Date created: 2025-07-17-Thu

2.6 Optimistic Initial Values

Initial values as prior knowledge

So far, all methods discussed are dependent to some extent on the initial action-value estimates, $Q_1(a)$.

- for sample-average methods, the bias disappears once all actions have been selected at least once,
- for constant α, the bias is permanent and decreases over time.

This is usually not a problem in practice because the initial estimates provide a way to supply prior knowledge about the level of expected rewards.

Initial values as a way to excourage exploration

Suppose we set initial action-value estimates to $+5$ (instead of 0) in the 10-armed testbed. Recall that $q_*(a) \sim N(0,1)$, so $Q_1(a) = 5$ is *wildly* optimistic. But this optimism encourages exploration since whatever actions are initially selected, it will feel disappointing and thus the learner will try other actions. Under this paradigm, even a greedy method does a fair amount of exploration.

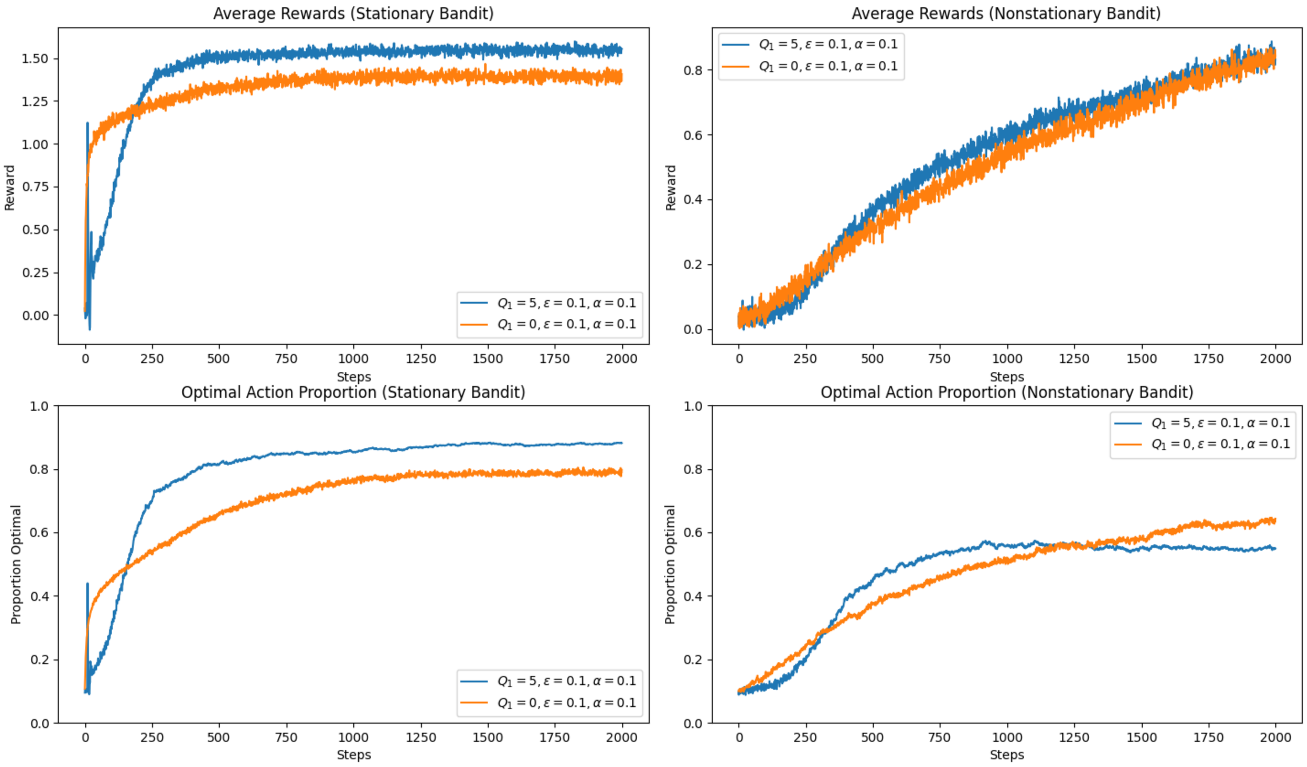

Figure 1: Fig 2.3: Optimistic initial values on 10-armed testbed

As shown, the optimistic initial value method initially does worse for the stationary task because it explores more, but eventually performs better. This method does not perform well on nonstationary problems, though, because the drive to explore is inherently temporary. Indeed, any method that focuses on the initial conditions in any special way is unlikely to help with the general nonstationary case.

Exercises

2.6: Mysterious Spikes

The results shown in Figure 2.3 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?

