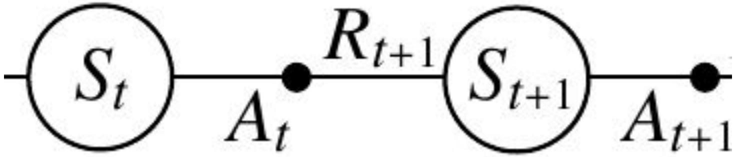
Aprendizaje Reforzado
Clase 6

Julián Martínez

Sarsa: TD con on-policy control



Sarsa

$$\hat{q}^{t+1}(S_t, A_t) = \hat{q}^{t}(S_t, A_t) + \alpha\left[R_{t+1} + \gamma\, \hat{q}^{t}(S_{t+1}, A_{t+1}) - \hat{q}^{t}(S_t, A_t)\right]$$

$$A_{t+1} \sim \pi_{\hat{q}^t} - \varepsilon \text{ greedy}$$

Pseudo-código del sutton

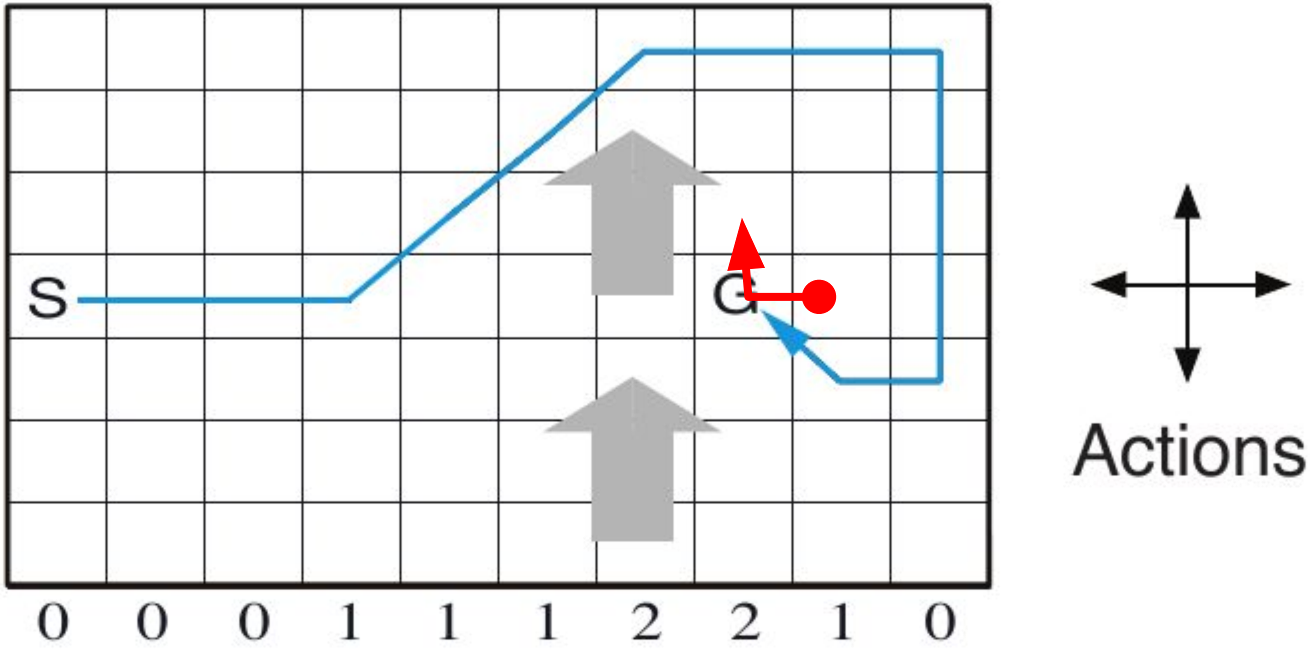
Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\mathit{terminal}, \cdot) = 0$

Loop for each episode:
    Initialize S
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Loop for each step of episode:
        Take action A, observe R, S'
        Choose A' from S' using policy derived from Q (e.g., ε-greedy)
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
        $S \leftarrow S'$; $A \leftarrow A'$;
    until S is terminal

