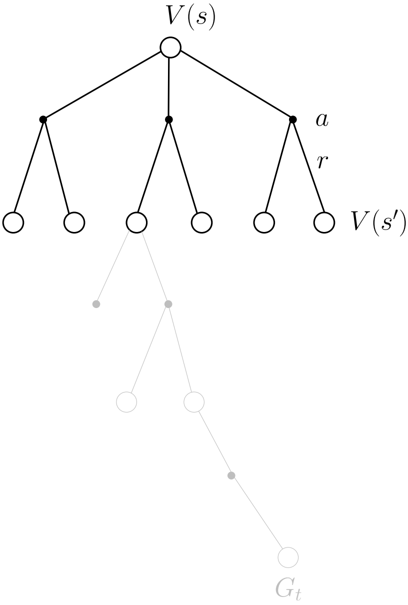
Aprendizaje por Diferencias Temporales

Fernando Lozano

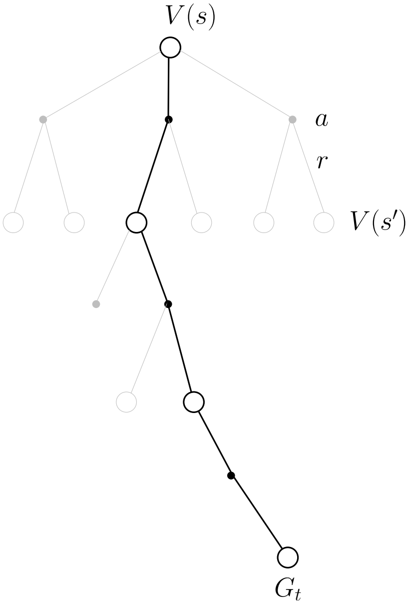
Universidad de los Andes

28 de febrero de 2023

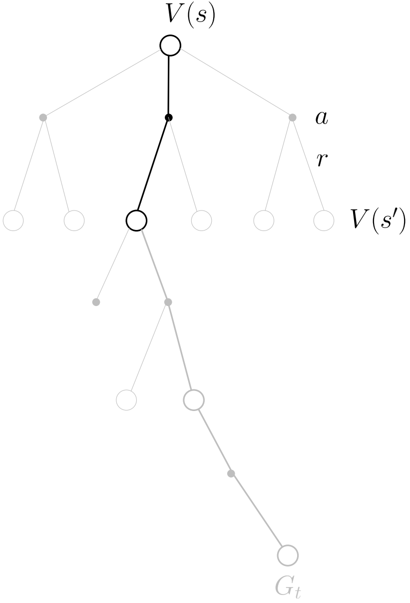
- Programación Dinámica.



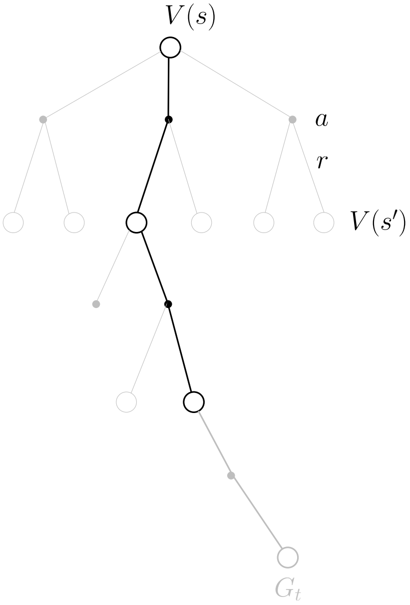
- Monte Carlo.



- Diferencias temporales ($TD(0)$).



- Diferencias temporales ($TD(\lambda)$).



Predicción

Predicción

- Actualización de Montecarlo:

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t)\right]$$

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(s_t)\right]$$

  - mover estimativo $V(S_t)$ hacia G_t.

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t)\right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi \left\{G_t \mid S_t = s\right\}$

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t) \right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi \left\{ G_t \mid S_t = s \right\}$
- TD: no esperar hasta el final del episodio,

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(s_t)\right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\}$

- TD: no esperar hasta el final del episodio,

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]$$

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t) \right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi \left\{ G_t \mid S_t = s \right\}$

- TD: no esperar hasta el final del episodio,

$$V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right]$$

  - mover estimativo $V(S_t)$ hacia $R_{t+1} + \gamma V(S_{t+1})$.

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t) \right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi \left\{ G_t \mid S_t = s \right\}$
- TD: no esperar hasta el final del episodio,

$$V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right]$$

  - mover estimativo $V(S_t)$ hacia $R_{t+1} + \gamma V(S_{t+1})$.
- Estimativo basado en estimativo: bootstrapping

Predicción

- Actualización de Montecarlo:

$$V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(s_t) \right]$$

  - mover estimativo $V(S_t)$ hacia G_t.
  - G_t es estimativo de $v_\pi(s) = \mathbb{E}_\pi \left\{ G_t \mid S_t = s \right\}$

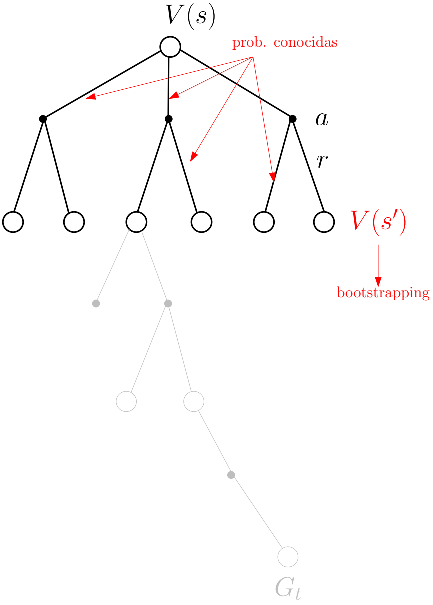
- TD: no esperar hasta el final del episodio,

