Métodos de Montecarlo

Fernando Lozano

Universidad de los Andes

16 de febrero de 2023

Métodos de Montecarlo

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.
  - Promediar muestras del retorno.

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.
  - Promediar muestras del retorno.
  - Tareas episódicas.

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.
  - Promediar muestras del retorno.
  - Tareas episódicas.
  - Estimación de $q_\pi(s, a)$: Búsqueda asociativa, Contextual bandits

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.
  - Promediar muestras del retorno.
  - Tareas episódicas.
  - Estimación de $q_\pi(s, a)$: Búsqueda asociativa, Contextual bandits(bandits relacionados por la dinámica del MDP).

Métodos de Montecarlo

- En general, las probabilidades $p(s', r \mid s, a)$ (o $p(s' \mid s, a), r(s, a, s')$) no se conocen explícitamente.
- Montecarlo: aprender a partir de experiencia:
  - Ejemplos de secuencias $s_t, a_t, r_{t+1}, s_{t+1}$.
  - Experiencia interactuando con el ambiente (on line) o por simulación.
  - Promediar muestras del retorno.
  - Tareas episódicas.
  - Estimación de $q_\pi(s, a)$: Búsqueda asociativa, Contextual bandits(bandits relacionados por la dinámica del MDP).
  - Iteración de política generalizada.

Evaluación de política π

- Generar episodios de acuerdo a π.

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.
- Estimativo de $v_\pi(s)$ es promedio de retornos obtenidos después de visitas a s.

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.
- Estimativo de $v_\pi(s)$ es promedio de retornos obtenidos después de visitas a s.
- Dos versiones:

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.
- Estimativo de $v_\pi(s)$ es promedio de retornos obtenidos después de visitas a s.
- Dos versiones:
  1. MC de primera visita.

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.
- Estimativo de $v_\pi(s)$ es promedio de retornos obtenidos después de visitas a s.
- Dos versiones:
  1. MC de primera visita.
  2. MC de todas las visitas.

Evaluación de política π

- Generar episodios de acuerdo a π.
- Una ocurrencia de s en un episodio se llama una visita a s.
- Estimativo de $v_\pi(s)$ es promedio de retornos obtenidos después de visitas a s.
- Dos versiones:
  1. MC de primera visita.
  2. MC de todas las visitas.

MC de primera visita

Incialice $V(s) \in \mathbb{R}$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
$G \leftarrow 0$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
    Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
    $G \leftarrow 0$
    for $t = T-1, T-2, \ldots 0$ **do**
        $G \leftarrow \gamma G + R_{t+1}$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
    Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
    $G \leftarrow 0$
    for $t = T-1, T-2, \ldots 0$ **do**
        $G \leftarrow \gamma G + R_{t+1}$
        if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
            Añada G a $Ret(S_t)$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
      Añada G a $Ret(S_t)$
      $V(S_t) \leftarrow$ promedio $Ret(S_t)$

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
      Añada G a $Ret(S_t)$
      $V(S_t) \leftarrow$ promedio $Ret(S_t)$
    end if

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
      Añada G a $Ret(S_t)$
      $V(S_t) \leftarrow$ promedio $Ret(S_t)$
    end if
  end for

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
      Añada G a $Ret(S_t)$
      $V(S_t) \leftarrow$ promedio $Ret(S_t)$
    end if
  end for
until ∞

MC de primera visita

Incialice $V(s) \in \mathbb{R}$
Incialice una lista vacía $Ret(s)$ para cada s.
repeat
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$ **then**
      Añada G a $Ret(S_t)$
      $V(S_t) \leftarrow$ promedio $Ret(S_t)$
    end if
  end for
until ∞

- Retornos a partir de $S_t = s$ son muestras i.i.d. de $v_\pi(s)$.

- Retornos a partir de $S_t = s$ son muestras i.i.d. de $v_\pi(s)$.
- $V(s)$ converge a $v_\pi(s)$.

- Retornos a partir de $S_t = s$ son muestras i.i.d. de $v_\pi(s)$.
- $V(s)$ converge a $v_\pi(s)$.
- $V(s)$ es estimador no sesgado de $v_\pi(s)$ con varianza $1/\sqrt{n}$, donde n es el número de visitas a s.

