Control on-policy con Aproximación de Funciones

Fernando Lozano

Universidad de los Andes

2 de mayo de 2022

Control con Semi gradientes

Control con Semi gradientes

- Aproximar q_*

Control con Semi gradientes

- Aproximar $q_* \approx \hat{q}(s, a, \mathbf{w})$

Control con Semi gradientes

- Aproximar $q_* \approx \hat{q}(s, a, \mathbf{w})$
- Target U_t (e.g. G_t, $G_{t:t+n}$).

Control con Semi gradientes

- Aproximar $q_* \approx \hat{q}(s, a, \mathbf{w})$
- Target U_t (e.g. G_t, $G_{t:t+n}$).
- En predicción:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[U_t - \hat{q}(S_t, A_t, \mathbf{w}_t) \right] \nabla_{\mathbf{w}} \hat{q}(S_t, A_t, \mathbf{w}_t)$$

Control con Semi gradientes

- Aproximar $q_* \approx \hat{q}(s, a, \mathbf{w})$
- Target U_t (e.g. G_t, $G_{t:t+n}$).
- En predicción:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[U_t - \hat{q}(S_t, A_t, \mathbf{w}_t)\right] \nabla_{\mathbf{w}} \hat{q}(S_t, A_t, \mathbf{w}_t)$$

- SARSA de un paso:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[R_t + \gamma \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t)\right] \nabla_{\mathbf{w}} \hat{q}(S_t, A_t, \mathbf{w}_t$$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
Inicialice $\mathbf{w}$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat $\triangleright$ para cada episodio

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $\mathbf{w}$
  repeat $\triangleright$ para cada episodio
    Inicialice S

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0,1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q $(\epsilon - \text{greedy})$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0,1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon-$greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w}) \right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inciaiice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla\hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inciale $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
      $S \leftarrow S'$, $A \leftarrow A'$

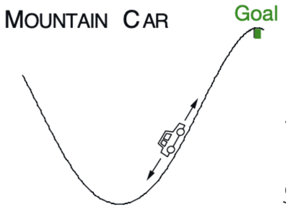
semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
      $S \leftarrow S'$, $A \leftarrow A'$
    until S es terminal

semi gradiente SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      if S' es terminal **then**
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
        break ▷ nuevo episodio
      end if
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
      $S \leftarrow S'$, $A \leftarrow A'$
    until S es terminal
  until ∞

Ejemplo: Mountain Car



- Acciones $A_t \in \{-1, 0, 1\}$, recompensa -1 en cada paso.
- Dinámica:

$$x_{t+1} \doteq \left[x_t + \dot{x}_{t+1}\right]\Big|_{-1,2}^{0,5}$$

$$\dot{x}_{t+1} \doteq \left[\dot{x} + 0{,}001 A_t - 0{,}0025\cos(3x_t)\right]\Big|_{-0,07}^{0,07}$$

si $x_{t+1} = -1{,}2 \Rightarrow \dot{x}_{t+1} = 0$

- 8 Tiles con ancho $\frac{1}{8}\times$rango.

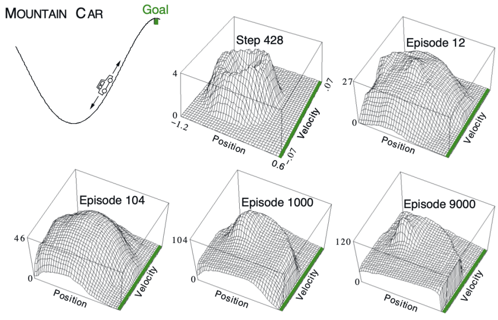
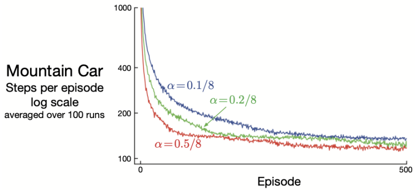
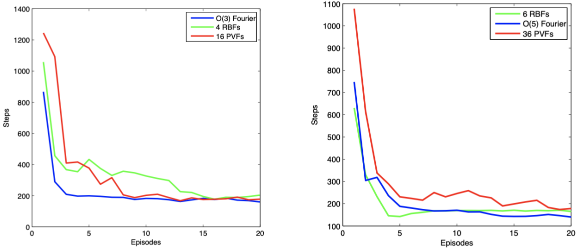


Figure 10.1: The Mountain Car task (upper left panel) and the cost-to-go function $(-\max_a \hat{q}(s, a, \mathbf{w}))$ learned during one run.

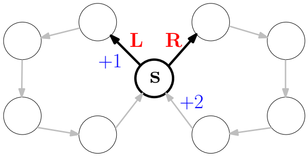




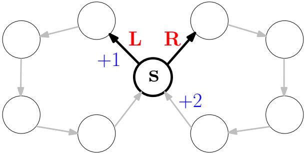
(a) (b)

Figure 7: Learning curves for agents using (a) order 3 (b) order 5 Fourier Bases, and RBFs and PVFs with corresponding number of basis functions.