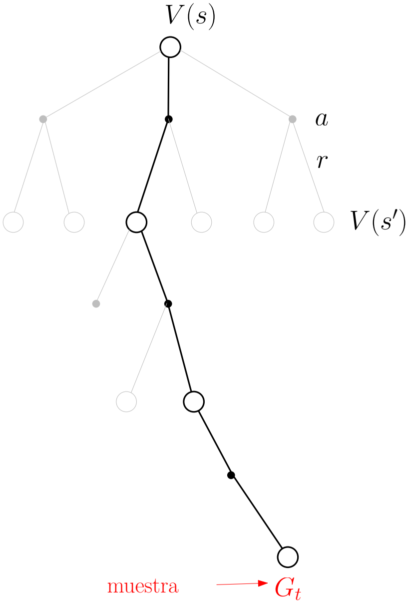
$$V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right]$$

  - mover estimativo $V(S_t)$ hacia $R_{t+1} + \gamma V(S_{t+1})$.

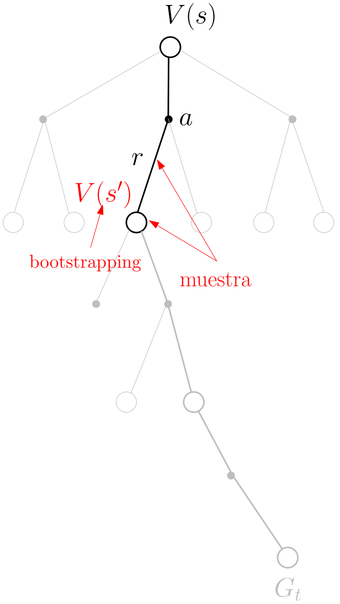
- Estimativo basado en estimativo: bootstrapping

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi \left\{ G_t \mid S_t = s \right\} \\ &= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right\} \end{aligned}$$


$V(s)$
prob. conocidas
a
r
$V(s')$
bootstrapping
G_t



$V(s)$
a
r
$V(s')$
muestra
G_t


$V(s)$
a
r
$V(s')$
bootstrapping
muestra
G_t

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$
Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$
Inicialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$
repeat ▷ para cada episodio

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

  Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

  repeat ▷ para cada episodio

    Inicialice S

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

repeat $\triangleright$ para cada episodio

  Inicialice S

  repeat $\triangleright$ para cada paso del episodio

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$
  Inicialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

$A \leftarrow$ acción dada por π en S

Tome acción A, observe R, y nuevo estado S'

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

repeat ▷ para cada episodio

  Inicialice S

  repeat ▷ para cada paso del episodio

    $A \leftarrow$ acción dada por π en S

    Tome acción A, observe R, y nuevo estado S'

    $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$
  Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S
      Tome acción A, observe R, y nuevo estado S'
      $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$
      $S \leftarrow S'$

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Incialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

repeat ▷ para cada episodio

    Inicialice S

    repeat ▷ para cada paso del episodio

        $A \leftarrow$ acción dada por π en S

        Tome acción A, observe R, y nuevo estado S'

        $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$

        $S \leftarrow S'$

    until S es terminal

Evaluación de política con TD(0) tabular

Require: Política π, tamaño de paso $\alpha \in (0, 1]$

Inicialice $V(s)\ \forall s \in \mathcal{S}^+$, $V(s_{\text{terminal}}) = 0$

repeat $\triangleright$ para cada episodio

  Inicialice S

  repeat $\triangleright$ para cada paso del episodio

    $A \leftarrow$ acción dada por π en S

    Tome acción A, observe R, y nuevo estado S'

    $V(S) \leftarrow V(S) + \alpha\,[R + \gamma V(S') - V(S)]$

    $S \leftarrow S'$

  until S es terminal

until ∞

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$G_t - V(S_t) = R_{t+1} + \gamma G_{t+1} - V(S_t)$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$G_t - V(S_t) = R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1})$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned} G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\ &= \delta_t + \gamma (G_{t+1} - V(S_{t+1})) \end{aligned}$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned} G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\ &= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2(G_{t+2} - V(S_{t+2})) \end{aligned}$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned} G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\ &= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2(G_{t+2} - V(S_{t+2})) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + +\gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(G_T - V(S_T)) \end{aligned}$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma (G_{t+1} - V(S_{t+1})) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 (G_{t+2} - V(S_{t+2})) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \cdots + + \gamma^{T-t-1} \delta_{T-1} + \gamma^{T-t} (G_T - V(S_T)) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \delta_{t+2} + \cdots + + \gamma^{T-t-1} \delta_{T-1} + \gamma^{T-t} (0 - 0)
\end{aligned}$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2(G_{t+2} - V(S_{t+2})) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + +\gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(G_T - V(S_T)) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + +\gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(0 - 0) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k
\end{aligned}$$

Error TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Error en estimativo de $V(S_t)$ disponible en $t+1$.
- En términos del error de Montecarlo:

$$\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2(G_{t+2} - V(S_{t+2})) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + +\gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(G_T - V(S_T)) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + +\gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(0 - 0) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k
\end{aligned}$$

(si V no cambia durante el episodio)

Ejemplo

State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

Ejemplo

State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

- Recompensa: Tiempo gastado en trayecto.

