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

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$
$N(A) \leftarrow N(A) + 1$
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

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

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

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

Soft-max distribution

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

Action preferences

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)), \quad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a), \quad \text{for all } a \neq A_t,$$

Policy Evaluation: *Sequence {vk} can be shown in general to converge to vpi as k-> ∞ under the same conditions that guarantee the existence of vpi. This algorithm is called **iterative policy evaluation**.*

Policy Improvement

New greedy policy, pi0

$$\begin{aligned} \pi'(s) &\doteq \arg\max_a q_\pi(s,a) \\ &= \arg\max_a \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \arg\max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], \end{aligned}$$

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

$$\begin{aligned} v_*(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_*(s')\Big], \text{ or} \end{aligned}$$

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1},a') \,\Big|\, S_t = s, A_t = a\Big] \\ &= \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \max_{a'} q_*(s',a')\Big], \end{aligned}$$

Some equations related to Bellman

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

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

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$$\begin{aligned} v_{\pi'}(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_{\pi'}(s')\Big]. \end{aligned}$$

$$\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_*,$$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))\left[r + \gamma V(s')\right]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Value Iteration

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_k(s')\Big], \end{aligned}$$

Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
$G \leftarrow \gamma G + R_{t+1}$
Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
Append G to $Returns(S_t, A_t)$
$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$
Initialize:
$\pi \leftarrow$ an arbitrary ε-soft policy
$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):
Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
$G \leftarrow \gamma G + R_{t+1}$
Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
Append G to $Returns(S_t, A_t)$
$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)
For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

Probability of nongreedy actions $\dfrac{\varepsilon}{|\mathcal{A}(s)|}$

- *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?*

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

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

$$
\begin{aligned}
\lim_{n\to\infty} Q_{n+1} &= \lim_{n\to\infty} (1-\alpha)^n Q_1 + \lim_{n\to\infty} \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i\\
&= \lim_{n\to\infty} \alpha \cdot \frac{1-(1-\alpha)^{n-1}}{1-(1-\alpha)} \cdot R_i\\
&= \lim_{n\to\infty} \alpha \frac{1-(1-\alpha)^{n-1}}{\alpha} \cdot R_i\\
&= \lim_{n\to\infty} 1-(1-\alpha)^{n-1} \cdot R_i\\
&= R_i
\end{aligned}
$$

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

- *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: $\int_a = \dfrac{\exp\{H_t(a)\}}{\sum_{b=1}^{k} \exp\{H_t(b)\}}$

case of two action:

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

if $Q = -(H_t(b) - H_t(a))$

then $\int_a = \dfrac{1}{1+\exp\{-Q\}} = \sigma(Q)$ → 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.

Expand the Bellman equation for the 2 states in the recycling robot, for an arbitrary policy π (a|s), discount factor γ, and domain parameters α, β, rsearch, rwait as described in the example.

$$
\begin{aligned}
v_\pi(high) = &\ \pi(search|high) \cdot P(high, r_{search}|high, search) \times [r_{search} + \gamma v_\pi(high)]\\
&+ \pi(search|high) \cdot P(low, r_{search}|high, search) \times [r_{search} + \gamma v_\pi(low)]\\
&+ \pi(wait|high) \cdot P(high, r_{wait}|high, wait) \times [r_{wait} + \gamma v_\pi(high)]\\
= &\ \pi(search|high) \cdot \alpha \times [r_{search} + \gamma v_\pi(high)]\\
&\ \pi(search|high) \cdot (1-\alpha) \cdot [r_{search} + \gamma v_\pi(low)]\\
&\ \pi(wait|high) \cdot [r_{wait} + \gamma v_\pi(high)]
\end{aligned}
$$

$$
\begin{aligned}
v_\pi(low) = &\ \pi(wait|low) \cdot P(low, r_{wait}|low, wait) \times [r_{wait} + \gamma v_\pi(low)]\\
&+ \pi(search|low) \cdot P(low, r_{search}|low, search) \times [r_{search} + \gamma v_\pi(low)]\\
&+ \pi(recharge|low) \cdot P(low, 0|low, recharge) \times [0 + \gamma v_\pi(high)]\\
&+ \pi(search|low) \cdot P(high, r_{search}|low, search) \times [-3 + \gamma v_\pi(low)]\\
= &\ \pi(wait|low) \cdot [r_{wait} + \gamma v_\pi(low)]\\
&+ \pi(search|low) \cdot \beta\, [r_{search} + \gamma v_\pi(low)]\\
&+ \pi(recharge|low) \cdot \gamma v_\pi(high)\\
&+ \pi(search|low) \cdot (1-\beta) \cdot [-3 + \gamma v_\pi(low)]
\end{aligned}
$$

You should now have two linear equations involving two unknowns, v(high) and v(low), as well as involving the policy π(a|s), γ, and the domain parameters. Let α = 0.8, β = 0.6, γ = 0.9, rsearch = 10, rwait = 3. Consider the policy π(search | high) = 1, π(wait | low) = 0.5, and π(recharge | low) = 0.5.

