Exercise 2.1 In ε-greedy action selection, for the case of two actions and $\varepsilon = 0.5$, what is the probability that the greedy action is selected? ☑

$$\text{Action} = \begin{cases} 50\%\ \text{Exploit} - \text{Greedy} \\ 50\%\ \text{Explore} \begin{cases} 50\%\ \text{Greedy} \\ 50\%\ \text{Explore} \end{cases} \end{cases} = \begin{cases} 75\%\ \text{Greedy} \\ 25\%\ \text{Explore} \end{cases}$$

Exercise 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 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 these time steps the ε 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? ☑

i	1:	2:	3:	4:	5:
A_i	1	2	2	2	3
R_i	−1	+1	−2	2	0
ε_i	?	?	?	✓	✓

? – possible

✓ – certain

	A_i	1	2	3	4	
	S^0	0	0	0	0	=> $Q^0 = [\underline{0^*}\ 0^*\ 0^*\ 0^*]$
	n^0	0	0	0	0	
1:	S^1	−1	0	0	0	=> $Q^1 = [-1\ \underline{0^*}\ 0^*\ 0^*]$
	n^1	1	0	0	0	
2:	S^2	−1	1	0	0	=> $Q^2 = [-1\ +\underline{1^*}\ 0\ 0]$
	n^2	1	1	0	0	
3:	S^3	−1	−1	0	0	=> $Q^3 = [-1\ \underline{-1/2}\ 0^*\ 0^*]$
	n^3	1	2	0	0	
4:	S^4	−1	+1	0	0	=> $Q^4 = [-1\ +1/3^*\ \underline{0}\ 0]$
	n^4	1	3	0	0	
5:	S^5	−1	+1	0	0	=> $Q^5 = [-1\ +1/3^*\ 0\ 0]$
	n^5	1	3	1	0	

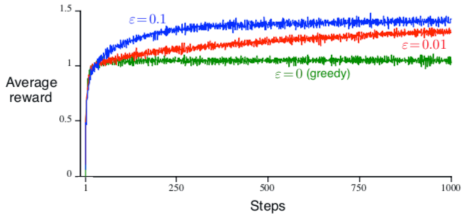
S^i ... sum of action rewards up to step i.

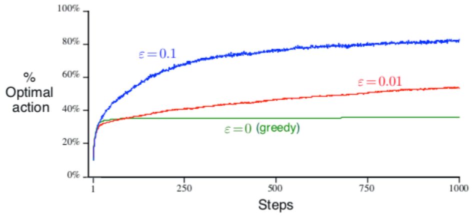
n^i ... number of times an action has been taken up to step i.

Q^i ... sample average action value after step i $(= S^i/n^i)$.

$*$ ... optimal actions based on current Q^i

Exercise 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 answer quantitatively. ☑





Once they have fully converged the exploration strategies pick the optimal action $a^* = \underset{i}{\operatorname{argmax}}\, \mathbb{E}[R_i]$ $1-\varepsilon$ times in the exploitation mode plus ε/n times in the random exploration move.

$$p(A_i=a^*) = 1-(1-1/n)\varepsilon = \begin{cases} 91\% & (\varepsilon=0.1) \\ 99.1\% & (\varepsilon=0.01) \end{cases}$$

Let's assume the expected maximum of cumulative rewards $q^* = \mathbb{E}[R_i|A_i=a^*]$ All the actions together have $\mathbb{E}[R_i|A_i]=0$ zero mean. Then the expectation of the reward is:

$$\mathbb{E}[R_i] = p(A_i=a^*)\cdot q^* + (1-p(A_i=a^*))\cdot 0 = \begin{cases} 91\%\, q^* & (\varepsilon=0.1) \\ 99.1\%\, q^* & (\varepsilon=0.01) \end{cases}$$

Exercise 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? ☑

Non-stationary

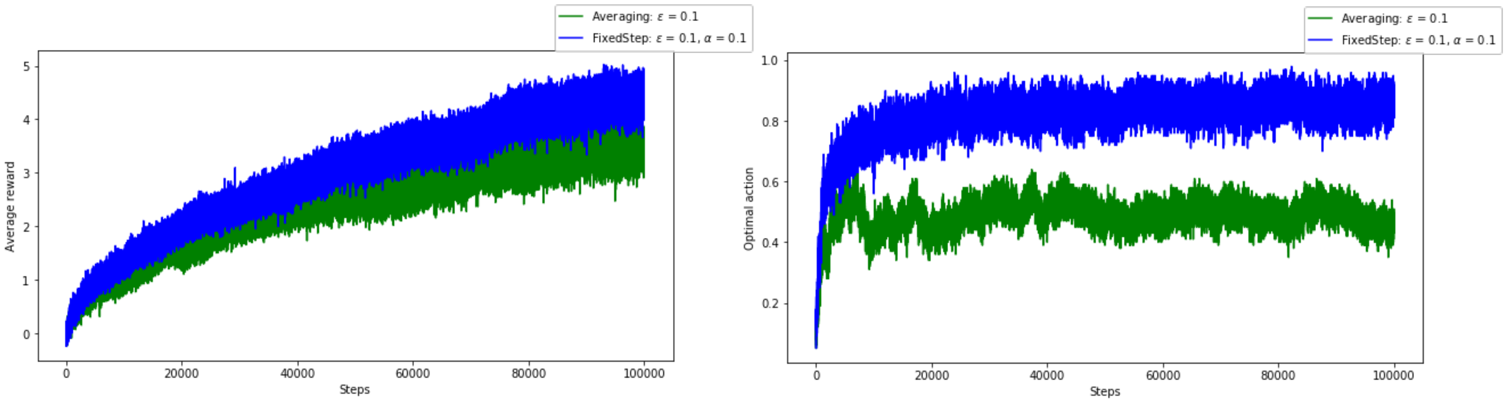
$$Q_{n+1} = Q_n + \alpha_n\left[R_n - Q_n\right]$$