- Retornos a partir de $S_t = s$ son muestras i.i.d. de $v_\pi(s)$.
- $V(s)$ converge a $v_\pi(s)$.
- $V(s)$ es estimador no sesgado de $v_\pi(s)$ con varianza $1/\sqrt{n}$, donde n es el número de visitas a s.
- MC de todas las visitas converge similarmente.

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).
  - Carta que muestra el dealer (baraja infinita).

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).
  - Carta que muestra el dealer (baraja infinita).
  - As usable: si o no.

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).
  - Carta que muestra el dealer (baraja infinita).
  - As usable: si o no.
- Dealer: política fija $\rightarrow$ pide carta sólo si total < 17.

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).
  - Carta que muestra el dealer (baraja infinita).
  - As usable: si o no.
- Dealer: política fija $\rightarrow$ pide carta sólo si total < 17.
- Recompensas $+1, -1, 0$ al ganar, perder o empatar.

Ejemplo: Blackjack

- Tarea episódica (juego=episodio).
- Estado:
  - Suma cartas (12,. . . ,21).
  - Carta que muestra el dealer (baraja infinita).
  - As usable: si o no.
- Dealer: política fija $\rightarrow$ pide carta sólo si total < 17.
- Recompensas $+1, -1, 0$ al ganar, perder o empatar.
- Política a evaluar: pedir carta si total < 20.