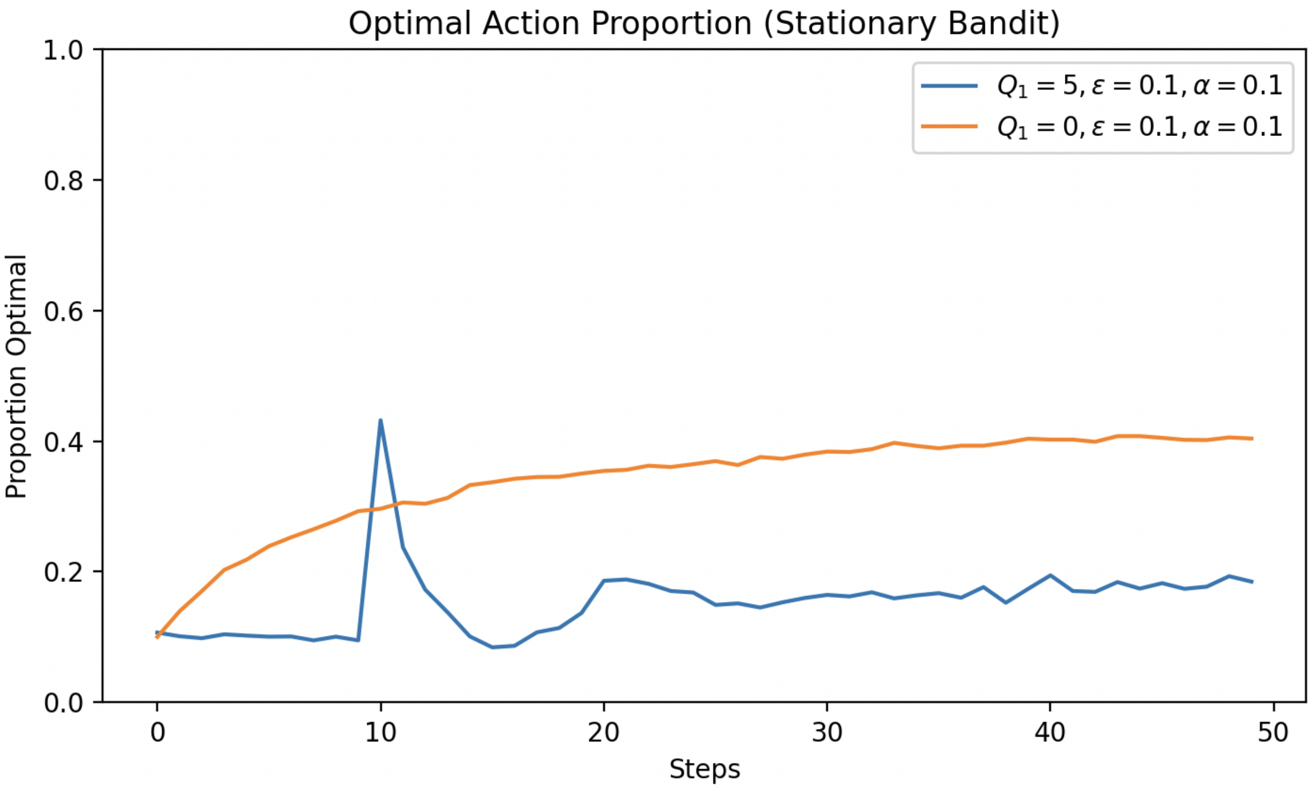

Figure 2: First 50 Steps

Solution As shown in the zoomed in figure, the spikes occur at steps 10 and 20 (and stabilize afterwards). If we were to perform the same experiment for a 5-armed testbed, we see that spikes occur at steps 5 and 10.

This indicates that on average, after k steps, each action has been selected once, and the action value estimates are on average

$$Q_2(a) \sim 5 + N(0, 1).$$

Obviously, $Q_2(a)$ is expected to be highest for the optimal action, and therefore on the kth step, the greedy algorithm is most likely to select the correct optimal action.

A similar argument applies to the $2k$th step, where on average

$$Q_3(a) \sim \frac{5 + 2N(0, 1)}{2}$$

implies that $Q_3(a^*)$ is highest, though with lower probability than on the kth step.