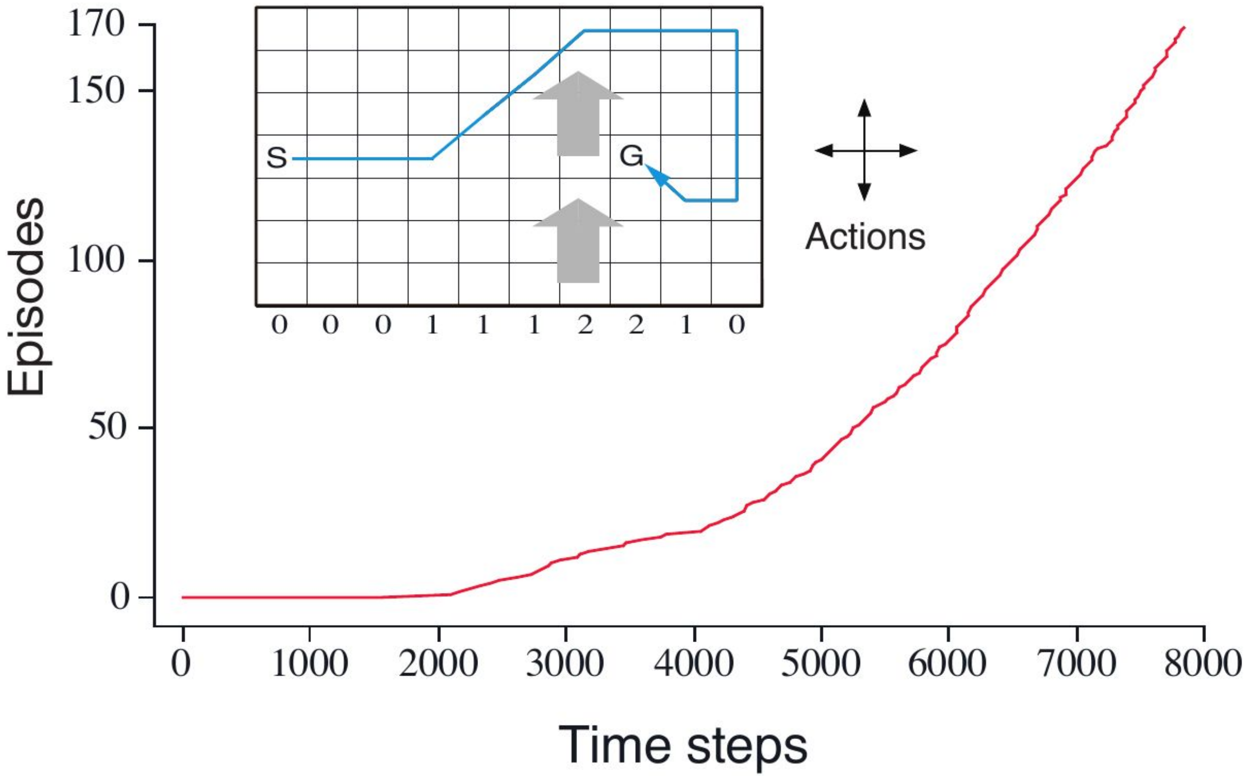
Windy GridWorld

El reward es -1
hasta llegar a G



Intensidad del viento

ε - GREEDY SARSA, ε = 0.1, α=0.5



Q-learning: TD con off-policy control

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \Big].$$

Q-learning: TD con off-policy control

$$\hat{q}^{t+1}(S_t, A_t) = \hat{q}^t(S_t, A_t) + \alpha\left[R_{t+1} + \gamma\, \hat{q}^t(S_{t+1}, A^*_{t+1}) - \hat{q}^t(S_t, A_t)\right]$$

$$A^*_{t+1} = \pi^*_{\hat{q}^t}(S_{t+1}) \quad ; \quad A_{t+1} \sim \pi^{*,\varepsilon}_{\hat{q}^t}(\cdot \mid S_{t+1})$$

