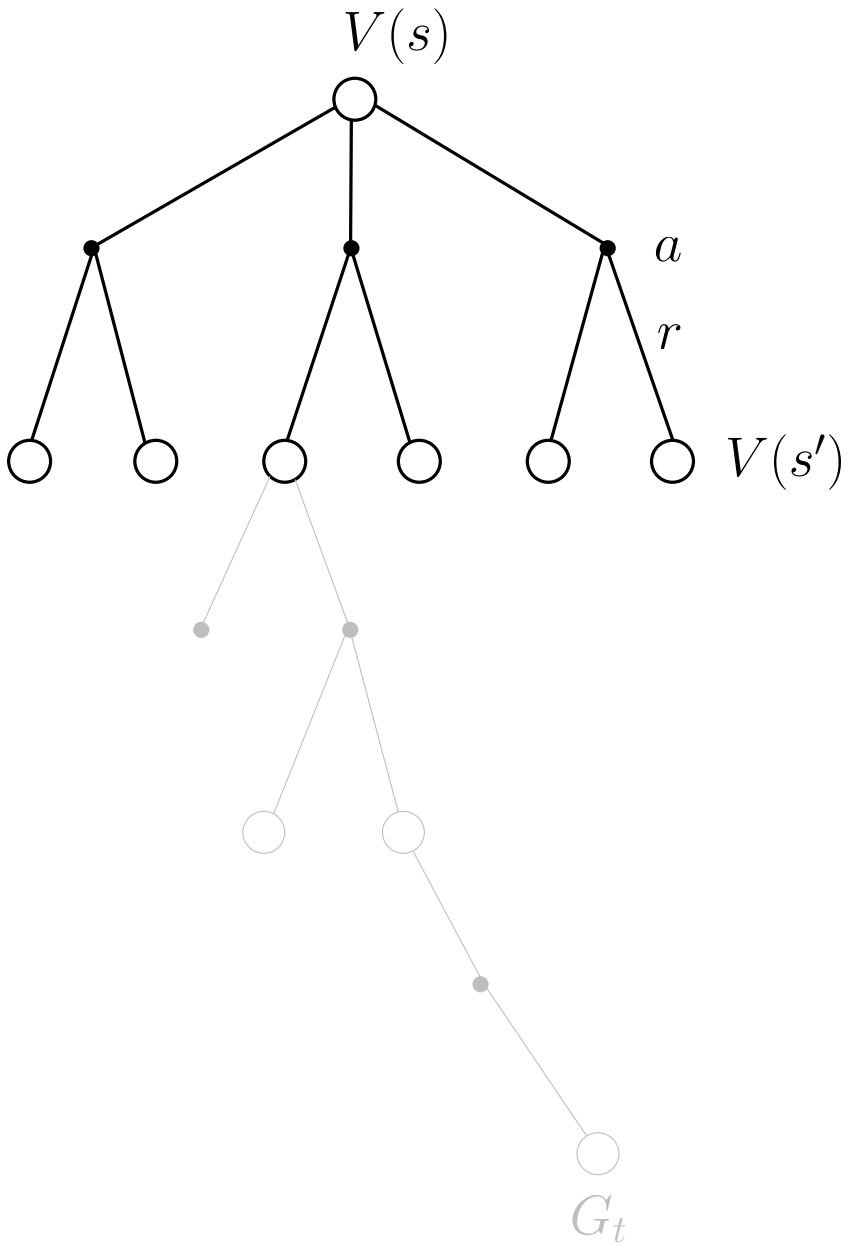
Aprendizaje por Diferencias Temporales con n pasos