Ejemplo

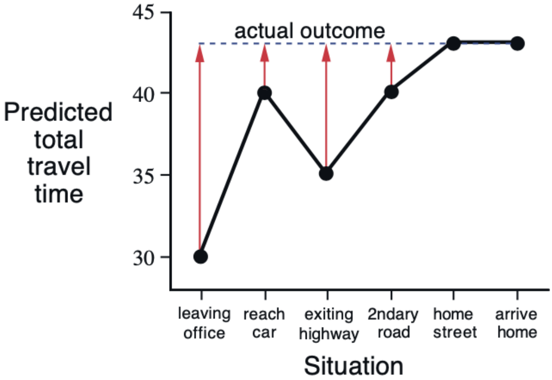
State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

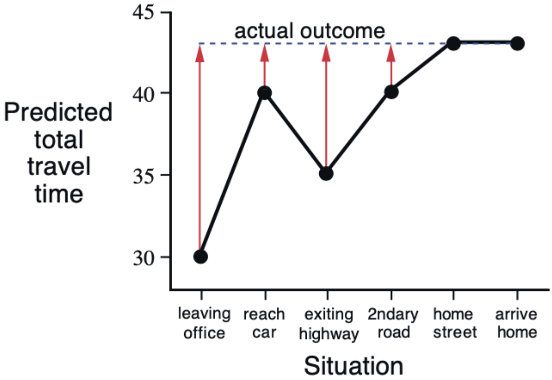
- Recompensa: Tiempo gastado en trayecto.
- Retorno G_t: Tiempo real faltante desde estado S_t.

Ejemplo

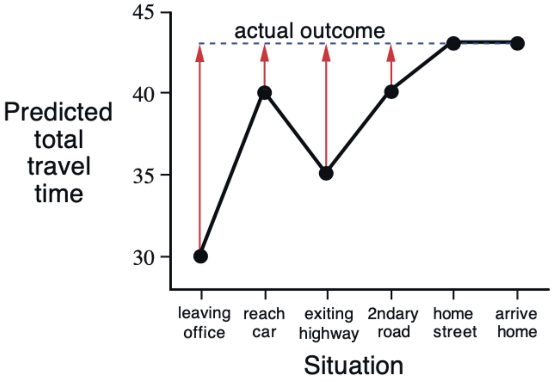
State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

- Recompensa: Tiempo gastado en trayecto.
- Retorno G_t: Tiempo real faltante desde estado S_t.

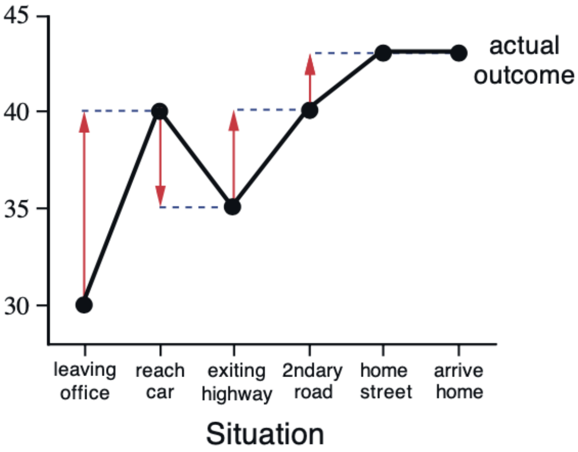

actual outcome
Predicted total travel time
45
40
35
30
leaving office
reach car
exiting highway
2ndary road
home street
arrive home
Situation

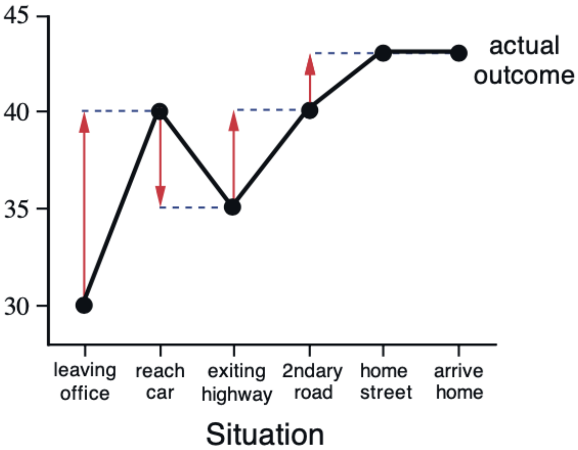


- Actualización de $\alpha-$Montecarlo: $G_t - V(S_t)$ $(\alpha = 1)$

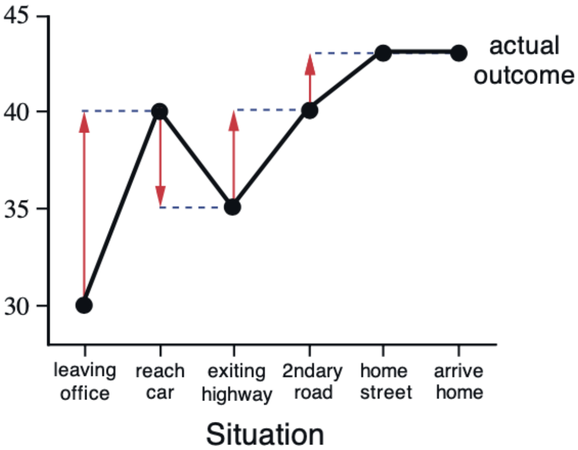


- Actualización de $\alpha-$Montecarlo: $G_t - V(S_t)$ $(\alpha = 1)$
- Qué cambia si $\alpha = \frac{1}{2}$?

