Ex. 2.1: $0.5 + 0.5 * 0.5 = 0.75$

Ex. 2.2: Random action is definite only where a non-optimal action is selected.

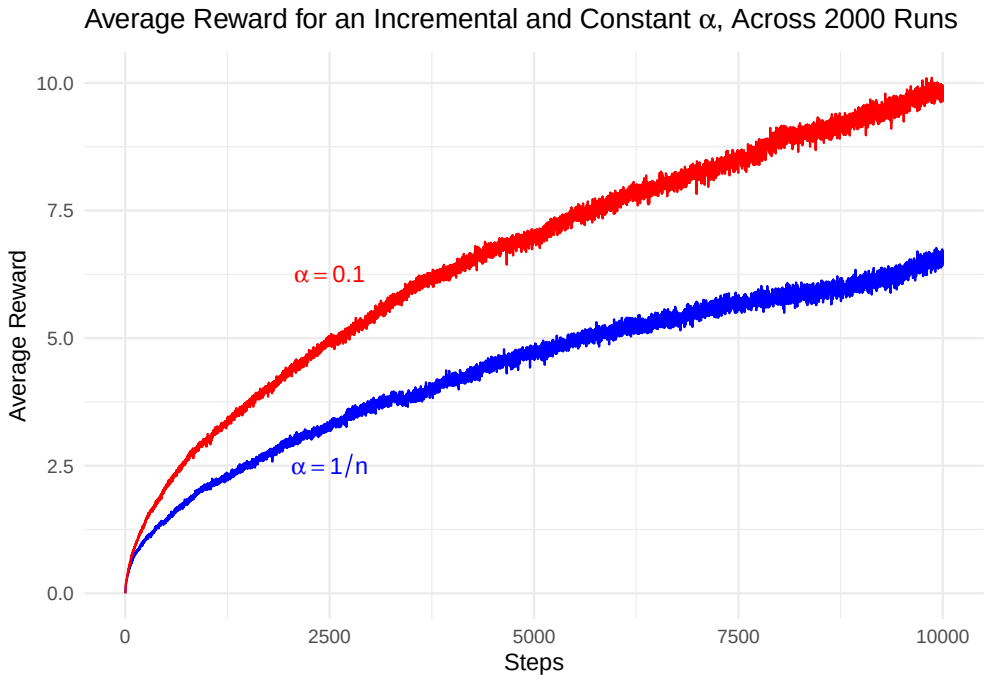
Time	Q.1.	Q.2.	Q.3.	Q.4.	Random.Action
0	0	0.00	0	0	Possible
1	-1	0.00	0	0	Possible
2	-1	1.00	0	0	Possible
3	-1	-0.50	0	0	Definite
4	-1	0.33	0	0	Definite
5	-1	0.33	0	0	

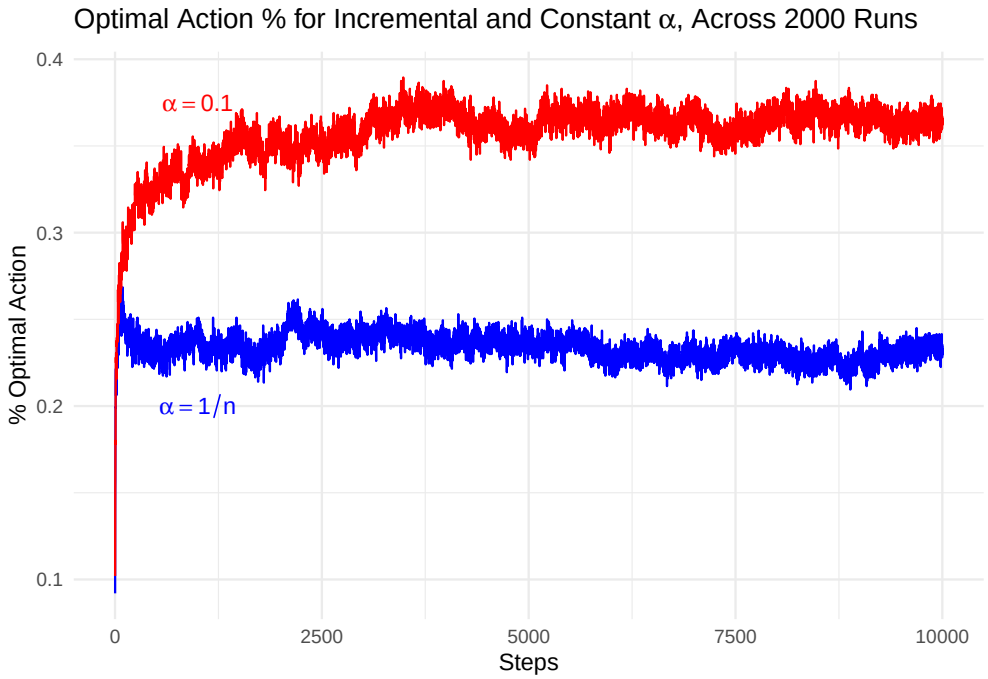
Ex. 2.3: $\epsilon = 0.01$. It will have a higher probability ($1 - 0.01 = 0.99$ vs. $1 - 0.1 = 0.9$) of selecting the best action relative to $\epsilon = 0.1$. The extent of improvement above a greedy policy is dependent on the greedy method's initial performance.