Fernando Lozano

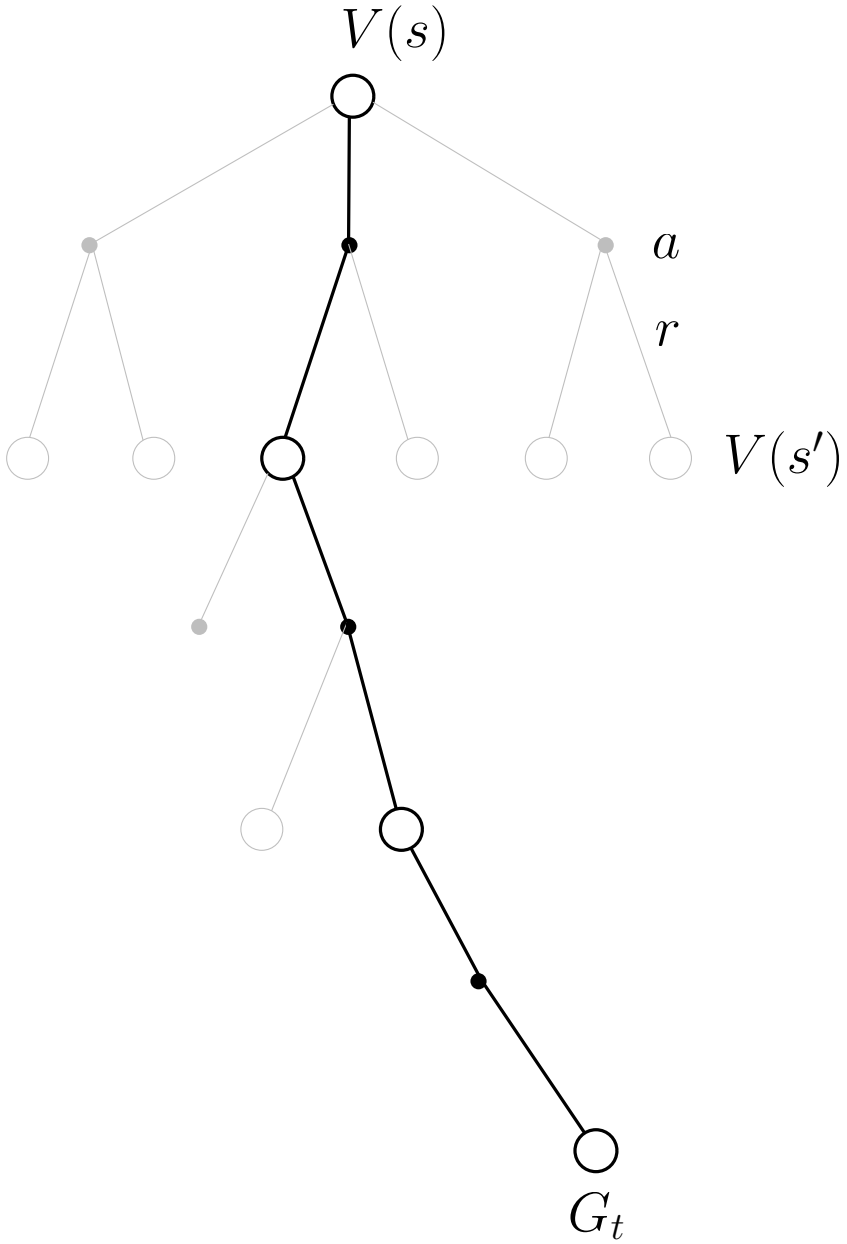
Universidad de los Andes

14 de marzo de 2023

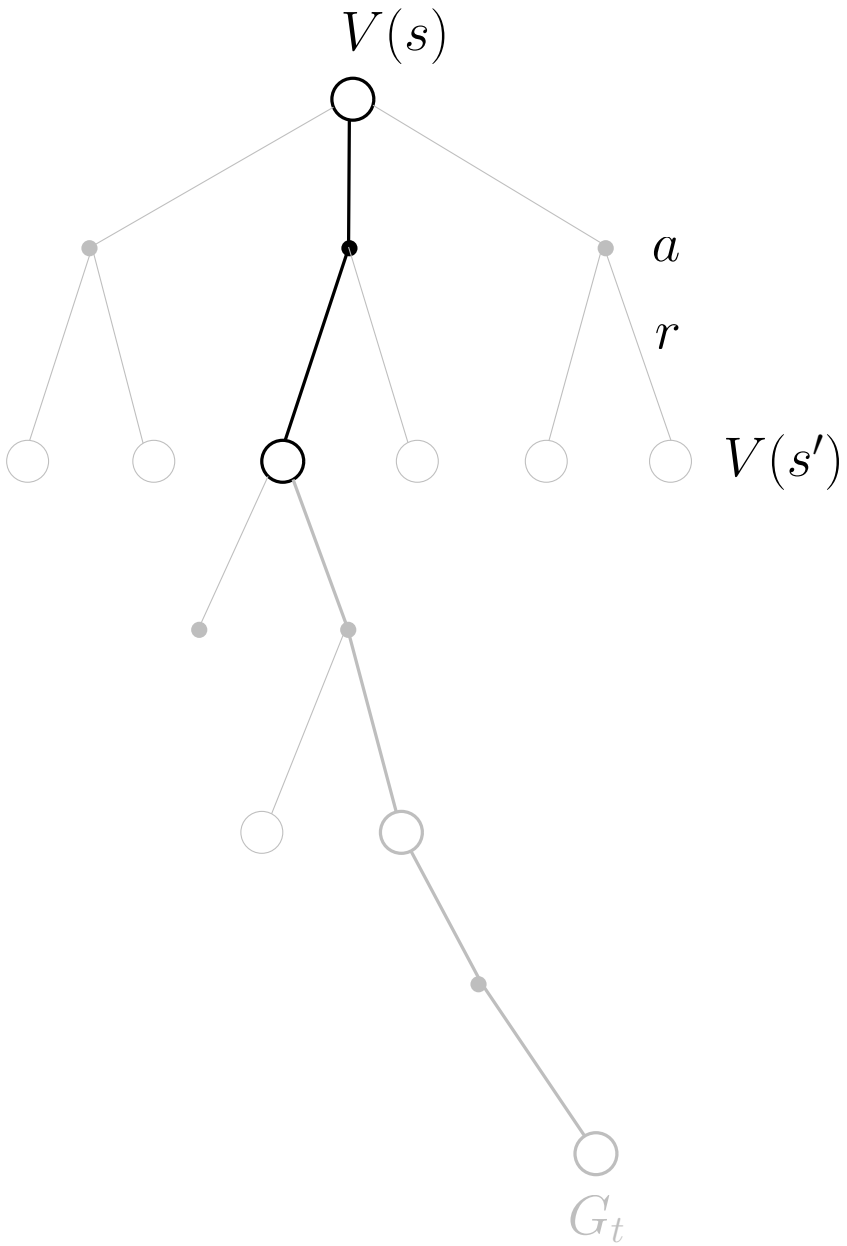
- Programación Dinámica.



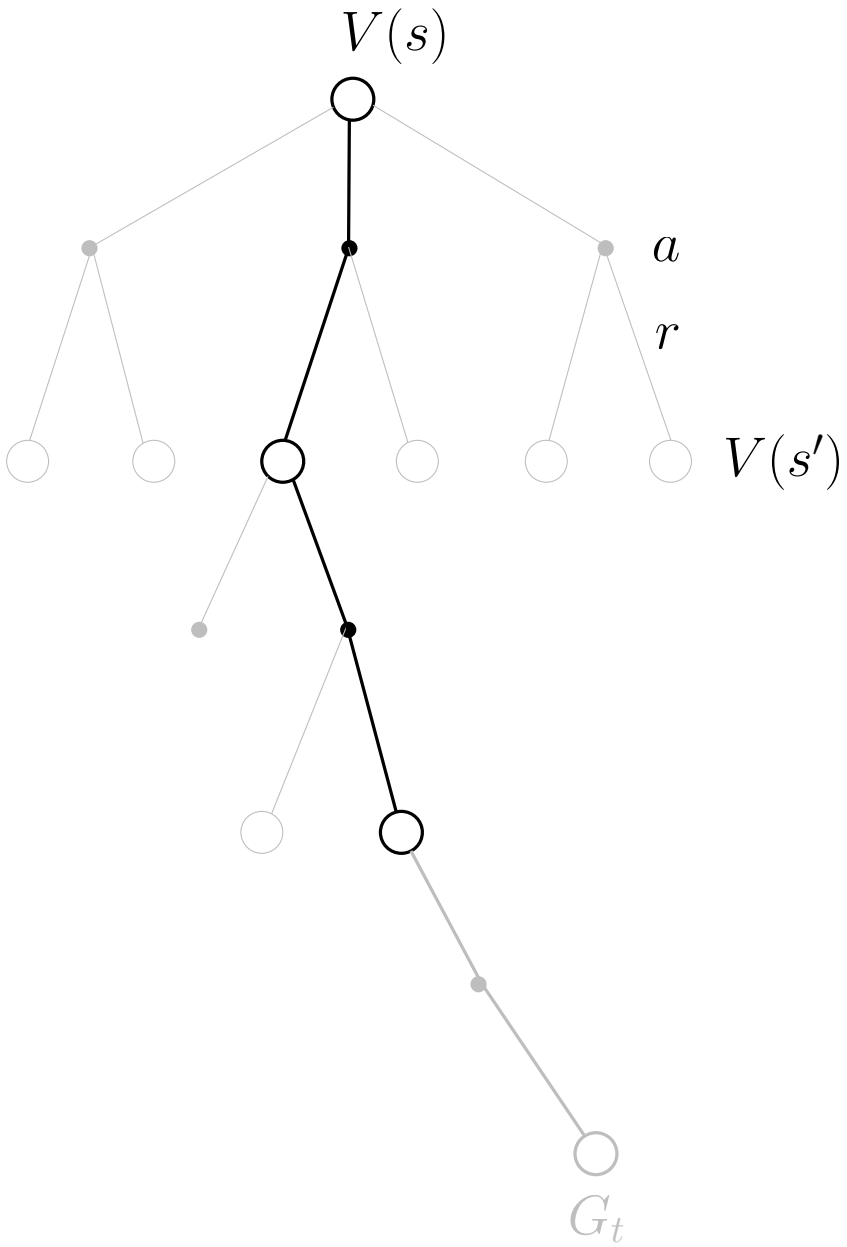
- Monte Carlo.