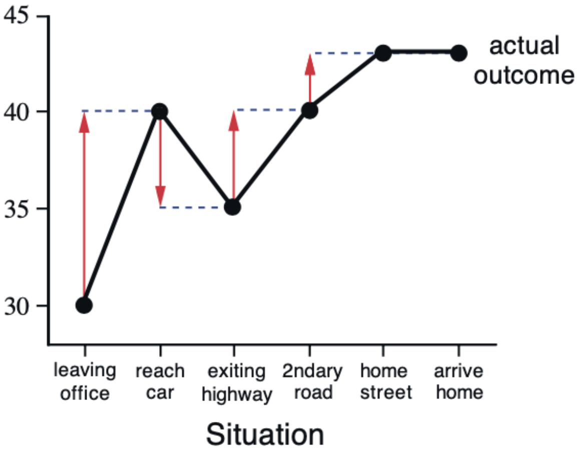

45
40
35
30
actual
outcome
leaving
office
reach
car
exiting
highway
2ndary
road
home
street
arrive
home
Situation



- Actualización de $TD(0)$: $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ $(\alpha = 1)$



- Actualización de $TD(0)$: $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ $(\alpha = 1)$
- Error proporcional al cambio en el tiempo de la predicción: δ_t.



- Actualización de $TD(0)$: $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ $(\alpha = 1)$
- Error proporcional al cambio en el tiempo de la predicción: δ_t.
- Qué sucede si cambia parcialmente la ruta?

Ventajas de TD

Ventajas de TD

- No requiere conocimiento del MDP.

Ventajas de TD

- No requiere conocimiento del MDP.
- No espera hasta el final del episodio para actualizar.

Ventajas de TD

- No requiere conocimiento del MDP.
- No espera hasta el final del episodio para actualizar.
- Aplicable en tareas no episódicas o con episodios muy largos.

Ventajas de TD

- No requiere conocimiento del MDP.
- No espera hasta el final del episodio para actualizar.
- Aplicable en tareas no episódicas o con episodios muy largos.
- Implementación incremental/en línea.

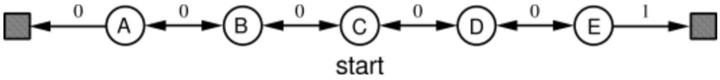
Ventajas de TD

- No requiere conocimiento del MDP.
- No espera hasta el final del episodio para actualizar.
- Aplicable en tareas no episódicas o con episodios muy largos.
- Implementación incremental/en línea.
- Convergencia (asimptótica) para cualquier política en el caso tabular.

Ventajas de TD

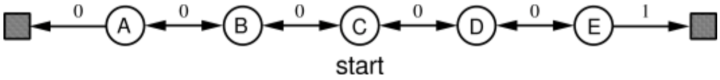
- No requiere conocimiento del MDP.
- No espera hasta el final del episodio para actualizar.
- Aplicable en tareas no episódicas o con episodios muy largos.
- Implementación incremental/en línea.
- Convergencia (asimptótica) para cualquier política en el caso tabular.
- En la práctica convergencia más rápida que MC.

Ejemplo: Random Walk



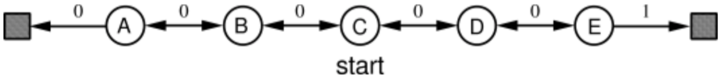
Ejemplo: Random Walk

Ejemplo: Random Walk



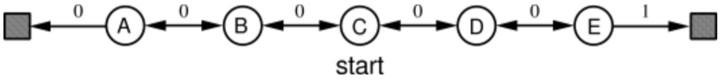
- Markov Reward Process (no hay acciones).

Ejemplo: Random Walk



- Markov Reward Process (no hay acciones).
- Comenzando en 0, se mueve a la izquierda o derecha con probabilidad $\frac{1}{2}$.

Ejemplo: Random Walk

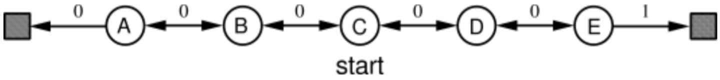


- Markov Reward Process (no hay acciones).
- Comenzando en 0, se mueve a la izquierda o derecha con probabilidad $\frac{1}{2}$.
- Recompensa 1 en el estado terminal de la derecha, 0 en otro caso.