$$Q_{n+1} = \alpha_n R_n + (1-\alpha_n)\underbrace{Q_n}_{\alpha_{n-1}R_{n-1} + (1-\alpha_{n-1})\underbrace{Q_{n-1}}_{\alpha_{n-2}R_{n-2} + (1-\alpha_{n-2})Q_{n-2}}}$$

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\ &= \alpha R_n + (1-\alpha)Q_n \\ &= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right] \\ &= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\alpha R_{n-2} + \\ &\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i. \end{aligned} \tag{2.6}$$

$$Q_{n+1} = Q_1 \prod_{i=1}^{n}(1-\alpha_i) + \sum_{i=1}^{n} \alpha_i \prod_{j=i+1}^{n}(1-\alpha_j)\, R_i$$

Exercise 2.5 (programming) Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for nonstationary 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 with 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. ☑



Exercise 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? ☑

Early in the iteration process all agents gradually try all optimistic Q=5 values, get disappointed and correct the initial optimistic value. The best actions receive the smallest correction and are therefore selected at the same time by all the agents. Because the Q value is still optimistic there's another round of corrections toward 0 (the mean reward for all arms). Due to random explorations the agents get out of sync over time so the following peaks are not as large as the first one occuring after k steps.

Exercise 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 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

$$\beta_n \doteq \alpha/\bar{o}_n, \tag{2.8}$$

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:

$\bar{o}_1 = \alpha$

$\bar{o}_2 = \alpha(2-\alpha)$

$\bar{o}_3 = \dots$

$$\bar{o}_n \doteq \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}), \quad \text{for } n \geq 0, \quad \text{with } \bar{o}_0 \doteq 0. \tag{2.9}$$

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

$$Q_{n+1} = Q_n + \overset{=\alpha/\bar{o}_n}{\beta_n}[R_n - Q_n]$$

$$Q_{n+1} = \beta_n R_n + (1-\beta_n)\underbrace{Q_n}_{\beta_{n-1}R_{n-1} + (1-\beta_{n-1})Q_{n-1}}$$

$\dots$

no initial bias because arbitrarily initialized Q_1 has no effect on Q_i $i \geq 2$

$$Q_{n+1} = \prod_{i=2}^{n}(1-\beta_i)R_1 + \sum_{i=2}^{n} \beta_i \prod_{j=i}^{n} (1-\beta_j) R_i$$

$$Q_2 = \beta_1 R_1 + (1-\beta_1)Q_1 = R_1$$

$$\beta_1 = \alpha/\bar{o}_1 = \frac{\alpha}{\underbrace{\bar{o}_0}_{=0} + \alpha(1-\underbrace{\bar{o}_0}_{=0})} = 1$$

Step size decay factor $\bar{o}_n$ is a geometric series converging to 1. So over time $\lim \beta_n = \alpha$ and the weighting quickly approaches classical constant step size update.

$$\bar{o}_n = \bar{o}_{n-1}(1-\alpha) + \alpha$$

$$\underbrace{[\bar{o}_n - \bar{o}_*]} = \underbrace{[\bar{o}_{n-1} - \bar{o}_*]}(1-\alpha)$$

$$\bar{o}_n^* = \bar{o}_{n-1}^* (1-\alpha)$$

$$\lambda^n = \lambda^{n-1}(1-\alpha)$$

$$\lambda = (1-\alpha)$$

$$\bar{o}_n^* = C \cdot \lambda^n$$

$$\bar{o}_n = 1 - (1-\alpha)^n$$

$$\beta_n = \frac{\alpha}{1-(1-\alpha)^n}$$

Steady state:

$$\bar{o}_* = \bar{o}_*(1-\alpha) + \alpha$$

$$\bar{o}_* = 1$$

Constant C:

$$\bar{o}_1 = \bar{o}_1^* + \bar{o}^* = C \cdot (1-\alpha)^1 + 1$$

$$\bar{o}_1 = \alpha$$

$$C = \frac{\alpha - 1}{1-\alpha} = -1$$

Exercise 2.8: UCB Spikes In Figure 2.4 the UCB algorithm shows a distinct spike in performance on the 11th step. Why is this? Note that for your answer to be fully satisfactory it must explain both why the reward increases on the 11th step and why it decreases on the subsequent steps. Hint: If $c = 1$, then the spike is less prominent. ☑

Initially no action has been visited so $N_t(a) = 0 \; \forall a$. All actions are maximizing ones. As the actions are visited $N_t(a)$ gets set to a non zero value the action selection at 11th steps uses the action with the highest Q value. Highest Q value option is very likely to yield good reward that causes the spike.