- Diferencias temporales 1 paso

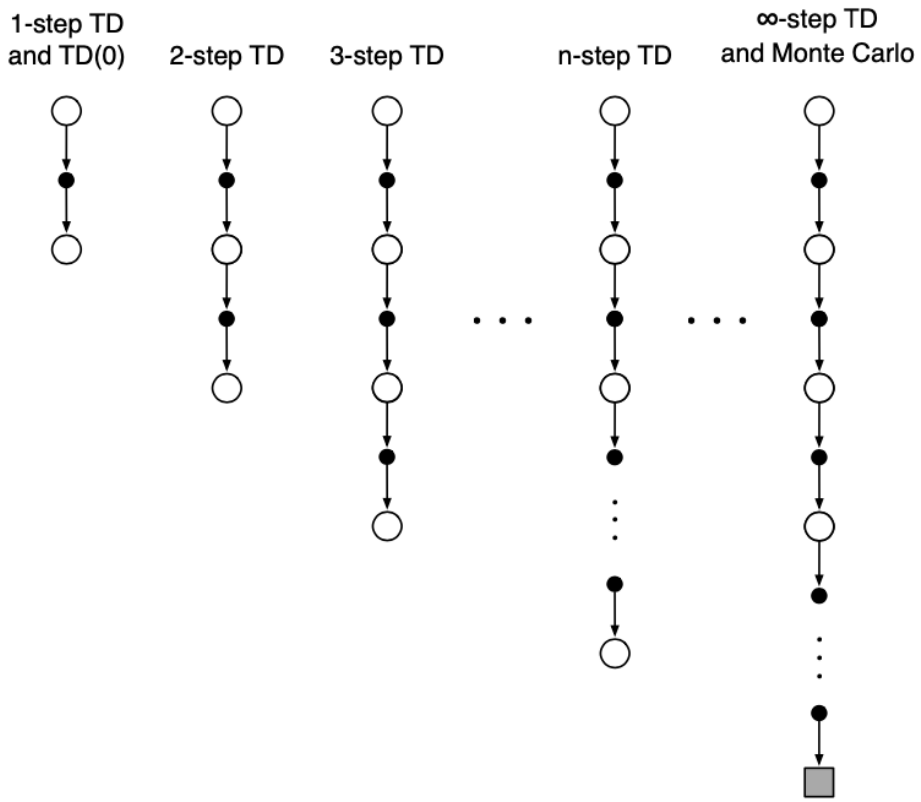


- Diferencias temporales n pasos.



Predicción TD de n pasos

Predicción TD de n pasos



- Target en Montecarlo:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

- Target en Montecarlo:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

- Target en TD de un paso:

$$G_{t:t+1} \doteq R_{t+1} + \gamma V_t(S_{t+1})$$

- Target en Montecarlo:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

- Target en TD de un paso:

$$G_{t:t+1} \doteq R_{t+1} + \gamma V_t(S_{t+1})$$

- Target en TD de dos pasos:

$$G_{t:t+2} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 V_{t+1}(S_{t+2})$$

- Target en Montecarlo:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

- Target en TD de un paso:

$$G_{t:t+1} \doteq R_{t+1} + \gamma V_t(S_{t+1})$$

- Target en TD de dos pasos:

$$G_{t:t+2} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 V_{t+1}(S_{t+2})$$

- En general, retorno de n pasos:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})$$

- Actualización después de observar R_{t+n}:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha \left[G_{t:t+n} - V_{t+n-1}(S_t) \right], \ 0 < T$$

- Actualización después de observar R_{t+n}:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha \left[G_{t:t+n} - V_{t+n-1}(S_t) \right], \ 0 < T$$

- En los primeros $n-1$ pasos del episodio no hay actualización.

- Actualización después de observar R_{t+n}:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha \left[G_{t:t+n} - V_{t+n-1}(S_t) \right], \ 0 < T$$

- En los primeros $n-1$ pasos del episodio no hay actualización.
- Al final del episodio se actualizan valores de los últimos $n-1$ estados.

TD de n pasos para estimar v_π

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

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
Incialice $V(s)$ $\forall s \in \mathcal{S}$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
Incialice $V(s)$ $\forall s \in \mathcal{S}$
repeat ▷ para cada episodio

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha \left[G - V(S_\tau)\right]$

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha \left[G - V(S_\tau)\right]$
      end if

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha \left[G - V(S_\tau)\right]$
      end if
      if $\tau = T - 1$ **then** break
      end if

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha \left[G - V(S_\tau)\right]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for

TD de n pasos para estimar v_π

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha \left[G - V(S_\tau)\right]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \end{array}$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \end{array} \qquad \tau = -2$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \end{array}$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \end{array} \qquad \tau = -1$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ S_3 & R_3 \end{array}$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ S_3 & R_3 \end{array} \qquad \tau = 0$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ S_3 & R_3 \end{array} \qquad \tau = 0$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ S_3 & R_3 \end{array} \qquad \tau = 0$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ S_3 & R_3 \end{array} \qquad \tau = 0$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 7$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 7$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 7$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 7$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 9, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 7$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 8$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 8$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 8$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 8$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 10, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 8$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10,$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 9$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 9$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 9$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**$T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 9$$

Ejemplo: $n = 3$, Episodio de duración 10

Require: Política π, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $V(s)$ $\forall s \in \mathcal{S}$
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal
    $T \leftarrow \infty$ ▷ Duración del episodio
    for $t = 0, 1, 2, \ldots, 11, \ldots$ **do**
      if $t < T$ **then**
        Tome acción de acuerdo a $\pi(. \mid S_t)$
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then** $T \leftarrow t + 1$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n V(S_{\tau+n})$
        end if
        $V(S_\tau) = V(S_\tau) + \alpha\,[G - V(S_\tau)]$
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

$$\begin{array}{c|c} S_0 & \\ S_1 & R_1 \\ S_2 & R_2 \\ \vdots & \vdots \\ S_{10} & R_{10} \end{array} \qquad T = 10, \quad \tau = 9$$