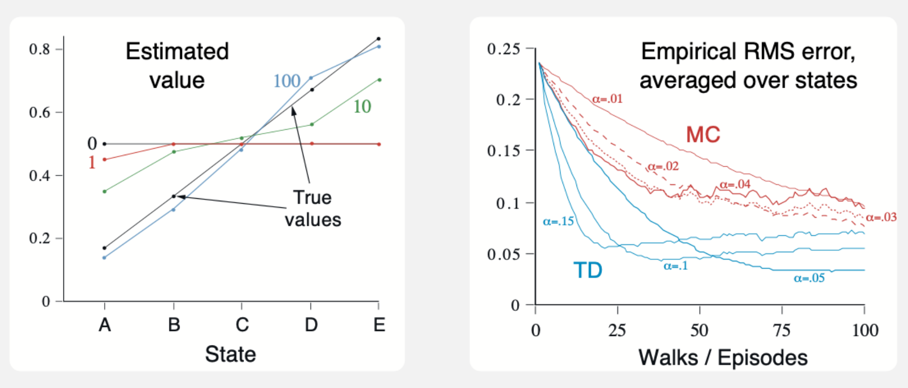
Ejemplo: Random Walk

- Markov Reward Process (no hay acciones).
- Comenzando en 0, se mueve a la izquierda o derecha con probabilidad $\frac{1}{2}$.
- Recompensa 1 en el estado terminal de la derecha, 0 en otro caso.
- $v_\pi(s)$ es igual a

Ejemplo: Random Walk



- Markov Reward Process (no hay acciones).
- Comenzando en 0, se mueve a la izquierda o derecha con probabilidad $\frac{1}{2}$.
- Recompensa 1 en el estado terminal de la derecha, 0 en otro caso.
- $v_\pi(s)$ es igual a la probabilidad de terminar a la derecha.


Estimated value
0.8
0.6
0.4
0.2
0
100
10
0
1
True values
A
B
C
D
E
State
Empirical RMS error, averaged over states
0.25
0.2
0.15
0.1
0.05
0
α=.01
MC
α=.02
α=.04
α=.03
α=.15
TD
α=.1
α=.05
0
25
50
75
100
Walks / Episodes

Control de política

- Iteración de política generalizada (GPI).

Control de política

- Iteración de política generalizada (GPI).
- Función de valor de acción $q_\pi(s, a)$ (en lugar de $v_\pi(s)$).

Control de política

- Iteración de política generalizada (GPI).
- Función de valor de acción $q_\pi(s, a)$ (en lugar de $v_\pi(s)$).
- TD para predicción.

Control de política

- Iteración de política generalizada (GPI).
- Función de valor de acción $q_\pi(s, a)$ (en lugar de $v_\pi(s)$).
- TD para predicción.
- Exploración/explotación:

Control de política

- Iteración de política generalizada (GPI).
- Función de valor de acción $q_\pi(s, a)$ (en lugar de $v_\pi(s)$).
- TD para predicción.
- Exploración/explotación: on-policy/off policy.

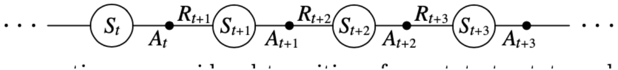
SARSA (On-policy)

SARSA (On-policy)

- Para aprender v_π TD(0) usa transiciones de estado a estado.

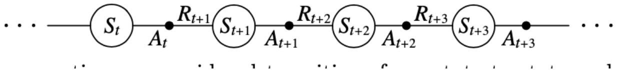
SARSA (On-policy)

- Para aprender v_π TD(0) usa transiciones de estado a estado.
- Para aprender q_π TD(0) usa transiciones entre pares (s, a)



SARSA (On-policy)

- Para aprender v_π TD(0) usa transiciones de estado a estado.
- Para aprender q_π TD(0) usa transiciones entre pares (s, a)



$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$$

con $Q(S_{t+1}, A_{t+1}) = 0$ para S_{t+1} terminal.

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
repeat $\triangleright$ para cada episodio

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)

SARSA

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$
  Inicialice $Q(s,a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q (ϵ − greedy)
    repeat ▷ para cada paso del episodio

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

  Inicialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

  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'.

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  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'.
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a Q ($\epsilon-$greedy)

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q $(\epsilon - \text{greedy})$
    repeat ▷ para cada paso del episodio
      Tome acción A, observe R, S'.
      Escoja A' de $\mathcal{A}(S')$, de acuerdo a Q $(\epsilon - \text{greedy})$
      $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma Q(S', A') - Q(S, A)\right]$

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
Inicialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
repeat ▷ para cada episodio
  Inicialice S
  Escoja A de $\mathcal{A}(S)$, de acuerdo a Q $(\epsilon - \text{greedy})$
  repeat ▷ para cada paso del episodio
    Tome acción A, observe R, S'.
    Escoja A' de $\mathcal{A}(S')$, de acuerdo a Q $(\epsilon - \text{greedy})$
    $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma Q(S', A') - Q(S, A)\right]$
    $S \leftarrow S'$, $A \leftarrow A'$

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

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'.

    Escoja A' de $\mathcal{A}(S')$, de acuerdo a Q ($\epsilon -$ greedy)

    $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma Q(S', A') - Q(S, A)\right]$

    $S \leftarrow S'$, $A \leftarrow A'$

  until S es terminal

SARSA

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

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'.

Escoja A' de $\mathcal{A}(S')$, de acuerdo a Q ($\epsilon -$ greedy)

$Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma Q(S', A') - Q(S, A)\right]$

$S \leftarrow S'$, $A \leftarrow A'$

until S es terminal

until ∞

Q-Learning (Off-Policy)

Q-Learning (Off-Policy)

- En lugar de usar estimativo de Q en el siguiente par estado-acción, usa el mejor valor de Q en ese estado.

Q-Learning (Off-Policy)

- En lugar de usar estimativo de Q en el siguiente par estado-acción, usa el mejor valor de Q en ese estado.

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

Q-Learning (Off-Policy)

- En lugar de usar estimativo de Q en el siguiente par estado-acción, usa el mejor valor de Q en ese estado.

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

- Q aproxima directamente q_*, independientemente de la política que se sigue.