Ex. 2.4:

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= \alpha_n R_n + (1 - \alpha_n)Q_n \\
&= \alpha_n R_n + (1 - \alpha_n) \times (\alpha_{n-1}R_{n-1} + (1 - \alpha_{n-1})Q_{n-1}) \\
&= \alpha_n R_n + (1 - \alpha_n)\alpha_{n-1}R_{n-1} + (1 - \alpha_n)(1 - \alpha_{n-1})Q_{n-1} \\
&= \alpha_n R_n + (1 - \alpha_n)\alpha_{n-1}R_{n-1} + (1 - \alpha_n)(1 - \alpha_{n-1})\alpha_{n-2}R_{n-2} + \cdots + \alpha_1 \prod_{i=1}^{n}(1 - \alpha_i)R_1 + \prod_{i=1}^{n}(1 - \alpha_i)Q_1 \\
&= \prod_{i=1}^{n}(1 - \alpha_i)Q_1 + \sum_{i=1}^{n}\alpha_i \prod_{i+1, i<n}^{n}(1 - \alpha_i)R_i
\end{aligned}
$$

Ex. 2.5:





Ex 2.6: Expect worse performance, on average, as the optimistic strategy encourages more exploration during earlier steps. The oscillations occur immediately after all states have been tested (10 steps). The optimistic strategy causes all states to be reached quicker, given the large initial Q-values of untouched states. At the 11th step and soon thereafter, more optimal action selection occurs given the models breadth of experience. This optimistic effect subsequently wears off and learning resembles a realistic scenario.

Ex 2.7: No initial bias exists as the Q_1 factor equates to 0.

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

$$\begin{aligned} \prod_{i=1}^{n}(1-\beta_i) &= \prod_{i=1}^{n}(1-\alpha/\theta_i) \\ &= (1-\alpha/\theta_1)\prod_{i=2}^{n}(1-\alpha/\theta_i) \qquad \text{where} \quad \theta_1 = \theta_0 + \alpha(1-\theta_0) = \alpha \\ &= (1-\alpha/\alpha)\prod_{i=2}^{n}(1-\alpha/\theta_i) \\ &= 0 \end{aligned}$$

Ex. 2.8: UCB (with c = 2) explores all bandits within 10 steps as a is maximising if $N_t(a) = 0$. The 11th step is more informed relative to a slower ϵ-greedy explorer (which is unlikely to have seen all states) thereby resulting in a reward spike. A drop in relative performance is subsequently experienced as the level of exploration is still high $Q_0(i) = 2\sqrt{\frac{ln(t)}{1}}$ where $t > 10$ (e.g. $t = 11, Q_{11}(i) + 3.1$) prior to an action's second selection, thus decreasing greedy selections. Exploration tapers off as t increases.

Ex 2.9:

$$\pi_t(a) = \frac{e^{H_t(a)}}{e^{H_t(a)} + e^{H_t(b)}} = \frac{1}{1 + e^{H_t(b) - H_t(a)}} = \frac{1}{1 + e^{-(H_t(a) - H_t(b))}}$$

Ex 2.10:

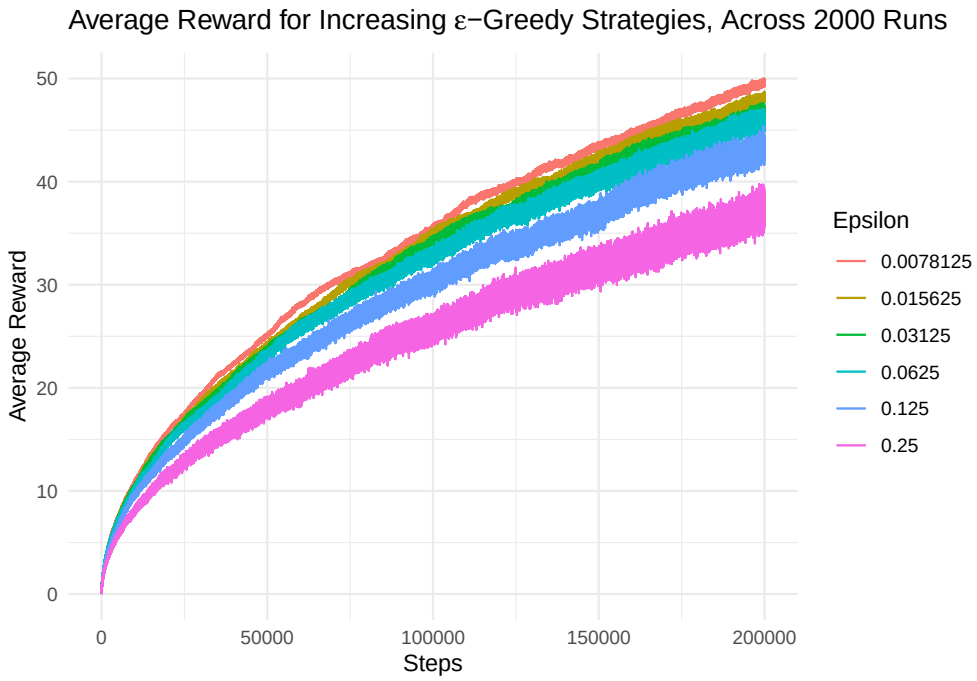
$A_1 = a : E[R] = 0.5 * (0.1 + 0.9) = 0.5$
$A_2 = a : E[R] = 0.5 * (0.2 + 0.8) = 0.5$

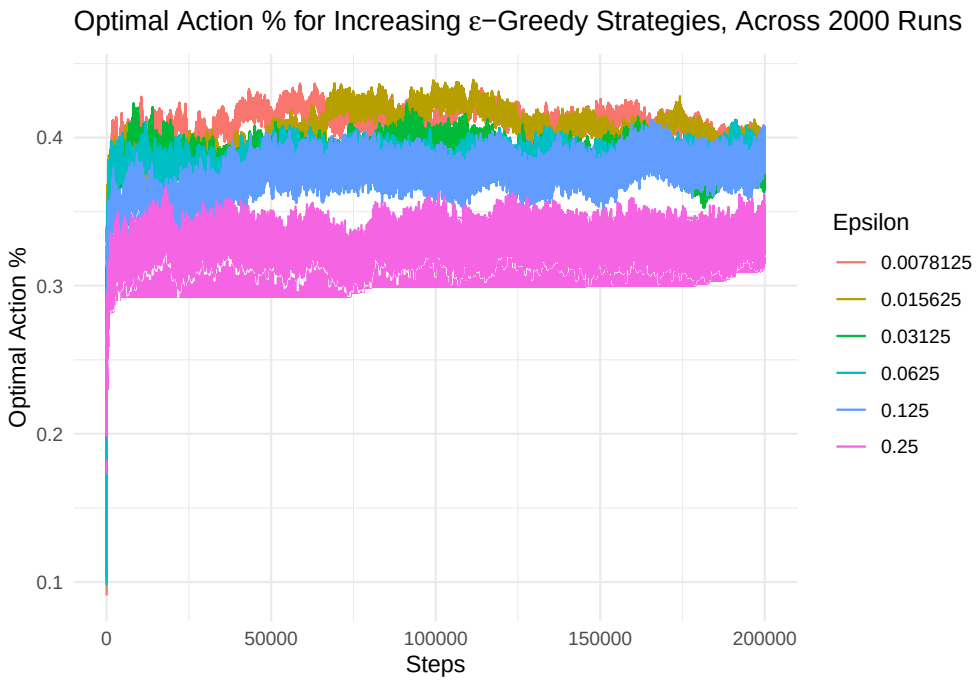
Best expectation of success: 0.5
Strategy: Choose randomly

Case A: $E[R] = 0.2$
Case B: $E[R] = 0.9$

Strategy: Depending on the case, select the action that has the highest expected value (in effect a 4 state problem).

Ex 2.11:




Optimal Action % for Increasing ε−Greedy Strategies, Across 2000 Runs
Optimal Action %
0.4
0.3
0.2
0.1
0
50000
100000
150000
200000
Steps
Epsilon
0.0078125
0.015625
0.03125
0.0625
0.125
0.25


Average Reward over the Last 100k Steps for Different ε−Greedy Strategies
Average reward over the last 100 000 steps
40
36
32
0.00
0.05
0.10
0.15
0.20
0.25
ε