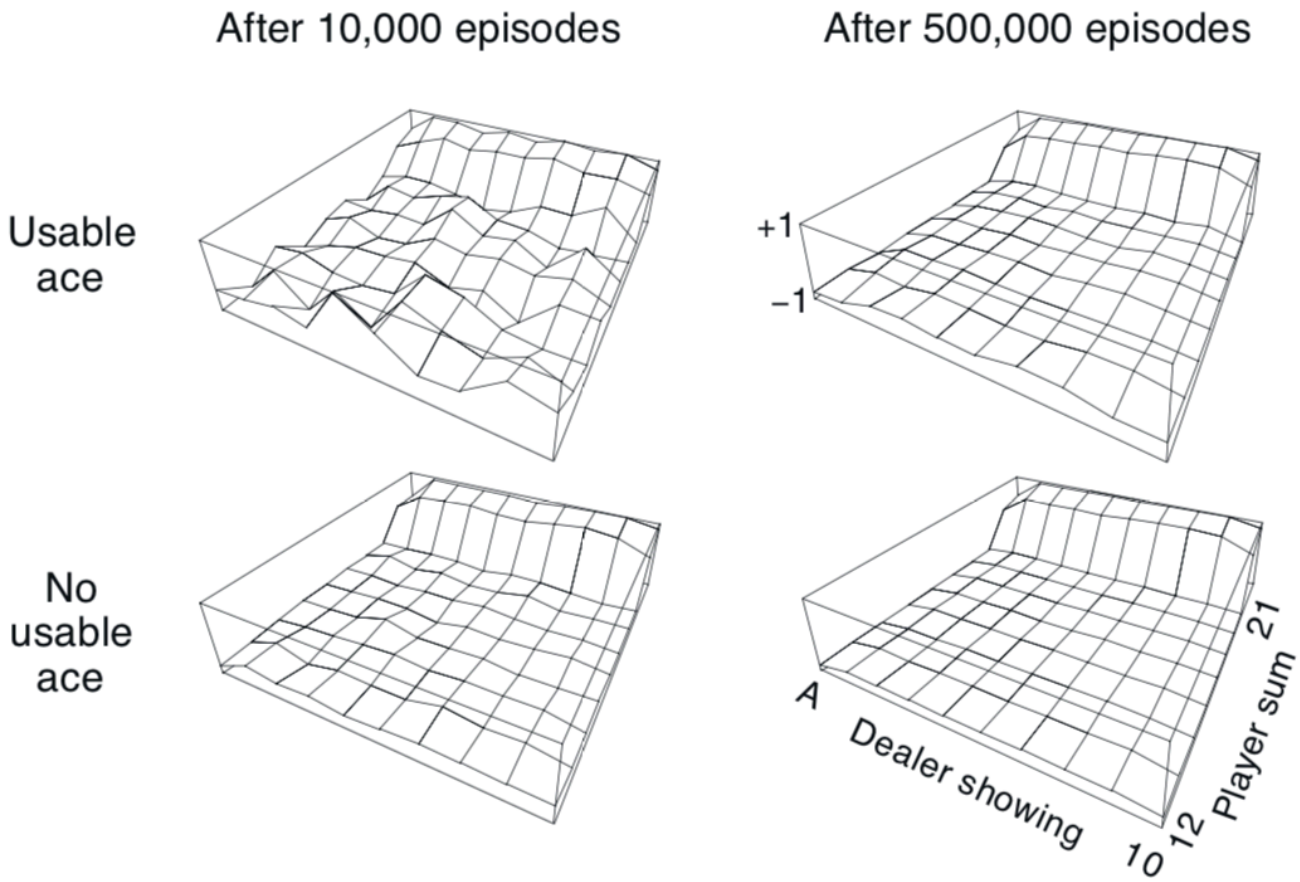


Figure 5.2: Approximate state-value functions for the blackjack policy that sticks only on 20 or 21, computed by Monte Carlo policy evaluation.

Estimación de valor de acciones

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.
- Es necesario estimar $q_\pi(s, a)$ para todas las acciones a posibles en s.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.
- Es necesario estimar $q_\pi(s, a)$ para todas las acciones a posibles en s.
- Mantener exploración.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.
- Es necesario estimar $q_\pi(s, a)$ para todas las acciones a posibles en s.
- Mantener exploración.
- Asumir episodios que comienzan en (s, a): supuesto de arranque explorativo.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.
- Es necesario estimar $q_\pi(s, a)$ para todas las acciones a posibles en s.
- Mantener exploración.
- Asumir episodios que comienzan en (s, a): supuesto de arranque explorativo.
- En la práctica, usar políticas estocásticas.

Estimación de valor de acciones

- Estimar q_*, en lugar de v_*.
- Estimar $q_\pi(s, a)$ promediando retornos después de visitar s, tomar acción a y seguir π.
- Problema: Pares (s, a) relevantes pueden no ser visitados.
- Es necesario estimar $q_\pi(s, a)$ para todas las acciones a posibles en s.
- Mantener exploración.
- Asumir episodios que comienzan en (s, a): supuesto de arranque explorativo.
- En la práctica, usar políticas estocásticas.

Control de política con Montecarlo

Control de política con Montecarlo

π_0

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \ldots \xrightarrow{\mathbf{I}} \pi^*$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \ldots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s, a)$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s,a)$$

- Note que:

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a))$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s,a)$$

- Note que:

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \end{aligned}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s,a)$$

- Note que:

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \end{aligned}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s,a)$$

- Note que:

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \\ &= v_{\pi_k}(s) \end{aligned}$$

Control de política con Montecarlo

$$\pi_0 \xrightarrow{\mathbf{E}} q_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} q_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi^* \xrightarrow{\mathbf{E}} q_{\pi^*}$$

- π_{k+1} es greedy con respecto a q_{π_k}:

$$\pi(s) = \arg\max_a q(s,a)$$

- Note que:

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \\ &= v_{\pi_k}(s) \end{aligned}$$

Se satisface teorema de mejoramiento de política.

MC de primera visita con ES

Inicialice $\pi(s)$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s, a) \in \mathbb{R}$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
    Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$
for $t = T-1, T-2, \ldots 0$ **do**

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    end if

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    end if
  end for

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    end if
  end for
until ∞

MC de primera visita con ES

Inicialice $\pi(s)$
Incialice $Q(s,a) \in \mathbb{R}$
Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat
  Escoja $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ aleatoriamente.
  Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**
      Añada G a $Ret(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    end if
  end for
until ∞