Q-Learning (Off-Policy)

- En lugar de usar estimativo de Q en el siguiente par estado-acción, usa el mejor valor de Q en ese estado.

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

- Q aproxima directamente q_*, independientemente de la política que se sigue.
- Política determina pares estado-acción visitados y actualizados.

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

  Incialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

  repeat ▷ para cada episodio

    Inicialice S

    repeat ▷ para cada paso del episodio

      Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      Escoja A de $\mathcal{A}(S)$, de acuerdo a Q $(\epsilon - \text{greedy})$
      Tome acción A, observe R, S'.

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $Q(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
      Tome acción A, observe R, S'.
      $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_a Q(S', a) - Q(S, A)\right]$

Q-Learning

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$

Inicialice $Q(s,a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

  Inicialice S

  repeat ▷ para cada paso del episodio

    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)

    Tome acción A, observe R, S'.

    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$

    $S \leftarrow S'$

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$

repeat $\triangleright$ para cada episodio

  Inicialice S

  repeat $\triangleright$ para cada paso del episodio

    Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon-$greedy)

    Tome acción A, observe R, S'.

    $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_a Q(S', a) - Q(S, A)\right]$

    $S \leftarrow S'$

  until S es terminal

Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$
  Inicialice $Q(s, a)\ \forall s \in \mathcal{S}^+$, con $Q(s_{\text{terminal}}, .) = 0$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      Escoja A de $\mathcal{A}(S)$, de acuerdo a Q ($\epsilon -$ greedy)
      Tome acción A, observe R, S'.
      $Q(S, A) \leftarrow Q(S, A) + \alpha\,[R + \gamma \max_a Q(S', a) - Q(S, A)]$
      $S \leftarrow S'$
    until S es terminal
  until ∞

On-Policy vs. Off-Policy

On-Policy vs. Off-Policy

- Suponga que se usa política π para comportamiento.

On-Policy vs. Off-Policy

- Suponga que se usa política π para comportamiento.
- SARSA:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, \underbrace{A_{t+1}}_{\pi}) - Q(S_t, A_t) \right]$$

On-Policy vs. Off-Policy

- Suponga que se usa política π para comportamiento.
- SARSA:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, \underbrace{A_{t+1}}_{\pi}) - Q(S_t, A_t) \right]$$

- Q-Learning:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, \underbrace{a}_{\sim\pi_* \neq \pi}) - Q(S_t, A_t) \right]$$

On-Policy vs. Off-Policy

- Suponga que se usa política π para comportamiento.
- SARSA:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, \underbrace{A_{t+1}}_{\pi}) - Q(S_t, A_t) \right]$$

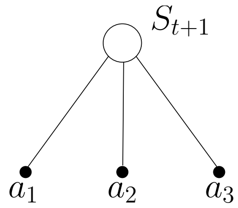
- Q-Learning:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, \underbrace{a}_{\sim\pi_* \neq \pi}) - Q(S_t, A_t) \right]$$

- Q-Learning: cualquier π que explore frecuentemente todos los pares (s, a).

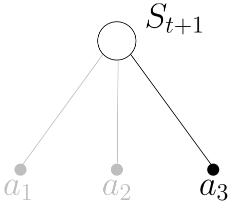
Muestreo por importancia?

Muestreo por importancia?

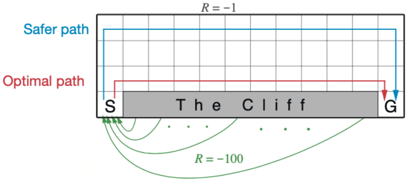


$$\mathbb{E}_{\pi}\left\{G_{t+1} \mid S_{t+1}\right\} = \sum_{a} \pi(a \mid S_{t+1}) Q(S_{t+1}, a)$$

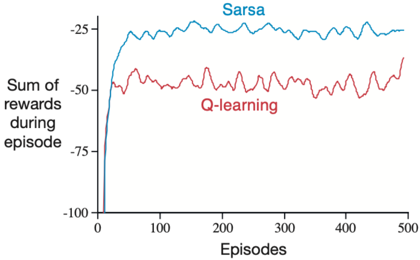
Muestreo por importancia?



$$\mathbb{E}_\pi\left\{G_{t+1} \mid S_{t+1}\right\} = \sum_a \pi(a \mid S_{t+1}) Q(S_{t+1}, a) = \max_a Q(S_{t+1}, a)$$



R = −1
Safer path
Optimal path
S
The Cliff
G
R = −100



Sarsa
Q-learning
Sum of rewards during episode
-25
-50
-75
-100
0
100
200
300
400
500
Episodes

SARSA Esperado

SARSA Esperado

- SARSA:

$$Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma Q(S', A') - Q(S, A) \right]$$

SARSA Esperado

- SARSA:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma Q(S',A') - Q(S,A)\right]$$

- Usar valor esperado con respecto a la política:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma \mathbb{E}_\pi \left\{Q(S',A') \mid S'\right\} - Q(S,A)\right]$$

SARSA Esperado

- SARSA:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma Q(S',A') - Q(S,A)\right]$$

- Usar valor esperado con respecto a la política:

$$\begin{aligned} Q(S,A) &\leftarrow Q(S,A) + \alpha \left[R + \gamma \mathbb{E}_\pi \left\{Q(S',A') \mid S'\right\} - Q(S,A)\right] \\ &\leftarrow Q(S,A) + \alpha \left[R + \gamma \sum_{a \in \mathcal{A}(S')} \pi(a \mid S') Q(S',a) - Q(S,A)\right] \end{aligned}$$

SARSA Esperado

- SARSA:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma Q(S',A') - Q(S,A)\right]$$

- Usar valor esperado con respecto a la política:

$$\begin{aligned} Q(S,A) &\leftarrow Q(S,A) + \alpha \left[R + \gamma \mathbb{E}_\pi \left\{Q(S',A') \mid S'\right\} - Q(S,A)\right] \\ &\leftarrow Q(S,A) + \alpha \left[R + \gamma \sum_{a \in \mathcal{A}(S')} \pi(a \mid S') Q(S',a) - Q(S,A)\right] \end{aligned}$$

- Actualización es determinística dado el siguiente estado S'.