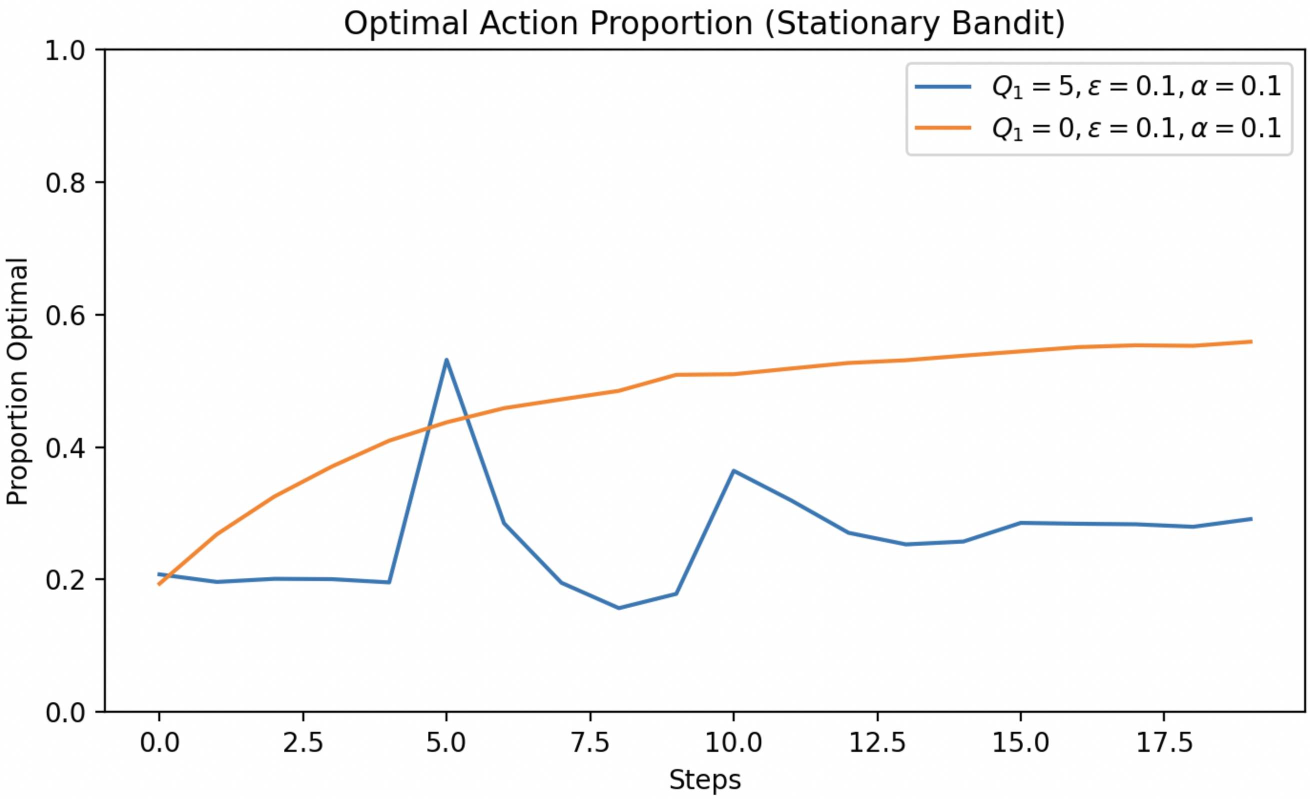

Figure 3: 5-Armed testbed

2.7: Unbiased Constant-Step-Size Trick

In most of this chapter we have used sample averages to estimate action values because sample averages do not produce the initial bias that constant step sizes do (see the analysis leading up to (2.6)). However, sample averages are not a completely satisfactory solution because they may perform poorly on nonstationary problems. Is it possible to avoid the bias of constant step sizes while retaining their advantages on nonstationary problems? One way is to use a step size of

$$\beta_n := \alpha/\bar{o}_n,$$

to process the nth reward for a particular action, where $\alpha > 0$ is a conventional constant step size, and $\bar{o}_n$ is a trace of one that starts at 0:

$$\bar{o}_n := \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}), \text{ for } n \geq 0, \text{ with } \bar{o}_0 := 0.$$

Carry out an analysis like that in (2.6) to show that Q_n is an exponential recency-weighted average *without initial bias.*

Solution By the results from exercise 2.4, we have

$$\begin{aligned} Q_{n+1} &= Q_n + \beta_n[R_n - Q_n] \\ &= Q_1 \prod_{i=1}^{n}(1-\beta_i) + \sum_{i=1}^{n} \beta_i \prod_{j=i+1}^{n}(1-\beta_j)R_i \end{aligned}$$

To show that this exponential recency-weighted average does not have initial bias, we need to show that $\prod_{i=1}^{n}(1-\beta_i) = 0$.

$$
\begin{aligned}
\prod_{i=1}^{n}(1-\beta_i) &= \prod_{i=1}^{n}\left(1-\frac{\alpha}{\bar{o}_{i-1}+\alpha(1-\bar{o}_{i-1})}\right) \\
&= \left(1-\frac{\alpha}{\bar{o}_0+\alpha(1-\bar{o}_0)}\right)\prod_{i=2}^{n}\left(1-\frac{\alpha}{\bar{o}_{i-1}+\alpha(1-\bar{o}_{i-1})}\right) \\
&= \left(1-\frac{\alpha}{0+\alpha(1-0)}\right)\prod_{i=2}^{n}\left(1-\frac{\alpha}{\bar{o}_{i-1}+\alpha(1-\bar{o}_{i-1})}\right) \\
&= 0
\end{aligned}
$$