$$V^{\pi}(s) = E^{\pi}[R_T \mid S_t = s]$$

$$V^{\pi}(s) = \sum_{s' \in S} T(s, \pi(s), s')[r(s, \pi(s), s') + \gamma V^{\pi}(s)]$$

$$V^{*}(s) = E^{\pi^{*}}[R_t \mid S_t = s]$$

$Q^{\pi}(S,a)$ Función de valor estado-acción

$$E^{\pi}[R_t \mid S_t = s, A_t = a]$$

$$Q^{\pi}(s,a) = \sum_{s' \in S} T(s,a,s') * (r(s,a,s') + \gamma Q^{\pi}(s', \pi(s')))$$

$$Q^{*}(S,a) = \sum_{s' \in S} T(s,a,s')[r(s,a,s') + \gamma \max_{a \in A(s')} Q^{*}(s',a')]$$

Políticas ε-greedy

$$\pi_{\varepsilon}(s) = \begin{cases} \pi(s) & \text{con prob } 1-\varepsilon \\ \text{Acción aleatoria} & \text{con prob } \varepsilon \end{cases}$$

$$Q^{*}(S,a) = \sum_{s' \in S} T(s,a,s')[r(s,a,s') + \gamma \max_{a \in A(s')} Q^{*}(s',a')]$$

1: Tengo el estado s

2: Selecciono la acción a con $\pi^*_\varepsilon(s) = \begin{cases} \max_{a \in A(s)} Q(s,a) & 1-\varepsilon \\ a \in A(s) & \varepsilon \end{cases}$

③: $\hat{Q}(s,a)$

4: Aplico a al sistema y obtengo s' y r

5: Selecciono la acción a' con $\pi^*_\varepsilon(s')$ $\rightarrow \begin{cases} \max_{a \in A(s)} Q(s',a') & 1-\varepsilon \\ \text{aleatoria} & \varepsilon \end{cases}$

6: $\hat{Q}(s,a|s',a') = r + \gamma Q(s',a')$

7.- $\delta = \hat{Q}(s,a|s',a') - \hat{Q}(s,a)$

8.- $\hat{Q}(s,a) = \hat{Q}(s,a) + \alpha\delta$

9.- $s = s'$, $a = a'$

10: Fuga a ③ pa

SARSA

On policy

```
def SARSA(mdp, α, ε, max_e, max_it, θ):
  Q = {(s,a): 0 for s in mdp.S for a in mdp.acciones_legales(s)}
  for _ in range(max_e):
     s = mdp.estado_inicial()
     a = max_{a∈A(s)} [Q(s,a)] if random() ≤ 1-ε else aleatorio(A(s))

     for _ range max_it
        s' = mdp.transición(s,a,s')
        a' = max_{a∈A(s)} [Q(s,a)] if random() ≤ 1-ε else choice(A(s))
        q = Q[[s,a]] + α[r + γ * Q([s',a'])) - Q([s,a]))
        Δ = max(Δ, abs(q - Q([s,a])))
        if terminal(s'):
              break
        s = s'
        a = a'
     if Δ < θ
        break
```

Q Learning

Off Policy

1.- Tengo el estado s

② Selecciono la acción a con $\pi^*_\varepsilon(s)$

③: $\hat{Q}(s,a)$

4.- Aplico a al sistema y obtengo s y r

5- $\hat{Q}(s,a|s',a') = r + \gamma \max_{a' \in A(s')} Q(s',a')$

6- $\delta = \hat{Q}(s,a|s',a') - \hat{Q}(s,a)$

7- $\hat{Q}(s,a) = \hat{Q}(s,a) + \alpha\delta$

8- $s = s'$, $a = a'$

9.- Fuga a ② pa