Descuento en tareas contínuas

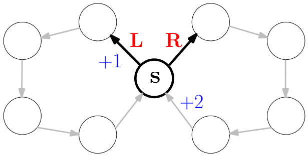


Descuento en tareas contínuas



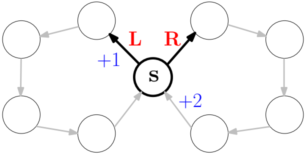
- 2 políticas determinísticas: Escoger **L**, escoger **R**

Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

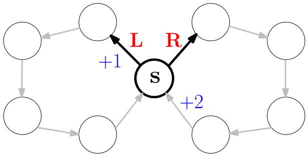
Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) =$$

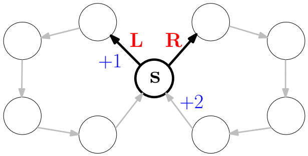
Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) = \frac{1}{1-\gamma^5},$$

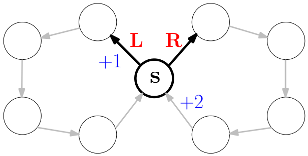
Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) = \frac{1}{1-\gamma^5}, \quad v_{\mathbf{R}}(s) =$$

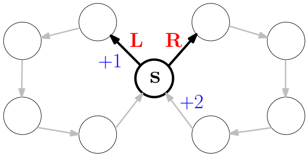
Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) = \frac{1}{1-\gamma^5}, \quad v_{\mathbf{R}}(s) = \frac{2\lambda^4}{1-\gamma^5}$$

Descuento en tareas contínuas

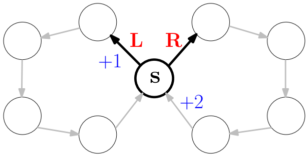


- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) = \frac{1}{1-\gamma^5}, \quad v_{\mathbf{R}}(s) = \frac{2\lambda^4}{1-\gamma^5}$$

- $v_{\mathbf{R}}(s) > v_{\mathbf{L}}(s)$ para $\gamma > 0{,}841$

Descuento en tareas contínuas



- 2 políticas determinísticas: Escoger **L**, escoger **R**
- Usando descuento γ:

$$v_{\mathbf{L}}(s) = \frac{1}{1-\gamma^5}, \quad v_{\mathbf{R}}(s) = \frac{2\lambda^4}{1-\gamma^5}$$

- $v_{\mathbf{R}}(s) > v_{\mathbf{L}}(s)$ para $\gamma > 0{,}841$
- Con 97 estados $\gamma > 0{,}993$!!

Recompensa promedio

$$r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right]$$

Recompensa promedio

$$
\begin{aligned}
r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h}\sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right] \\
&= \sum_{s} \mu_\pi(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r
\end{aligned}
$$

Recompensa promedio

$$r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \ : \ S_0, A_{0:t-1} \sim \pi\right]$$
$$= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r$$

- $\mu_\pi(s)$ es la distribución de estado estable, independiente de S_0:

$$\mu_\pi(s) = \lim_{t\to\infty} \mathbf{P}\left\{S_t = s \mid A_{0:t-1} \sim \pi\right\}$$

(MDP ergódico)

Recompensa promedio

$$r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \ : \ S_0, A_{0:t-1} \sim \pi\right]$$
$$= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r$$

- $\mu_\pi(s)$ es la distribución de estado estable, independiente de S_0:

$$\mu_\pi(s) = \lim_{t\to\infty} \mathbf{P}\left\{S_t = s \mid A_{0:t-1} \sim \pi\right\}$$

(MDP ergódico)

- Satisface:

$$\sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) = \mu_\pi(s')$$

Descuento?

$$J(\pi) = \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}}$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right]
\end{aligned}
$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s')\right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s')
\end{aligned}
$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s)\\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a)
\end{aligned}
$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s')
\end{aligned}
$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s')\right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s') \\
&= r(\pi) + \gamma J(\pi)
\end{aligned}
$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s)\, \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s)\\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s') \\
&= r(\pi) + \gamma J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 J(\pi) \\
&\vdots
\end{aligned}
$$

Descuento?

$$\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s') \\
&= r(\pi) + \gamma J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 J(\pi) \\
&\vdots \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 r(\pi) + \gamma^3 r(\pi) + \dots
\end{aligned}$$

Descuento?