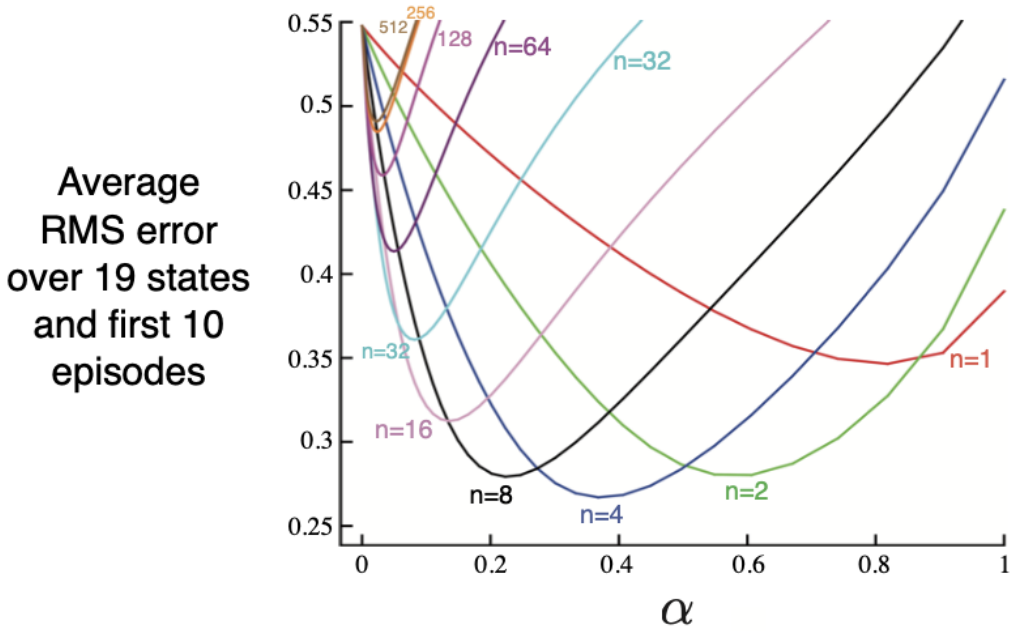
Ejemplo: Random Walk

Ejemplo: Random Walk

- 19 estados, $r(A) = -1{,}10$ Episodios, 100 repeticiones:

Ejemplo: Random Walk

- 19 estados, $r(A) = -1$,10 Episodios, 100 repeticiones:



SARSA de n pasos

SARSA de n pasos

- Transiciones entre pares (s, a).

SARSA de n pasos

- Transiciones entre pares (s, a).
- Política ϵ-greedy.

SARSA de n pasos

- Transiciones entre pares (s, a).
- Política ϵ-greedy.
- Retorno de n pasos:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

SARSA de n pasos

- Transiciones entre pares (s, a).
- Política ϵ-greedy.
- Retorno de n pasos:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

con $G_{t:t+n} \doteq G_t$ si $t + n \geq T$

SARSA de n pasos

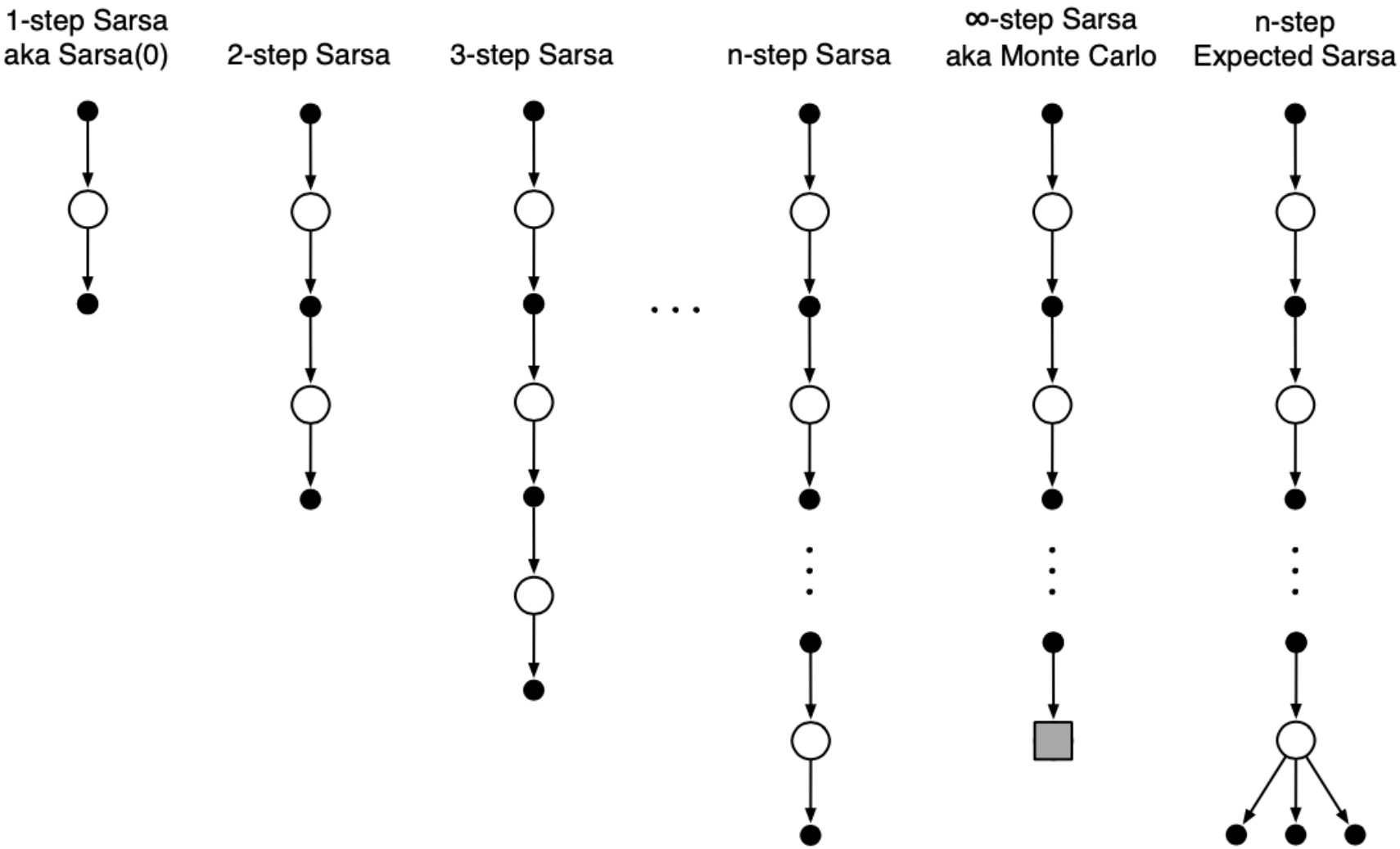
- Transiciones entre pares (s, a).
- Política ϵ-greedy.
- Retorno de n pasos:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

con $G_{t:t+n} \doteq G_t$ si $t + n \geq T$

- Actualización:

$$Q_{t+n}(S_t, A_t) \doteq Q_{t+n-1}(S_t, A_t) + \alpha \left[G_{t:t+n} - Q_{t+n-1}(S_t, A_t) \right], \; 0 \leq t < T$$


1-step Sarsa
aka Sarsa(0)
2-step Sarsa
3-step Sarsa
n-step Sarsa
∞-step Sarsa
aka Monte Carlo
n-step
Expected Sarsa

SARSA de n pasos, estimar $Q \approx q_\pi$

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

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\text{mín}(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \alpha \left[G - Q(S_\tau, A_\tau)\right]$

