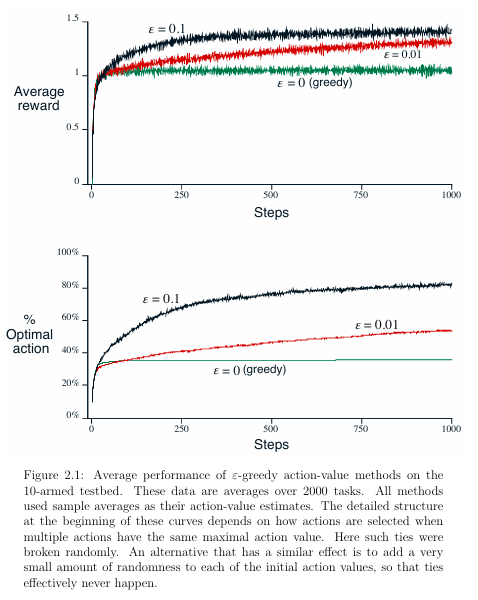
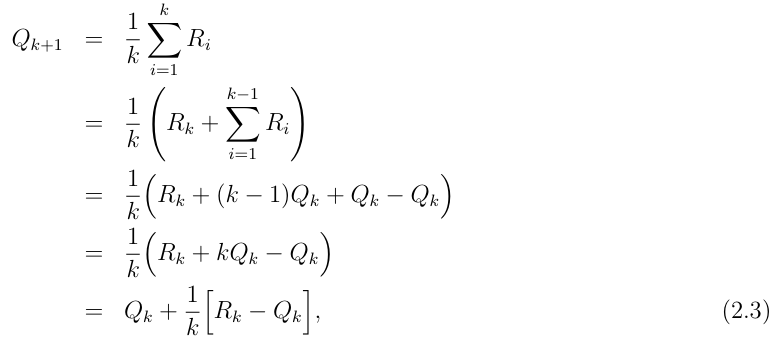
**Chapter 2 Multi-arm Bandits**

2.2 Action-Value Methods



2.3 Incremental Implementation (Tracking a Stationary Problem)

Given this average and a kth reward for the action, Rk, then the average of all k rewards can be computed by:



which holds even for k = 1, obtaining Q2 = R1 for arbitrary Q1. This imple mentation requires memory only for Qk and k, and only the small computation (2.3) for each new reward.

The update rule (2.3) is of a form that occurs frequently throughout this book. The general form is:



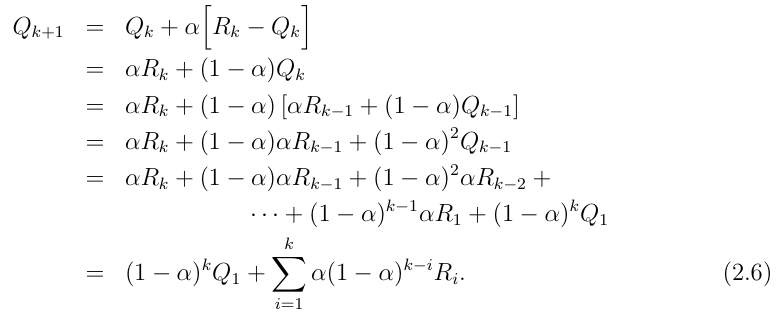
We sometimes use the informal shorthand = 1 k to refer to this case, leaving the dependence of k on the action implicit.

2.4 Tracking a Nonstationary Problem

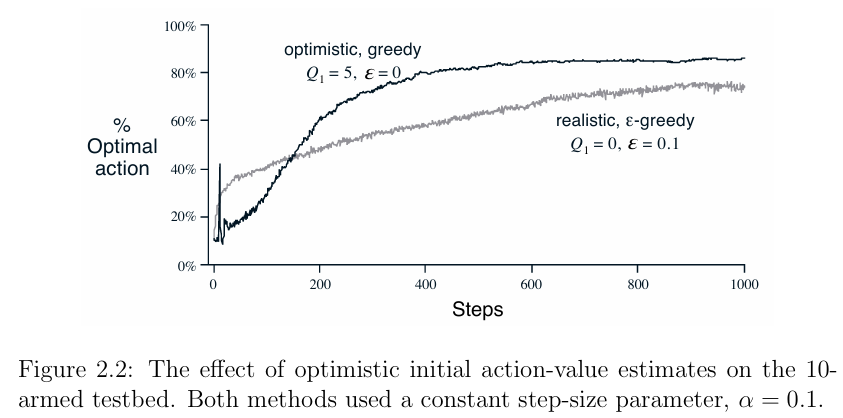
the incremental update rule (2.3) for updating an average Qk of the k 1 past rewards is modified to be:



where the step-size parameter is constant. This results in Qk+1 being a weighted average of past rewards and the initial estimate Q1:



2.5 Optimistic Initial Values



For the sample-average methods, the bias disappears once all actions have been selected at least once, but for methods with constant e , the bias is permanent, though decreasing over time as given by (2.6).

Figure 2.2 shows the performance on the 10-armed bandit testbed of a greedy method using Q1(a) = +5, for all a. For comparison, also shown is an-greedy method with Q1(a) = 0. Initially, the optimistic method performs worse because it explores more, but eventually it performs better because its exploration decreases with time. We call this technique for encouraging exploration optimistic initial values.