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) \overbrace{v_\pi^\gamma(s)}^{\substack{v_\pi(s) \\ \text{con } \gamma}} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi^\gamma(s') \right] \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \gamma v_\pi^\gamma(s') \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s') \\
&= r(\pi) + \gamma J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 J(\pi) \\
&\vdots \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 r(\pi) + \gamma^3 r(\pi) + \ldots = \frac{1}{1-\gamma} r(\pi)
\end{aligned}
$$

$$r(\pi_1) < r(\pi_2) \Rightarrow J(\pi_1) < J(\pi_2)$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \dots$$

- Función de valor:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Función de valor:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

- Ecuaciones de Bellman:

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

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \dots$$

- Función de valor:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

- Ecuaciones de Bellman:

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

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

- Error TD:

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)$$

- Error TD:

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, A_t, \mathbf{w}_t)$$

- Error TD:

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, A_t, \mathbf{w}_t)$$

- $\bar{R}_t$: Estimativo de $r(\pi)$ en tiempo t.

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
Incialice $\mathbf{w}$

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Incialice **w**
  repeat $\triangleright$ para cada paso

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Incialice $\mathbf{w}$
  repeat $\triangleright$ para cada paso
    Tome acción A, observe R, S'.

semi gradiente diferencial SARSA

Require: Función $\hat{q}\,:\,\mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ (ϵ − greedy)
    $\delta \leftarrow R - \bar{R} + \hat{q}(s', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
    $\delta \leftarrow R - \bar{R} + \hat{q}(s', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
    $\bar{R} \leftarrow \bar{R} + \beta\delta$

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
    $\delta \leftarrow R - \bar{R} + \hat{q}(s', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
    $\bar{R} \leftarrow \bar{R} + \beta\delta$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$

semi gradiente diferencial SARSA

Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ ($\epsilon -$ greedy)
    $\delta \leftarrow R - \bar{R} + \hat{q}(s', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
    $\bar{R} \leftarrow \bar{R} + \beta\delta$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$
    $S \leftarrow S'$, $A \leftarrow A'$

semi gradiente diferencial SARSA

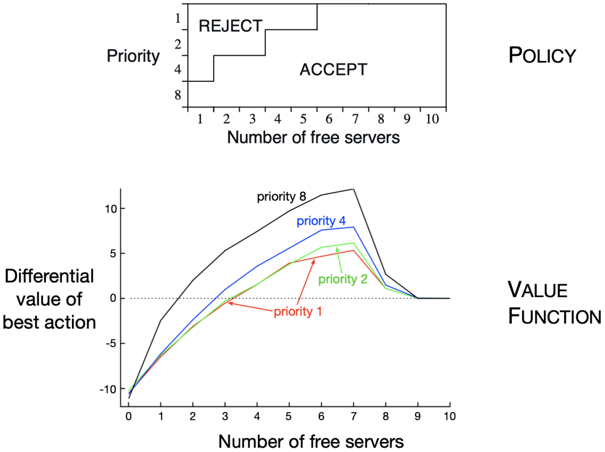
Require: Función $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Require: $\alpha, \beta \in (0, 1]$, $\epsilon > 0$
  Inicialice $\mathbf{w}$
  repeat ▷ para cada paso
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a $\hat{q}$ $(\epsilon - \text{greedy})$
    $\delta \leftarrow R - \bar{R} + \hat{q}(s', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
    $\bar{R} \leftarrow \bar{R} + \beta\delta$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$
    $S \leftarrow S'$, $A \leftarrow A'$
  until ∞

Ejemplo

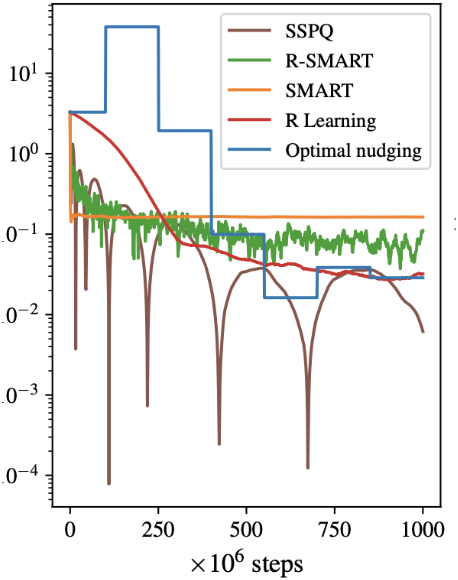
- 10 Servidores atienden cola de clientes con pripridades 1,2,...8.
- Recompensa $\propto$ prioridad.
- Aceptar o rechazar cliente.

Ejemplo

- 10 Servidores atienden cola de clientes con pripridades 1,2,...8.
- Recompensa $\propto$ prioridad.
- Aceptar o rechazar cliente.



Nudging



SSPQ
R-SMART
SMART
R Learning
Optimal nudging
10^1
10^0
0^{-1}
0^{-2}
0^{-3}
0^{-4}
0
250
500
750
1000
$\times 10^6$ steps