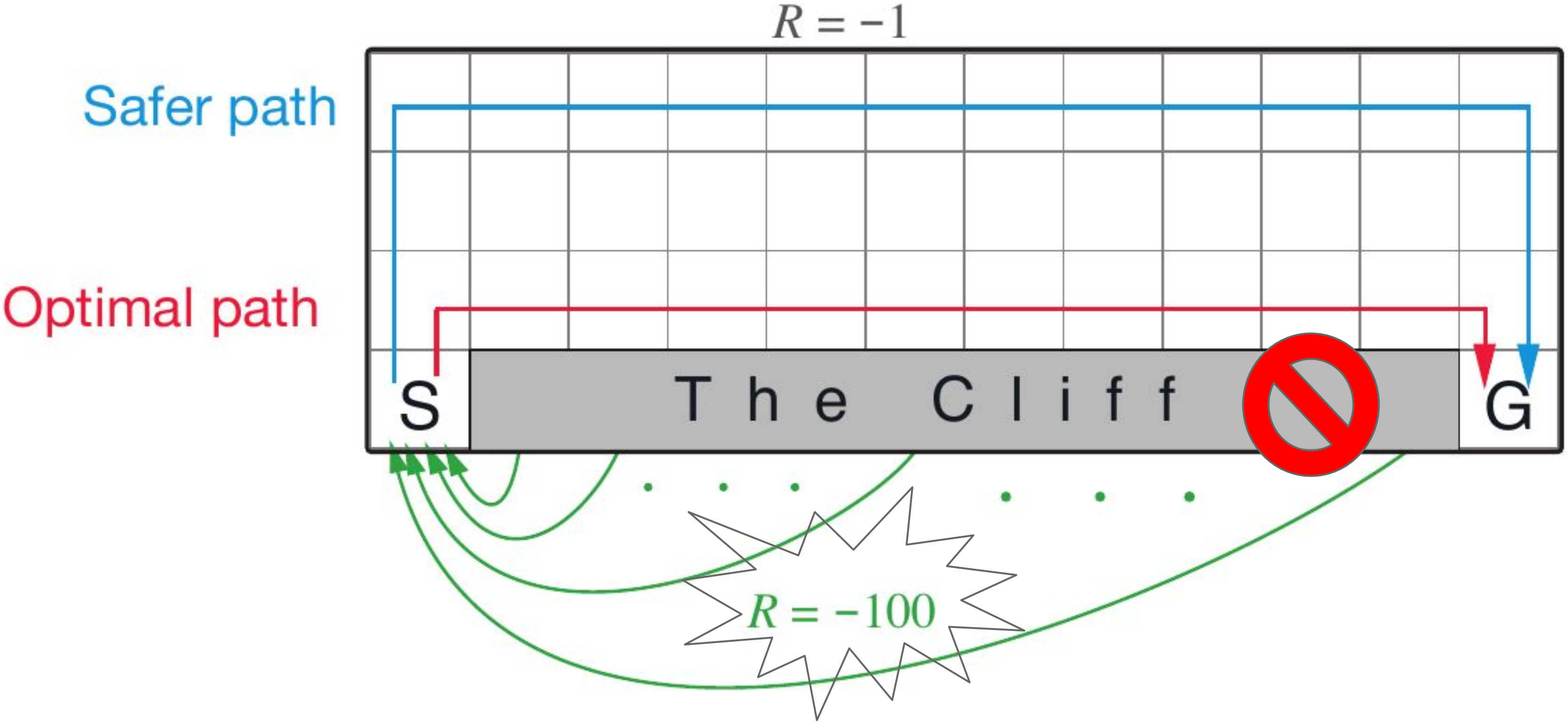
Pseudo-código del sutton

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

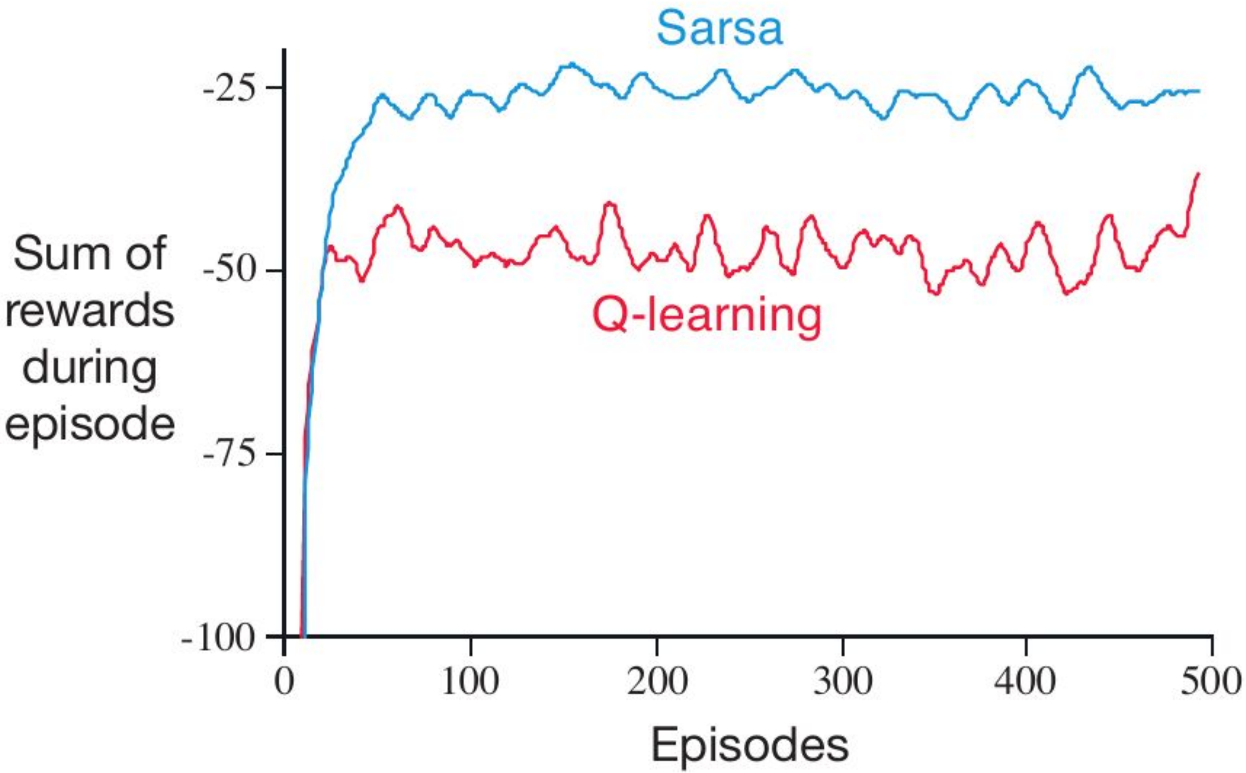
Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Ejemplo 6.6 - Sutton (Acantilado Walking)