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \alpha \left[G - Q(S_\tau, A_\tau)\right]$
        Haga π ϵ-greedy con respecto a Q
      end if

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \alpha \left[G - Q(S_\tau, A_\tau)\right]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if

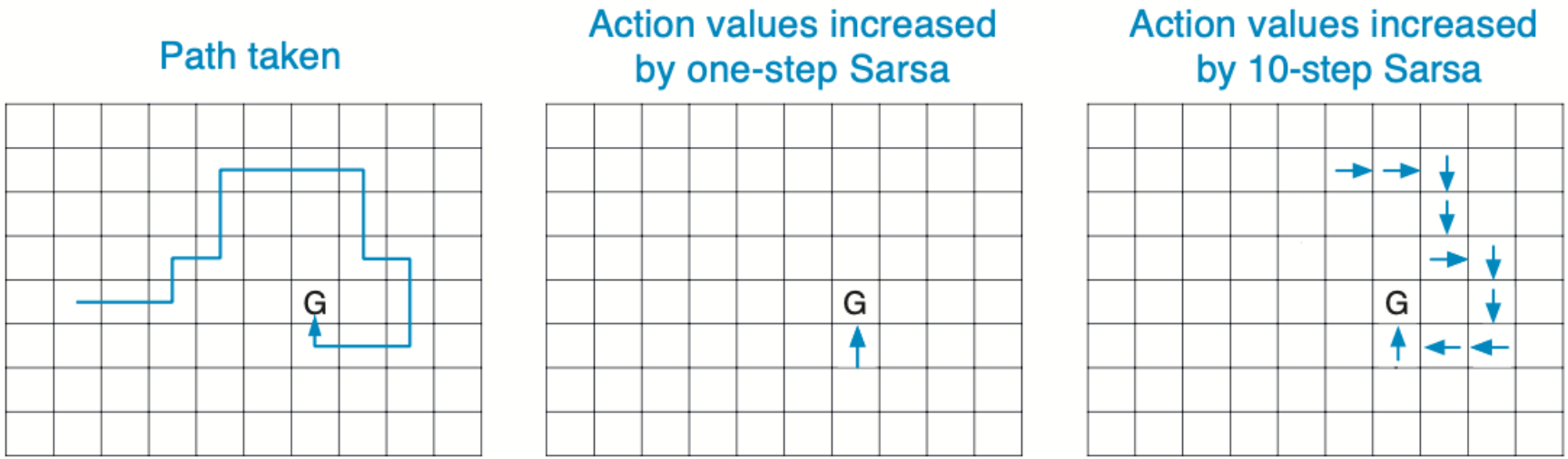
SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for

SARSA de n pasos, estimar $Q \approx q_\pi$

Require: Tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim \pi(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim \pi(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Ejemplo: Gridworld



Control off-policy de n pasos

Control off-policy de n pasos

- Separar en dos políticas:

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)
- Actualización de V de n pasos:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \rho_{t:t+n-1}\alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right], \; 0 \leq < T$$

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)
- Actualización de V de n pasos:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \rho_{t:t+n-1}\alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right], \quad 0 \leq < T$$

- Con razón de muestreo por importancia:

$$\rho_{t:h} \doteq \prod_{k=t}^{\min(h,T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$$

Control off-policy de n pasos

- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)
- Actualización de V de n pasos:

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \rho_{t:t+n-1}\alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right], \quad 0 \leq < T$$

- Con razón de muestreo por importancia:

$$\rho_{t:h} \doteq \prod_{k=t}^{\min(h,T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$$

- SARSA off-policy:

$$Q_{t+n}(S_t, A_t) \doteq Q_{t+n-1}(S_t, A_t) + \rho_{t:t+n-1}\alpha\left[G_{t:t+n} - Q_{t+n-1}(S_t, A_t)\right],$$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
Inicialice π (ϵ-greedy con respecto a Q)
repeat ▷ para cada episodio

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
Inicialice π (ϵ-greedy con respecto a Q)
repeat ▷ para cada episodio
Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \rho\alpha \left[G - Q(S_\tau, A_\tau)\right]$

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \rho\alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \rho\alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if

SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \rho\alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for

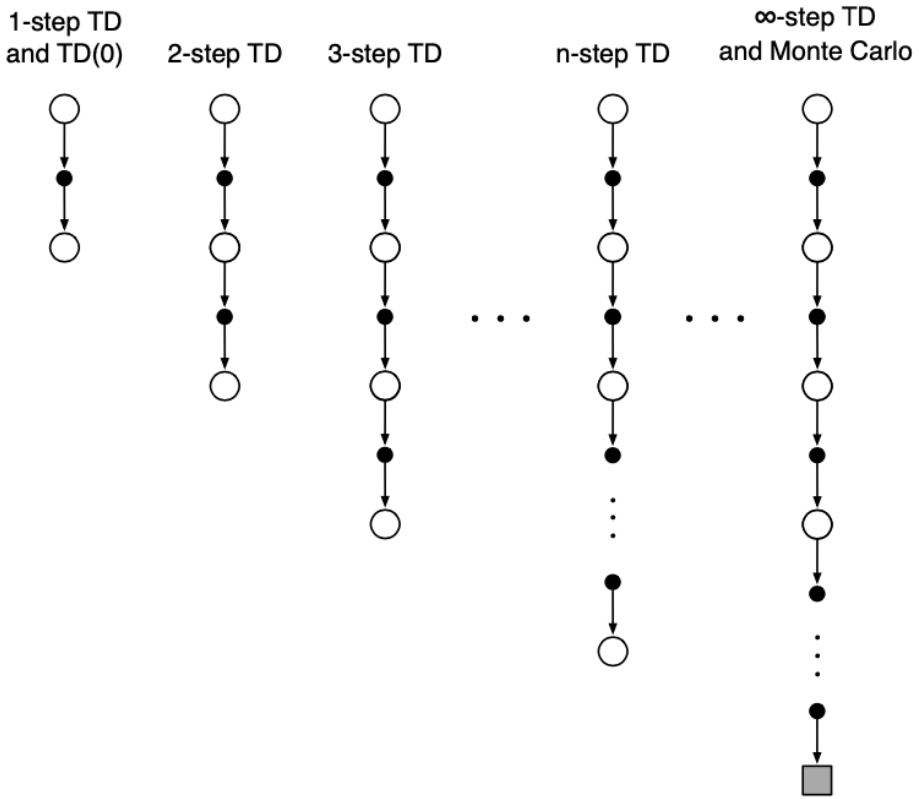
SARSA de n pasos off-policy, estimar $Q \approx q_\pi$