SARSA Esperado

- SARSA:

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q(S',A') - Q(S,A)\right]$$

- Usar valor esperado con respecto a la política:

$$\begin{aligned} Q(S,A) &\leftarrow Q(S,A) + \alpha\left[R + \gamma \mathbb{E}_\pi\left\{Q(S',A') \mid S'\right\} - Q(S,A)\right] \\ &\leftarrow Q(S,A) + \alpha\left[R + \gamma \sum_{a\in\mathcal{A}(S')} \pi(a \mid S')Q(S',a) - Q(S,A)\right] \end{aligned}$$

- Actualización es determinística dado el siguiente estado S'.
- Elimina varianza debida a selección de acción en S'.

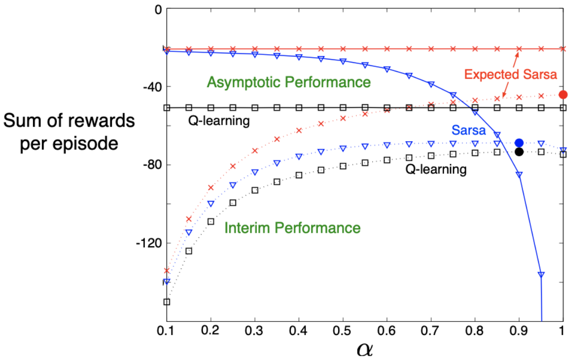


Figure 6.3: Interim and asymptotic performance of TD control methods on the cliff-walking task as a function of α. All algorithms used an ε-greedy policy with $\varepsilon = 0.1$. Asymptotic performance is an average over 100,000 episodes whereas interim performance is an average over the first 100 episodes. These data are averages of over 50,000 and 10 runs for the interim and asymptotic cases respectively. The solid circles mark the best interim performance of each method. Adapted from van Seijen et al. (2009).

Sesgo de Maximización

- Algoritmos de TD usan maximización sobre acciones en un estado:

Sesgo de Maximización

- Algoritmos de TD usan maximización sobre acciones en un estado:
  - Q-Learning actualiza usando maximización sobre acciones.

Sesgo de Maximización

- Algoritmos de TD usan maximización sobre acciones en un estado:
  - Q-Learning actualiza usando maximización sobre acciones.
  - Selección de acción $\epsilon-$ greedy.

Sesgo de Maximización

- Algoritmos de TD usan maximización sobre acciones en un estado:
  - Q-Learning actualiza usando maximización sobre acciones.
  - Selección de acción $\epsilon-$ greedy.
- Se usa Máximo sobre estimativos como Estimativo del máximo $\Rightarrow$ Sesgo de maximización.

Sesgo de Maximización

- Algoritmos de TD usan maximización sobre acciones en un estado:
  - Q-Learning actualiza usando maximización sobre acciones.
  - Selección de acción $\epsilon-$ greedy.
- Se usa Máximo sobre estimativos como Estimativo del máximo $\Rightarrow$ Sesgo de maximización.

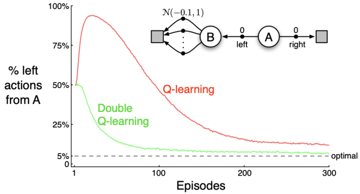


Figure 6.5: Comparison of Q-learning and Double Q-learning on a simple episodic MDP (shown inset). Q-learning initially learns to take the left action much more often than the right action, and always takes it significantly more often than the 5% minimum probability enforced by ε-greedy action selection with $\varepsilon = 0.1$. In contrast, Double Q-learning is essentially unaffected by maximization bias. These data are averaged over 10,000 runs. The initial action-value estimates were zero. Any ties in ε-greedy action selection were broken randomly.

Aprendizaje Doble

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$
- Idea:
  - Mantener 2 estimativos independientes de $q(a)$, $Q_1(a)$, $Q_2(a)$ $\forall a \in \mathcal{A}$.