Sarsa vs Q-learning



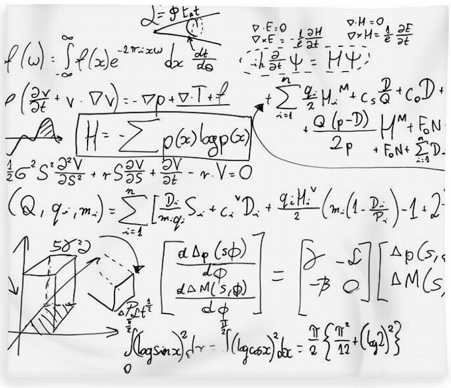
¿Por qué es off-Policy?

$$q_*(s,a) = \mathbb{E}_{\pi'_*}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \mathbb{E}_{\pi_*}\left[R_{t+1} + \gamma\, q_{\pi_*}(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \mathbb{E}_{\pi^{\varepsilon}_*}\left[\left\{R_{t+1} + \gamma\, q_{\pi_*}(S_{t+1}, A_{t+1})\right\} w_{t+1:T} \mid S_t = s, A_t = a\right]$$

$$\omega(s', a') = \frac{\mathbb{1}_{\{\pi^*(s') = a'\}}}{(1-\varepsilon)\,\mathbb{1}_{\{\pi^*(s') = a'\}} + \varepsilon \cdot \frac{1}{|A|}}$$



Resumiendo

$$\hat{q}^{t+1}(S_t, A_t) = \hat{q}^t(S_t, A_t) + \alpha\left[R_{t+1} + \gamma\, \hat{q}^t(S_{t+1}, A^*_{t+1}) - \hat{q}^t(S_t, A_t)\right]$$

$$A^*_{t+1} = \pi^*_{\hat{q}^t}(S_{t+1}) \quad ; \quad A_{t+1} \sim \pi^{*,\varepsilon}_{\hat{q}^t}(\cdot \mid S_{t+1})$$

$$= \mathbb{E}_{\pi^\varepsilon_*}\left[\left\{R_{t+1} + \gamma\, q_{\pi_*}(S_{t+1}, A_{t+1})\right\} w(S_{t+1}, A_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

$$w(s', a') = \frac{\mathbb{1}_{\{\pi^*(s') = a'\}}}{(1-\varepsilon)\,\mathbb{1}_{\{\pi^*(s') = a'\}} + \varepsilon \cdot \frac{1}{|\mathcal{A}|}}$$

Stochastic Approximation

$$E\left[f(\theta, w)\right]\Big|_{\theta=\theta^*} = 0$$

Robbins-Monro, 1951

$$\theta(n+1) = \theta(n) + \alpha_n f\big(\theta(n), w(n+1)\big)$$

$$\sum \alpha_n = \infty \qquad \sum \alpha_n^2 < \infty \qquad \alpha_n = \frac{1}{n}$$