Require: Política soft b, tamaño de paso $\alpha \in (0, 1]$, n
  Incialice $Q(s, a)$ $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$
  Inicialice π (ϵ-greedy con respecto a Q)
  repeat ▷ para cada episodio
    Inicialice S_0 no terminal, $A_0 \sim b(. \mid S_0)$
    $T \leftarrow \infty$
    for $t = 0, 1, 2, \ldots$ **do**
      if $t < T$ **then**
        Tome acción A_t
        Observe y almacene R_{t+1}, S_{t+1}
        if S_{t+1} es terminal **then**
          $T \leftarrow t + 1$
        else
          Seleccione y almacene $A_t \sim b(. \mid S_{t+1})$
        end if
      end if
      $\tau \leftarrow t - n + 1$ ▷ tiempo del estado actualizado
      if $\tau \geq 0$ **then**
        $\rho \leftarrow \prod_{k=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$
        $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**
          $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) = Q(S_\tau, A_\tau) + \rho\alpha\,[G - Q(S_\tau, A_\tau)]$
        Haga π ϵ-greedy con respecto a Q
      end if
      if $\tau = T - 1$ **then** break
      end if
    end for
  until ∞

Trazas de elegibilidad

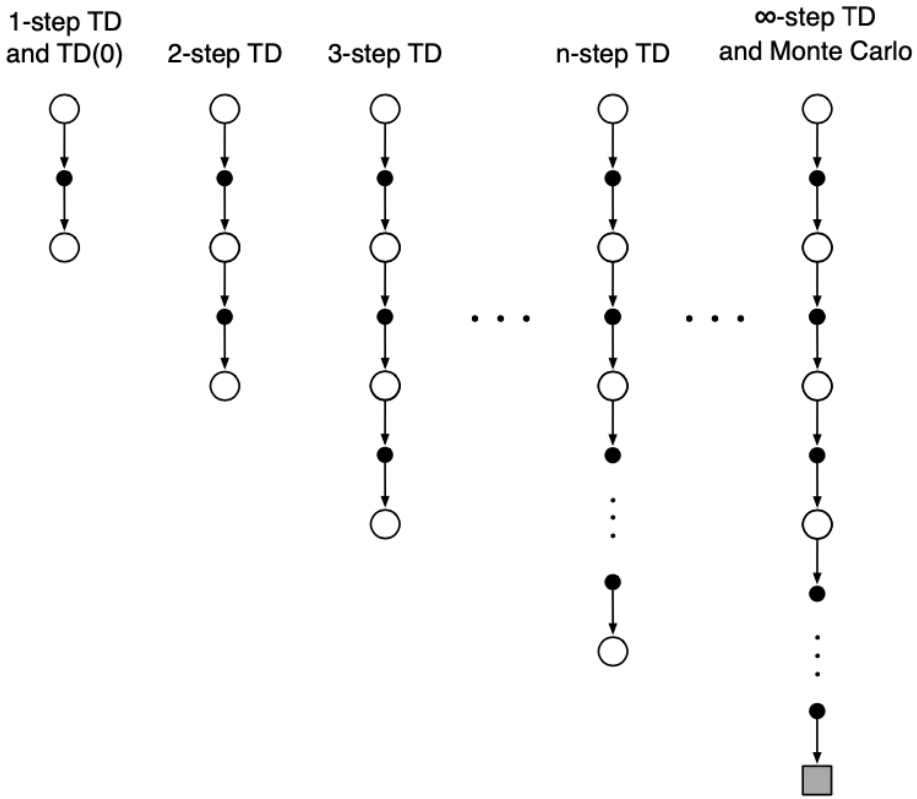
Trazas de elegibilidad

- TD con backup de n pasos:



Trazas de elegibilidad

- TD con backup de n pasos:



- Backup complejo: combinación convexa de retornos $G_{t:t+n}$.

$TD(\lambda)$

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$$G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n}$$

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

  - $\lambda = 1 \Rightarrow$

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} = (1-\lambda)\sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

  - $\lambda = 1 \Rightarrow$Montecarlo.

$TD(\lambda)$

- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

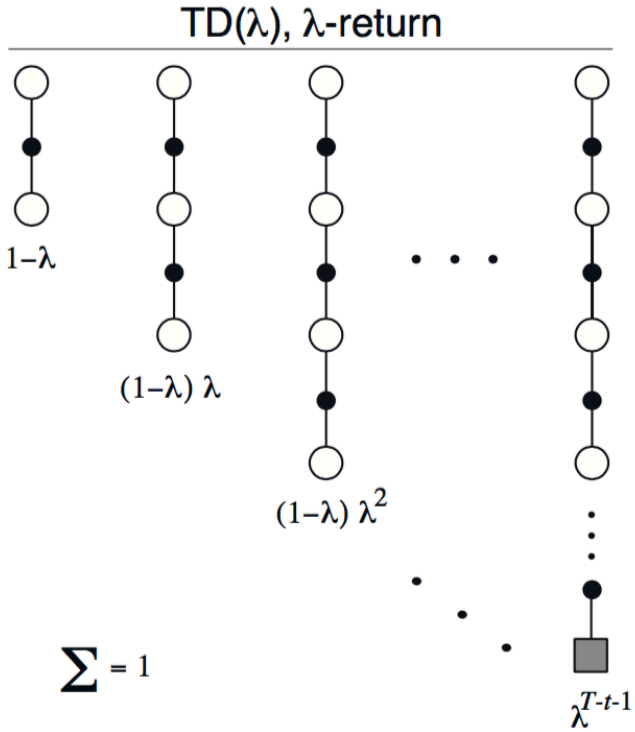
  - $\lambda = 1 \Rightarrow$Montecarlo.
  - $\lambda = 0 \Rightarrow$