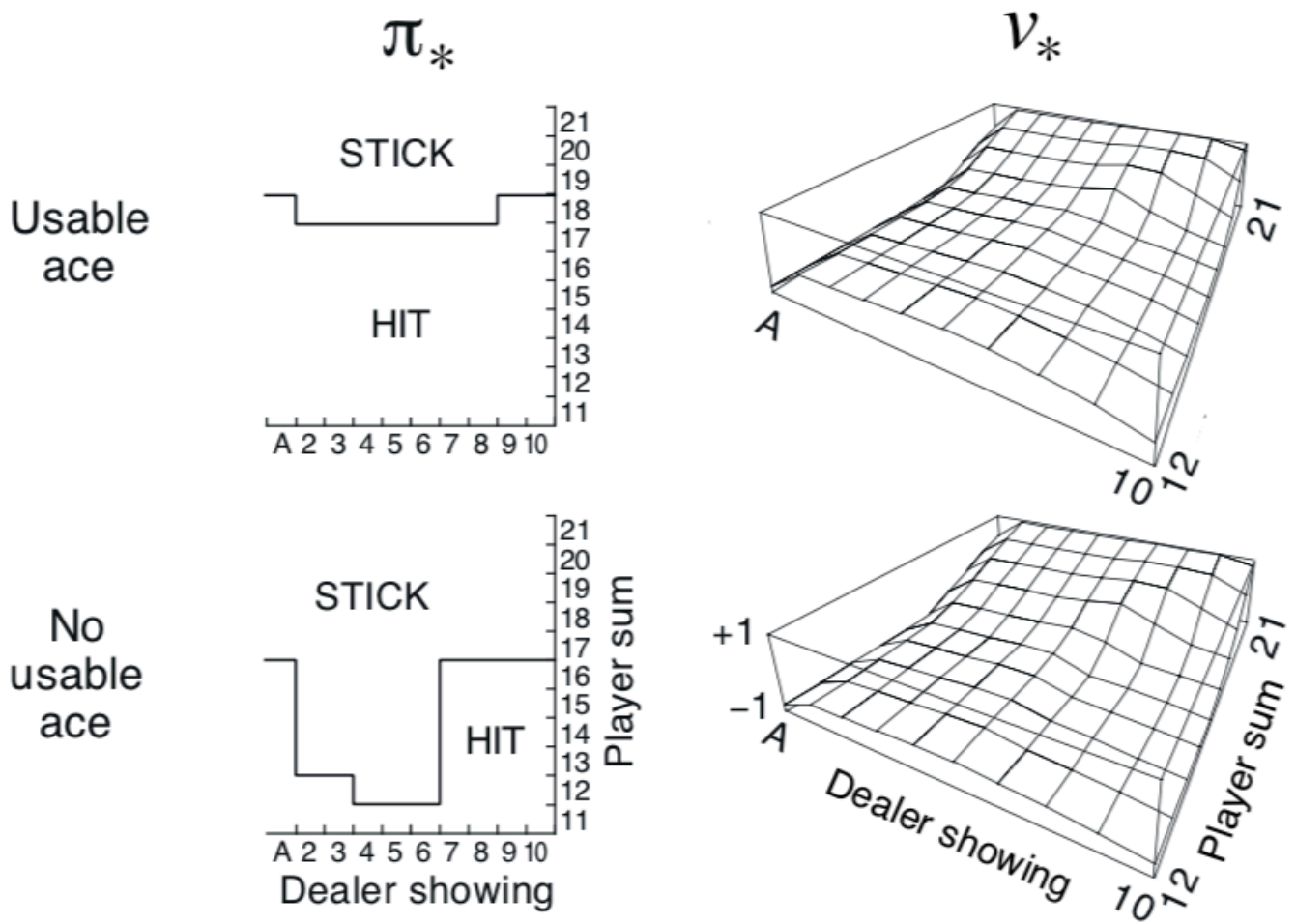


Figure 5.5: The optimal policy and state-value function for blackjack, found by Monte Carlo ES (Figure 5.4). The state-value function shown was computed from the action-value function found by Monte Carlo ES.

Control On-policy

Control On-policy

- Suposición de arranque explorativo no es práctica.

Control On-policy

- Suposición de arranque explorativo no es práctica.
- Control On-policy: evalúa y mejora la política usada para tomar decisiones.

Control On-policy

- Suposición de arranque explorativo no es práctica.
- Control On-policy: evalúa y mejora la política usada para tomar decisiones.
- Políticas soft:

$$\pi(a \mid s) > 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

Control On-policy

- Suposición de arranque explorativo no es práctica.
- Control On-policy: evalúa y mejora la política usada para tomar decisiones.
- Políticas soft:

$$\pi(a \mid s) > 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

- Política ϵ-greedy:

Control On-policy

- Suposición de arranque explorativo no es práctica.
- Control On-policy: evalúa y mejora la política usada para tomar decisiones.
- Políticas soft:

$$\pi(a \mid s) > 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

- Política ϵ-greedy:

$$\pi(a \mid s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(s)|} & a = \arg\max_a q(s,a), \\ \frac{\epsilon}{|\mathcal{A}(s)|} & \text{demás acciones.} \end{cases}$$

- Una política π' que es $\epsilon - \text{greedy}$ con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