ODES / Método de euler

$$\bar{f}(\theta) = E[f(\theta, w)]$$

$$\frac{d\,x(t)}{dt} = \bar{f}(x(t))$$

$$\frac{\theta(n+1) - \theta(n)}{1/n} = \bar{f}(\theta(n))$$

$$\theta(n+1) = \theta(n) + \alpha_n \left[\bar{f}(\theta(n)) + \Delta(n+1) \right]$$

$$\Delta(n+1) = f(\theta(n), W(n+1)) - \bar{f}(\theta(n))$$

<u>THM</u> If the ODE has a unique asymptotically stable equilibrium x^* $\Rightarrow$ $x_n \to x^*$ with probability one

$$P\left(\lim_{n\to\infty} x_n = x^*\right) = 1$$

Ejemplo 1 - Monte Carlo

$$E[c(X)] \stackrel{?}{=}$$

$$f(\theta, x) = c(x) - \theta$$

$$\theta_{n+1} = \theta_n + \alpha_n \left[c(X_{n+1}) - \theta_n \right]$$

Ejemplo 2 - Ajuste

$$F(\theta) := \frac{1}{2} E\left[\left(Y - f_\theta(X)\right)^2\right]$$

$$\nabla_\theta F(\theta) = E\left[\left(Y - f_\theta(X)\right) \cdot \nabla_\theta f_\theta(X)\right]$$

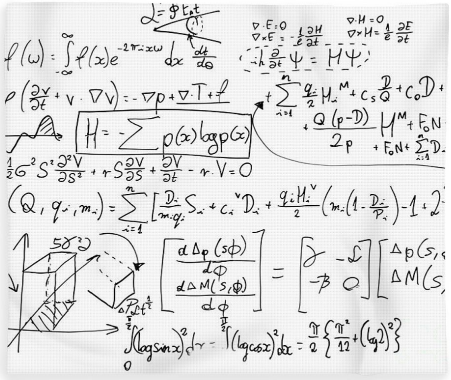
$$\theta_{n+1} = \theta_n + \alpha_n \left[\left(Y_n - f_{\theta_n}(X_n)\right) \nabla_\theta f_{\theta_n}(X_n)\right]$$

Ejemplo 3 - Temporal Difference

$$E\left[R_s^+ + \gamma \sum_{s'} v(s')\, \mathbb{1}_{\{S_s^+ = s'\}} - v(s)\right] = 0 \quad \forall s$$

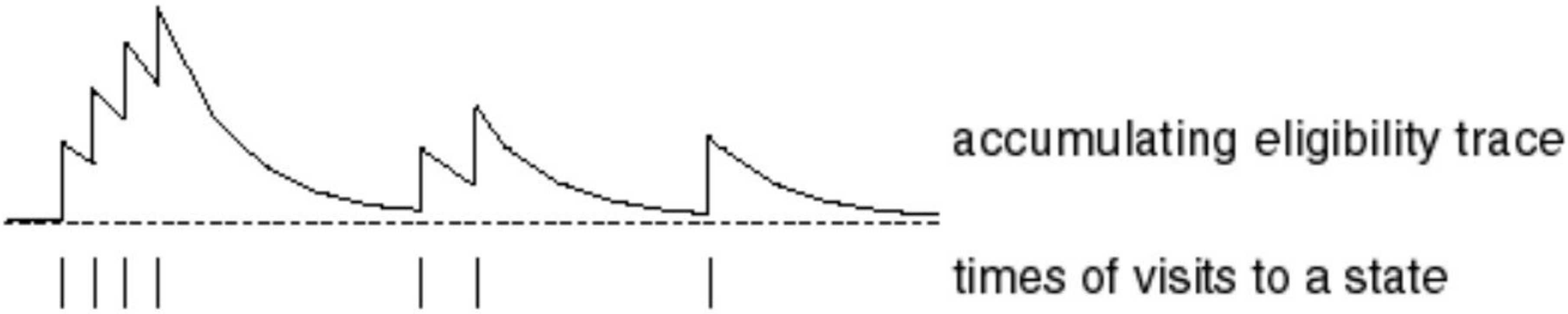
$$v_{n+1} = v_n + \alpha\left[R_{n+1} + \gamma\, v_n(S_{n+1}) - v_n(S_n)\right]$$

Eligibility traces



$$E_0(s) = 0$$

$$E_t(s) = \gamma \lambda E_{t-1}(s) + \mathbf{1}(S_t = s)$$



El lado oscuro...

- Artículo de divulgación sobre Stochastic Approximation https://www.ias.ac.in/article/fulltext/reso/018/12/1086-1094
- THE O.D.E. METHOD FOR CONVERGENCE OF STOCHASTICAPPROXIMATION AND REINFORCEMENT LEARNING
  http://repository.ias.ac.in/5333/1/351.pdf
- Reinforcement Learning: Hidden Theory and New Super-Fast Algorithms (Charla de Meyn sobre Stochastic Approximation en el Simons Institute) https://www.youtube.com/watch?v=dhEF5pfYmvc


TENGO HASTA AHÍ...
GENERADORMEMES.COM