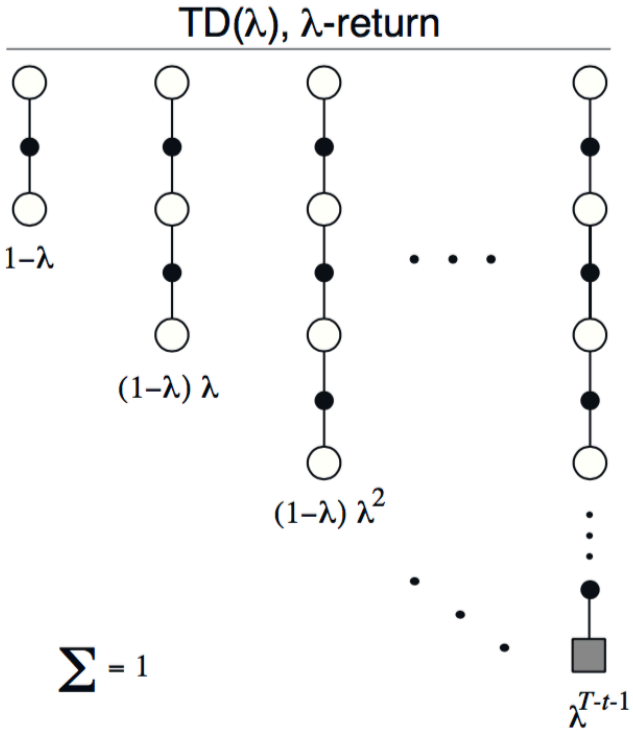
$TD(\lambda)$

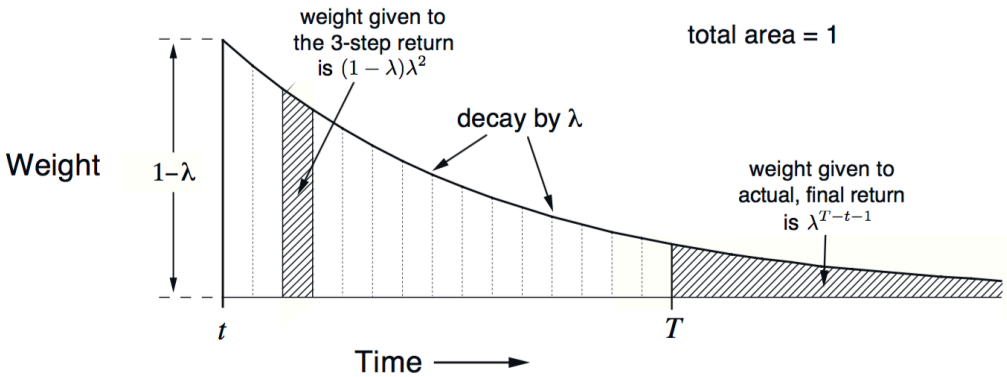
- Combinación de todos los retornos de n pasos con peso λ^{n-1}, con $0 \leq \lambda \leq 1$.

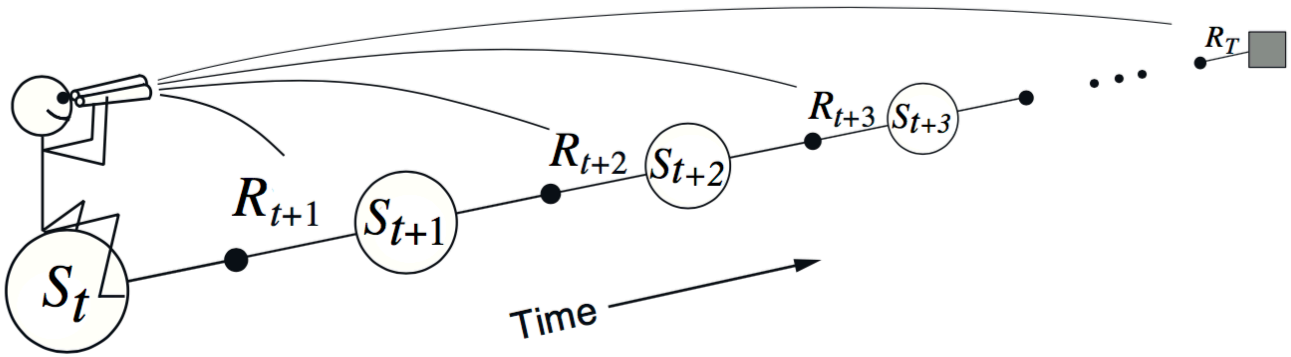
$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

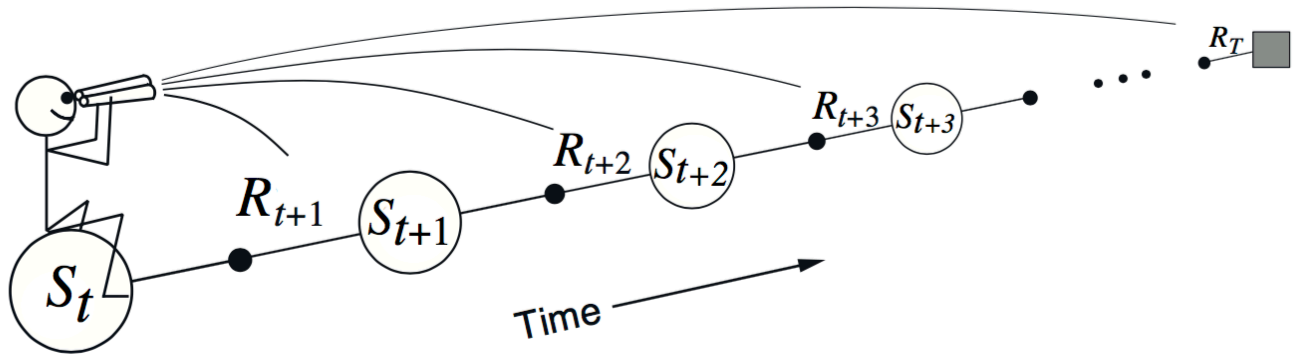
  - $\lambda = 1 \Rightarrow$Montecarlo.
  - $\lambda = 0 \Rightarrow$TD(0).

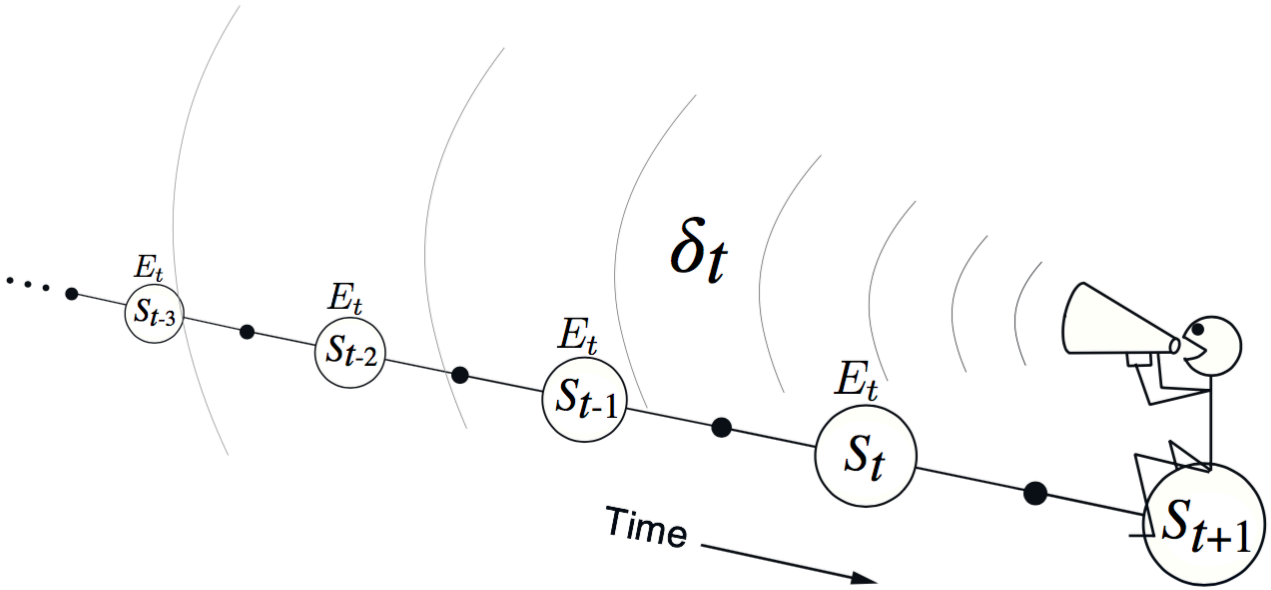

TD(λ), λ-return
1–λ
(1–λ) λ
(1–λ) λ2
∑ = 1
λT-t-1


TD(λ), λ-return
1–λ
(1–λ) λ
(1–λ) λ2
∑ = 1
λT-t-1


weight given to
the 3-step return
is $(1-\lambda)\lambda^2$
total area = 1
decay by λ
Weight
1–λ
weight given to
actual, final return
is λ^{T-t-1}
t
T
Time


