Initialize, for $a = 1$ to k:
$Q(a) \leftarrow 0$
$N(a) \leftarrow 0$

Loop forever:
$A \leftarrow \begin{cases} \text{argmax}_a\, Q(a) & \text{with probability } 1-\varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
$R \leftarrow bandit(A)$
$N(A) \leftarrow N(A) + 1$
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}[R - Q(A)]$

$$Q_{n+1} = \quad (1-\alpha)^n Q_1 + \sum_n^{} \alpha(1-\alpha)^{n-i} R_i$$

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

$$A_t \doteq \underset{a}{\text{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Dynamics of the MDP

$$p(s', r \,|\, s, a) \;\doteq\; \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

State-transition probabilities

$$p(s' \,|\, s, a) \;\doteq\; \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} \;=\; \sum_{r \in \mathcal{R}} p(s', r \,|\, s, a).$$

Expected rewards state-action

$$r(s, a) \;\doteq\; \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] \;=\; \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \,|\, s, a)$$

Expected rewards state-action-state

$$r(s, a, s') \;\doteq\; \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] \;=\; \sum_{r \in \mathcal{R}} r \frac{p(s', r \,|\, s, a)}{p(s' \,|\, s, a)}.$$

State-value function for policy pi

$$\begin{aligned} v_\pi(s) \;&\doteq\; \mathbb{E}_\pi[G_t \mid S_t = s] \;=\; \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s \right] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \,|\, s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \Big] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r \,|\, s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all } s \in \mathcal{S}, \end{aligned}$$

Action-value function for policy pi

$$q_\pi(s, a) \;\doteq\; \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \;=\; \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a \right]$$

Bellman optimality equation

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\ &= \max_a \sum_{s', r} p(s', r \,|\, s, a) \big[r + \gamma v_*(s') \big] \\ q_*(s, a) \;&=\; \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\Big|\; S_t = s, A_t = a \Big] \\ &=\; \sum_{s', r} p(s', r \,|\, s, a) \Big[r + \gamma \max_{a'} q_*(s', a') \Big]. \end{aligned}$$

Some equations related to Bellman

$$v_\pi(s) = \sum_{a} \pi(a|s) \sum_{s'} p(s'|s, a) \left[r(s, a) + v_\pi(s') \right]$$

$$v_*(s) = \max_a \left\{ \sum p(s'|s, a) \left[r(s, a) + v_*(s') \right] \right\}$$

$$q_\pi(s, a) = \sum_{s'} p(s'|s, a) \left[r(s, a) + \gamma \sum_{a'} \pi(a'|s')\, q_\pi(s', a') \right]$$

$$q_*(s, a) = \sum_{s'} p(s'|s, a) \left[r(s, a) + \gamma \max_{a'} q_*(s', a') \right]$$

- If the step-size parameters, an, are not constant, then the estimate Qn is a weighted average of previously received rewards with a weighting different from that given by Equation 2.6. What is the weighting on each prior reward for the general case, analogous to Equation 2.6, in terms of the sequence of step-size parameters?

$$
\begin{aligned}
&Q_{n+1} \\
&= Q_n + \alpha_n \left(R_n - Q_n\right) \\
&= \alpha_n R_n + (1 - \alpha_n)\left[Q_{n-1} + \alpha_{n-1}\left(R_{n-1} - Q_{n-1}\right)\right] \\
&= \ldots \\
&= \sum_{i=1}^{n} \alpha_i R_i \prod_{i=i+1}^{n} (1 - \alpha_i) + Q_1 \prod_{i=1}^{n} (1 - \alpha_i)
\end{aligned}
$$

$Q_t(a) = \frac{sumofrewardswhenatakenpriortot}{numberoftimesatakenpriortot} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} = \mathbb{E}[R_i | A_i = a] = q^*(a)$

- If Q1 = 0, is Qn (for n > 1) biased? Justify your answer with brief words and equations

(b) equation 2.6

$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

when $Q_1 = 0$

$Q_{n+1} = \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

$\Rightarrow E(Q_{n+1}) = E\left(\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i\right)$

$= \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} E(R_i)$

$\Rightarrow$ when $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$

$E(Q_{n+1}) = E(R_i)$

$\Rightarrow$ equation is unbiased

- Derive condition(s) for Q1 for when Qn will be unbiased.

(c) Q_n will be unbiased if $\begin{cases} Q_1 = 0 \\ \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1 \end{cases}$

- Show that Qn is an unbiased estimator as n. (which is often referred to as asymptotically unbiased).

(d) $Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

$\lim_{n\to\infty} Q_{n+1} = \lim_{n\to\infty} (1-\alpha)^n Q_1 + \lim_{n\to\infty} \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

$= \lim_{n\to\infty} \alpha \cdot \frac{1-(1-\alpha)^{n+1}}{1-(1-\alpha)} \cdot R_i$

$= \lim_{n\to\infty} \alpha \frac{1-(1-\alpha)^{n+1}}{\alpha} \cdot R_i$

$= \lim_{n\to\infty} 1 - (1-\alpha)^{n-1} \cdot R_i$

$= R_i$

$\Rightarrow \lim_{n\to\infty} E(Q_{n+1}) = E(R_i)$

- Why should we expect that the exponential recency-weighted average will be biased in practice?

For a, we are able to control the learning rate by modify a, the learning process could be speed up by increasing a and vice versa. For Q1, the agent could start explore more in the beginning rather then always choosing the best choice by increase Q1.

- 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

4. Softmax: $\int_a = \frac{\exp[H_t(a)]}{\sum_{b=1}^{k} \exp[H_t(b)]}$

case of two actions:

$\int_a = \frac{\exp[H_t(a)]}{\exp[H_t(a)] + \exp[H_t(b)]} = \frac{1}{1 + \exp[H_t(b) - H_t(a)]}$

if $Q = -\left(H_t(b) - H_t(a)\right)$

then $\int_a = \frac{1}{1+\exp[-Q]} = \sigma(Q)$ → sigmoid

- UCB also produce spikes in the very beginning in both the two reproduced figures. Explain in your own words why the spikes appear (both the sharp increase and sharp decrease)

- In the optimistic initialization, the first few steps are not really random. Instead, we loop through all the actions multiple times, picking our large optimistic value at random. In the first round, all actions have equal optimization percentages (10%).The next action will have a spike, because on average the best action will have the largest value. This will be repeated with decreasing values until the effect of the optimal initial value fades away.