- Una política π' que es $\epsilon - \text{greedy}$ con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$q_\pi(s, \pi'(s)) = \sum_a \pi'(a \mid s) q_\pi(s, a)$$

- Una política π' que es $\epsilon - \text{greedy}$ con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a \mid s) q_\pi(s, a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon) \max_a q_\pi(s, a)
\end{aligned}
$$

- Una política π' que es $\epsilon-$ greedy con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a \mid s) q_\pi(s, a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \max_a q_\pi(s, a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \sum_a \frac{\pi(a \mid s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s, a)
\end{aligned}
$$

$$
\left.\begin{matrix} \alpha_i \geq 0 \\ \sum_i \alpha_i = 1 \end{matrix}\right\} \Rightarrow \sum_i \alpha_i x_i \leq \max_i x_i
$$

- Una política π' que es $\epsilon - \text{greedy}$ con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a \mid s) q_\pi(s, a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon) \max_a q_\pi(s, a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon) \sum_a \frac{\pi(a \mid s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1 - \epsilon} q_\pi(s, a) \\
&\qquad \uparrow \\
&\qquad \left.\begin{matrix} \alpha_i \geq 0 \\ \sum_i \alpha_i = 1 \end{matrix}\right\} \Rightarrow \textstyle\sum_i \alpha_i x_i \leq \max_i x_i \\
&= \sum_a \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

- Una política π' que es $\epsilon-$ greedy con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a \mid s) q_\pi(s, a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \max_a q_\pi(s, a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \sum_a \frac{\pi(a \mid s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s, a) \\
&\qquad\qquad \uparrow \\
&\qquad\qquad \left.\begin{matrix} \alpha_i \geq 0 \\ \sum_i \alpha_i = 1 \end{matrix}\right\} \Rightarrow \sum_i \alpha_i x_i \leq \max_i x_i \\
&= \sum_a \pi(a \mid s) q_\pi(s, a) \\
&= v_\pi(s)
\end{aligned}
$$

- Una política π' que es $\epsilon-$ greedy con respecto a q_π, mejora cualquier otra política $\epsilon-$soft.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a \mid s) q_\pi(s, a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \max_a q_\pi(s, a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1-\epsilon) \sum_a \frac{\pi(a \mid s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s, a) \\
&\qquad\qquad \uparrow \\
&\qquad\qquad \left.\begin{matrix} \alpha_i \geq 0 \\ \sum_i \alpha_i = 1 \end{matrix}\right\} \Rightarrow \sum_i \alpha_i x_i \leq \max_i x_i \\
&= \sum_a \pi(a \mid s) q_\pi(s, a) \\
&= v_\pi(s)
\end{aligned}
$$

- Se puede demostrar que $q_\pi(s, \pi'(s)) = v_\pi(s) \Leftrightarrow \pi, \pi'$ son óptimas entre todas las políticas $\epsilon-$soft

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon -$ greedy

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon -$ greedy
Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy
Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).
repeat

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1}$,

$G \leftarrow 0$

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

$$\pi(a \mid S_t) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(S_t)|} & a = \arg\max_a Q(S_t, a) \\ \frac{\epsilon}{|\mathcal{A}(s)|} & a \neq \arg\max_a Q(S_t, a) \end{cases}$$

end if

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

$$\pi(a \mid S_t) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(S_t)|} & a = \arg\max_a Q(S_t, a) \\ \frac{\epsilon}{|\mathcal{A}(s)|} & a \neq \arg\max_a Q(S_t, a) \end{cases}$$

end if

end for

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

$$\pi(a \mid S_t) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(S_t)|} & a = \arg\max_a Q(S_t, a) \\ \frac{\epsilon}{|\mathcal{A}(s)|} & a \neq \arg\max_a Q(S_t, a) \end{cases}$$

end if

end for

until ∞

On policy control MC de primera visita

Inicialice $\pi(s)$ $\epsilon-$ greedy

Incialice $Q(s,a) \in \mathbb{R}$, Incialice una lista vacía $Ret(s,a)$ para cada (s,a).

repeat

Episodio π : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_{T-1},$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots 0$ **do**

$G \leftarrow \gamma G + R_{t+1}$

if S_t, A_t no aparece en $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ **then**

Añada G a $Ret(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ promedio $Ret(S_t, A_t)$

$$\pi(a \mid S_t) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(S_t)|} & a = \arg\max_a Q(S_t, a) \\ \frac{\epsilon}{|\mathcal{A}(s)|} & a \neq \arg\max_a Q(S_t, a) \end{cases}$$

end if

end for

until ∞

Control off-policy

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!)