$$
\begin{aligned}
v_\pi(high) = &\ \pi(search|high) \cdot \alpha \times [r_{search} + \gamma v_\pi(high)]\\
&+ \pi(search|high) \cdot (1-\alpha) \cdot [r_{search} + \gamma v_\pi(low)]\\
&+ \pi(wait|high) \cdot [r_{wait} + \gamma v_\pi(high)]\\
= &\ 1 \times 0.8 \times [10 + 0.9 v_\pi(high)]\\
&+ 1 \times 0.2 \times [10 + 0.9 v_\pi(low)]
\end{aligned}
$$

$$v_\pi(high) = 8 + 0.72 v_\pi(high) + 2 + 0.18 v_\pi(low)$$

$$0.28 v_\pi(high) = 10 + 0.18 v_\pi(low) \quad ①$$

$$
\begin{aligned}
v_\pi(low) = &\ \pi(wait|low) \cdot [r_{wait} + \gamma v_\pi(low)]\\
&+ \pi(search|low) \cdot \beta\, [r_{search} + \gamma v_\pi(low)]\\
&+ \pi(recharge|low) \cdot \gamma v_\pi(high)\\
&+ \pi(search|low) \cdot (1-\beta) \cdot [-3 + \gamma v_\pi(high)]\\
= &\ 0.5 \times [3 + 0.9 \times v_\pi(low)]\\
&+ 0.5 \times 0.9 \times v_\pi(high)
\end{aligned}
$$

$$v_\pi(low) = 1.5 + 0.45 v_\pi(low) + 0.45 v_\pi(high)$$

$$0.55 v_\pi(low) = 1.5 + 0.45 v_\pi(high)$$

$$11 v_\pi(low) = 30 + 9 v_\pi(high) \quad ② \qquad \Rightarrow 77 v_{low} = 210 + 63 v_{high}$$

Combine ① and ②

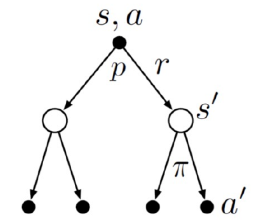
$$\Rightarrow 14 v_\pi(high) = 500 + 9 v_\pi(low) \qquad \Rightarrow 63 v_{high} = 2250 + 40.5 v_{low}$$

$$365 v_{low} = 2460$$

$$\Rightarrow v_{low} = 67$$

$$v_{high} = 79$$

$$14 v_{high} = 500 + 603$$



q_π backup diagram

Give an equation for vπ in terms of qπ and π.

$$
\begin{aligned}
v_\pi &= E_\pi[G_t | S_t = s]\\
&= \sum_a E_\pi[G_t | S_t = s, A_t = a]\, P[A_t = a | S_t = s]\\
&= \sum_a q_\pi(s, a)\, \pi(a|s)
\end{aligned}
$$

Give an equation for qπ in terms of vπ and the four-argument p.

$$
\begin{aligned}
\text{(b)}\ q(s, a) &= E_\pi[G_t | S_t = s, A_t = a]\\
&= \sum_{s'} [G_t | S_t = s, A_t = a, S_{t+1} = s']\, P[S_{t+1} = s' | S_t = s, A_t = a]\\
&= \sum_{s'} [G_t | S_{t+1} = s']\, P[S_{t+1} = s' | S_t = s, A_t = a]\\
&= \sum_{s', r} \{r + \gamma E_\pi[G_{t+1} | S_{t+1} = s']\}\, p(s', r | s, a)\\
&= \sum_{s', r} [r + \gamma v_\pi(s')]\, p(s', r | s, a)
\end{aligned}
$$

What is the Bellman equation for action values, that is, for qπ? It must give the action value qπ(s, a) in terms of the action values, qπ(s′, a′), of possible successors to the state–action pair (s, a).

$$
\begin{aligned}
q_\pi(s, a) &= \sum_{s', r} p(s', r | s, a)\, [r + \gamma v_\pi(s')]\\
&= \sum_{s', r} p(s', r | s, a)\left[r + \sum_{a' \in A} \pi(a'|s') \cdot \gamma q_\pi(s', a')\right]
\end{aligned}
$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}.$$

$$\sum_{k=0}^{\infty} y^k = 1 + y + y^2 + y^3 + \ldots =$$

$$= 1 + y(1 + y + y^2 + y^3 + \ldots) = 1 + y \left(\sum_{k=0}^{\infty} y^k\right)$$

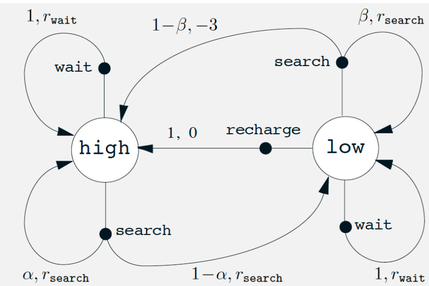
$$\sum_{k=0}^{\infty} y^k = -\frac{1}{y-1} = \frac{1}{1-y}$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

$$G'_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+1+k} + c) = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} + \sum_{k=0}^{\infty} c \cdot \gamma^k =$$

$$= G_t + \frac{c}{1-\gamma}$$

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	r_{search}
high	search	low	$1-\alpha$	r_{search}
low	search	high	$1-\beta$	-3
low	search	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	-



s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
low	wait	low	r_{wait}	1
low	recharge	high	0	1