After that the 2000 averaged agents get a little bit out of sync becuse some of the highest Q actions turned out to be noise and not signal. Agent then has to adjust Q values to the new reward. This follow up process is ditermined by chance so agents get out of sync. Each one sees high values at different times so the peak is very unlikely to happen again.

Exercise 2.9 Show that in the case of two actions, the soft-max distribution is the same as that given by the logistic, or sigmoid, function often used in statistics and artificial neural networks. ☑

$$\Pr(A_t = 1) = \frac{e^{H_t(1)}}{\sum_{b=1}^{2} e^{H_t(b)}} = \frac{e^{H_t( )}}{e^{H_t(1)} + e^{H_t(2)}} = \frac{1}{1 + e^{H_t(2)/H_t(1)}} = \text{sigmoid}\left(-\frac{H_t(2)}{H_t(1)}\right)$$

$$\Pr(A_t = 2) = \frac{e^{H_t(2)}}{\sum_{b=1}^{2} e^{H_t(b)}} = \frac{e^{H_t(2)}}{e^{H_t(1)} + e^{H_t(2)}} = \frac{1}{1 + e^{H_t(1)/H_t(2)}} = \text{sigmoid}\left(-\frac{H_t(1)}{H_t(2)}\right)$$

Exercise 2.10 Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it? ☑

A: 1: 0.1 50%, 2: 0.2 50%

B: 1: 0.9 50%, 2: 0.8 50%

1) A and B are indistinguishable

$$\left.\begin{aligned} \mathbb{E}[R|1] &= 0.5 \\ \mathbb{E}[R|2] &= 0.5 \end{aligned}\right\}$$

Both actions yield equal return

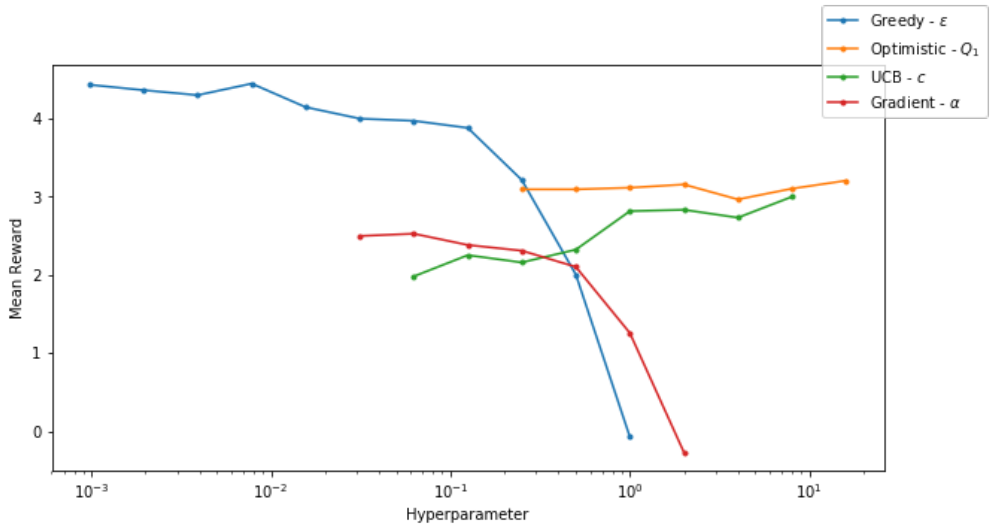
2) A and B are distinguishable

$\mathbb{E}[R|1,A] = 0.1$ $\quad$ $\mathbb{E}[R|1,B] = 0.9$

$\mathbb{E}[R|2,A] = 0.2$ $\quad$ $\mathbb{E}[R|2,B] = 0.8$

Optimal strateg yields expected reward 0.55 by taking action 2 in A and 1 in B.

Exercise 2.11 (programming) Make a figure analogous to Figure 2.6 for the nonstationary case outlined in Exercise 2.5. Include the constant-step-size ε-greedy algorithm with $\alpha = 0.1$. Use runs of 200,000 steps and, as a performance measure for each algorithm and parameter setting, use the average reward over the last 100,000 steps. ☑



ε-Greedy does the best job at adapting to a non-stationary environment. Other methods level off exploration after the initial phase and don't discover the evolving optimal arm.