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$
- Idea:
  - Mantener 2 estimativos independientes de $q(a)$, $Q_1(a)$, $Q_2(a)$ $\forall a \in \mathcal{A}$.
  - Usar Q_1 para maximizar, Q_2 para estimar valor de acción resultante:

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$
- Idea:
  - Mantener 2 estimativos independientes de $q(a)$, $Q_1(a)$, $Q_2(a)$ $\forall a \in \mathcal{A}$.
  - Usar Q_1 para maximizar, Q_2 para estimar valor de acción resultante:

$$
\begin{aligned}
A^* &= \arg\max Q_1(a) \\
Q_2(A^*) &= Q_2(\arg\max Q_1(a))
\end{aligned}
$$

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$
- Idea:
  - Mantener 2 estimativos independientes de $q(a)$, $Q_1(a)$, $Q_2(a)$ $\forall a \in \mathcal{A}$.
  - Usar Q_1 para maximizar, Q_2 para estimar valor de acción resultante:

$$
\begin{aligned}
A^* &= \arg\max Q_1(a) \\
Q_2(A^*) &= Q_2(\arg\max Q_1(a))
\end{aligned}
$$

  - Tenemos $\mathbb{E}\left[Q_2(A^*)\right] = q(A^*)$

Aprendizaje Doble

- Considere un Multiarmed-Bandit con valores $q(a)$
- Idea:
  - Mantener 2 estimativos independientes de $q(a)$, $Q_1(a)$, $Q_2(a)$ $\forall a \in \mathcal{A}$.
  - Usar Q_1 para maximizar, Q_2 para estimar valor de acción resultante:

$$A^* = \arg\max Q_1(a)$$
$$Q_2(A^*) = Q_2(\arg\max Q_1(a))$$

  - Tenemos $\mathbb{E}\left[Q_2(A^*)\right] = q(A^*)$
- Intercambiar Q_1 y Q_2.

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Inicialice $Q_1(s, a), Q_2(s, a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$

Incialice $Q_1(s,a), Q_2(s,a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat $\triangleright$ para cada episodio

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Inicialice $Q_1(s, a), Q_2(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat $\triangleright$ para cada episodio

Inicialice S

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q_1(s, a), Q_2(s, a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$

Inicialice $Q_1(s,a), Q_2(s,a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ $(\epsilon - \text{greedy})$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q_1(s, a), Q_2(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon -$ greedy)

Tome acción A, observe R, S'.

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$

Incialice $Q_1(s,a), Q_2(s,a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon-$greedy)

Tome acción A, observe R, S'.

Lanzar moneda

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Inicialice $Q_1(s, a), Q_2(s, a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ (ϵ − greedy)

Tome acción A, observe R, S'.

Lanzar moneda

if Cara **then**

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q_1(s,a), Q_2(s,a)$ $\forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ $(\epsilon - \text{greedy})$

Tome acción A, observe R, S'.

Lanzar moneda

if Cara **then**

$$Q_1(S,A) \leftarrow Q_1(S,A) + \alpha \left[R + \gamma Q_2(S', \arg\max_a Q_1(S', a)) - Q_1(S,A)\right]$$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Inicialice $Q_1(s, a), Q_2(s, a)$ $\forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

  Inicialice S

  repeat ▷ para cada paso del episodio

    Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon -$ greedy)

    Tome acción A, observe R, S'.

    Lanzar moneda

    if Cara **then**

$$Q_1(S, A) \leftarrow Q_1(S, A) + \alpha \left[R + \gamma Q_2(S', \arg\max_a Q_1(S', a)) - Q_1(S, A)\right]$$

    else

$$Q_2(S, A) \leftarrow Q_2(S, A) + \alpha \left[R + \gamma Q_1(S', \arg\max_a Q_2(S', a)) - Q_2(S, A)\right]$$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0,1]$, $\epsilon > 0$

Inciali ce $Q_1(s,a), Q_2(s,a)$ $\forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon -$ greedy)

Tome acción A, observe R, S'.

Lanzar moneda

if Cara **then**

$$Q_1(S,A) \leftarrow Q_1(S,A) + \alpha \left[R + \gamma Q_2(S', \arg\max_a Q_1(S',a)) - Q_1(S,A)\right]$$

else

$$Q_2(S,A) \leftarrow Q_2(S,A) + \alpha \left[R + \gamma Q_1(S', \arg\max_a Q_2(S',a)) - Q_2(S,A)\right]$$

end if

$S \leftarrow S'$

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Incialice $Q_1(s, a), Q_2(s, a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

  Inicialice S

  repeat ▷ para cada paso del episodio

    Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon -$ greedy)

    Tome acción A, observe R, S'.

    Lanzar moneda

    if Cara **then**

$$Q_1(S, A) \leftarrow Q_1(S, A) + \alpha \left[R + \gamma Q_2(S', \arg\max_a Q_1(S', a)) - Q_1(S, A)\right]$$

    else

$$Q_2(S, A) \leftarrow Q_2(S, A) + \alpha \left[R + \gamma Q_1(S', \arg\max_a Q_2(S', a)) - Q_2(S, A)\right]$$

    end if

    $S \leftarrow S'$

  until S es terminal

Double Q-Learning

Require: Tamaño de paso $\alpha \in (0, 1]$, $\epsilon > 0$

Inicialice $Q_1(s, a), Q_2(s, a)\ \forall s \in \mathcal{S}^+$, con $Q_{1,2}(s_{\text{terminal}}, .) = 0$

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Escoja A de $\mathcal{A}(S)$, de acuerdo a $Q_1 + Q_2$ ($\epsilon -$ greedy)

Tome acción A, observe R, S'.

Lanzar moneda

if Cara **then**

$$Q_1(S, A) \leftarrow Q_1(S, A) + \alpha \left[R + \gamma Q_2(S', \arg\max_a Q_1(S', a)) - Q_1(S, A)\right]$$

else

$$Q_2(S, A) \leftarrow Q_2(S, A) + \alpha \left[R + \gamma Q_1(S', \arg\max_a Q_2(S', a)) - Q_2(S, A)\right]$$

end if

$S \leftarrow S'$

until S es terminal

until ∞