R_T
R_{t+3}
S_{t+3}
R_{t+2}
S_{t+2}
R_{t+1}
S_{t+1}
S_t
Time


R_T
R_{t+3}
S_{t+3}
R_{t+2}
S_{t+2}
R_{t+1}
S_{t+1}
S_t
Time


δ_t
E_t
S_{t-3}
E_t
S_{t-2}
E_t
S_{t-1}
E_t
S_t
S_{t+1}
Time

Trazas de elegibilidad

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{si } s \neq S_t \ , \\ \gamma\lambda e_{t-1}(s) + 1 & \text{si } s = S_t \ . \end{cases}$$

Trazas de elegibilidad

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{si } s \neq S_t \ , \\ \gamma\lambda e_{t-1}(s) + 1 & \text{si } s = S_t \ . \end{cases}$$

- Error TD causa actualización proporcional de estados recientemente visitados:

Trazas de elegibilidad

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{si } s \neq S_t \ , \\ \gamma\lambda e_{t-1}(s) + 1 & \text{si } s = S_t \ . \end{cases}$$

- Error TD causa actualización proporcional de estados recientemente visitados:

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

Trazas de elegibilidad

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{si } s \neq S_t \ , \\ \gamma\lambda e_{t-1}(s) + 1 & \text{si } s = S_t \ . \end{cases}$$

- Error TD causa actualización proporcional de estados recientemente visitados:

$$\delta_t = R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t) \rightarrow \Delta V_t(s) = \alpha\delta_t e_t(s)$$

Evaluación de política con $TD(\lambda)$

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

Evaluación de política con $TD(\lambda)$

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

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

Evaluación de política con $TD(\lambda)$

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

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

repeat ▷ para cada episodio

Evaluación de política con $TD(\lambda)$

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

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

repeat ▷ para cada episodio

Inicialice S

Evaluación de política con $TD(\lambda)$

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

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

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

Evaluación de política con $TD(\lambda)$

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

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

repeat ▷ para cada episodio

Inicialice S

repeat ▷ para cada paso del episodio

$A \leftarrow$ acción dada por π en S

Evaluación de política con $TD(\lambda)$

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

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

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 $TD(\lambda)$

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

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

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'

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

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

$V(s) \leftarrow V(s) + \alpha\delta e(s)$

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

$V(s) \leftarrow V(s) + \alpha\delta e(s)$

$e(s) \leftarrow \gamma\lambda e(s)$

end for

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

$V(s) \leftarrow V(s) + \alpha\delta e(s)$

$e(s) \leftarrow \gamma\lambda e(s)$

end for

$S \leftarrow S'$

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

$V(s) \leftarrow V(s) + \alpha\delta e(s)$

$e(s) \leftarrow \gamma\lambda e(s)$

end for

$S \leftarrow S'$

until S es terminal

Evaluación de política con $TD(\lambda)$

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

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

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'

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

$e(S) \leftarrow e(S) + 1$

for $s \in \mathcal{S}$ **do**

$V(s) \leftarrow V(s) + \alpha\delta e(s)$

$e(s) \leftarrow \gamma\lambda e(s)$

end for

$S \leftarrow S'$

until S es terminal

until ∞

SARSA(λ)

SARSA(λ)

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

SARSA(λ)

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

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

SARSA(λ)

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

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

repeat ▷ para cada episodio

SARSA(λ)

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

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

repeat ▷ para cada episodio

Inicialice S

SARSA(λ)

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

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

repeat ▷ para cada episodio

Inicialice S

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

SARSA(λ)

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

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

repeat ▷ para cada episodio

Inicialice S

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

repeat ▷ para cada paso del episodio

SARSA(λ)

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

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

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(λ)

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

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

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(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

SARSA(λ)

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

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

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)

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

$e(S, A) \leftarrow e(S, A) + 1$

SARSA(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

SARSA(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

$Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$

SARSA(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

$Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$

$e(s, a) \leftarrow \gamma\lambda e(s, a)$

end for

SARSA(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

$Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$

$e(s, a) \leftarrow \gamma\lambda e(s, a)$

end for

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

SARSA(λ)

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

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

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 $(\epsilon - \text{greedy})$

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

$Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$

$e(s, a) \leftarrow \gamma\lambda e(s, a)$

end for

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

until S es terminal

SARSA(λ)

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

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

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)

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

$e(S, A) \leftarrow e(S, A) + 1$

for all s, a **do**

$Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$

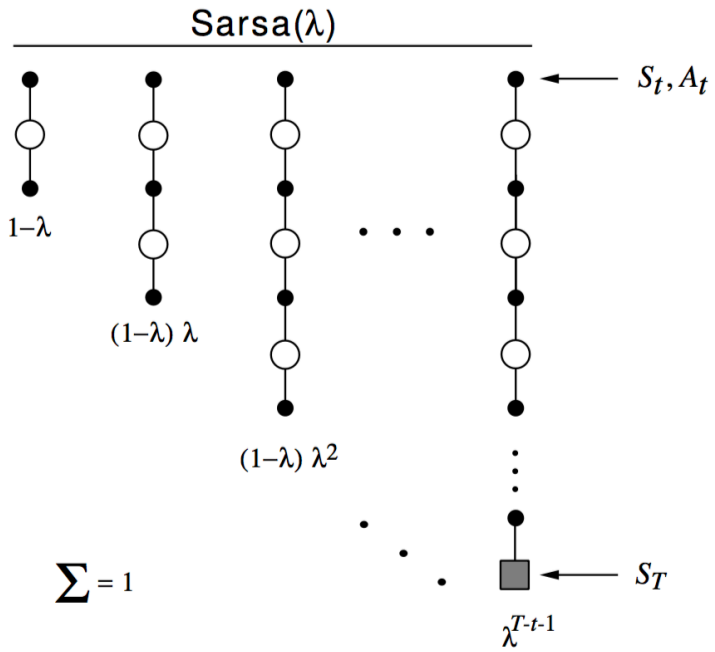
$e(s, a) \leftarrow \gamma \lambda e(s, a)$

end for

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

until S es terminal

until ∞


Sarsa(λ)
S_t, A_t
1–λ
(1–λ) λ
(1–λ) λ²
$\sum = 1$
S_T
λ^{T-t-1}