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)
- Aproximación más general.

Control off-policy

- Aprendizaje de q(s,a) condicionado a comportamiento futuro óptimo.
- Mantener exploración (acciones no óptimas!): $\epsilon -$ greedy.
- Separar en dos políticas:
  - Política que se aprende π (política objetivo), puede ser determinística.
  - Política de comportamiento b (soft)
- Aproximación más general.
- Métodos aplicables a aprendizaje de otras fuentes (p.ej. comportamiento humano).

Estimación

Estimación

- Estimar v_π o q_π usando experiencia de acuerdo a $b \neq \pi$.

Estimación

- Estimar v_π o q_π usando experiencia de acuerdo a $b \neq \pi$.
- Considere un episodio generado usando π:

$$S_0, A_0, R_1; S_1, A_1, R_2 \ldots$$
$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

Estimación

- Estimar v_π o q_π usando experiencia de acuerdo a $b \neq \pi$.
- Considere un episodio generado usando π:

$$S_0, A_0, R_1; S_1, A_1, R_2 \ldots$$
$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

- Si $\forall s \in \mathcal{S}$, $\pi(a \mid s) > 0 \Rightarrow b(a \mid s) > 0$, es probable observar la subsequencia:

$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

usando b.

Estimación

- Estimar v_π o q_π usando experiencia de acuerdo a $b \neq \pi$.
- Considere un episodio generado usando π:

$$S_0, A_0, R_1; S_1, A_1, R_2 \ldots$$
$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

- Si $\forall s \in \mathcal{S}$, $\pi(a \mid s) > 0 \Rightarrow b(a \mid s) > 0$, es probable observar la subsequencia:

$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

usando b. (b tiene cubrimiento de π).

Estimación

- Estimar v_π o q_π usando experiencia de acuerdo a $b \neq \pi$.
- Considere un episodio generado usando π:

$$S_0, A_0, R_1; S_1, A_1, R_2 \ldots$$
$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

- Si $\forall s \in \mathcal{S}$, $\pi(a \mid s) > 0 \Rightarrow b(a \mid s) > 0$, es probable observar la subsequencia:

$$S_\tau, A_\tau, R_{\tau+1}; S_{\tau+1}, A_{\tau+1}, R_{\tau+2} \ldots S_{T-1}, A_{T-1}, R_T; S_T$$

  usando b. (b tiene cubrimiento de π).
- Aunque es la misma secuencia, en general sucede con probabilidades diferentes bajo π que bajo b.

Muestreo por importancia (Importance Sampling)

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es:

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t$$

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

  - Estimador no sesgado.

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

  - Estimador no sesgado.

$$\mathbb{E}\left[X_t \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}\right]$$

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

  - Estimador no sesgado.

$$\mathbb{E}\left[X_t \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}\right] = \sum_{i=1}^{6} \mathbf{P}_u\{X_i\} X_i \frac{\mathbf{P}_c\{X_i\}}{\mathbf{P}_u\{X_i\}}$$

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

  - Estimador no sesgado.

$$\mathbb{E}\left[X_t \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}\right] = \sum_{i=1}^{6} \mathbf{P}_u\{X_i\} X_i \frac{\mathbf{P}_c\{X_i\}}{\mathbf{P}_u\{X_i\}} = \sum_{i=1}^{6} \mathbf{P}_c\{X_i\} X_i$$

Muestreo por importancia (Importance Sampling)

- Lanzar dado justo, valor esperado del número mostrado es: 3.5.
- Suponga un segundo dado en el que la probabilidad de 4,5,6 es $\frac{2}{9}$ y la probabilidad de 1,2, 3 es $\frac{1}{9}$.(valor esperado 4)
- Cómo usamos el primer dado para estimar el valor esperado del número mostrado por el segundo?
- Muestreo por importancia simple:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \qquad \underbrace{\times \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}_{\text{importance sampling ratio}}$$

  - Estimador no sesgado.

$$\mathbb{E}\left[X_t \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}\right] = \sum_{i=1}^{6} \mathbf{P}_u\{X_i\} X_i \frac{\mathbf{P}_c\{X_i\}}{\mathbf{P}_u\{X_i\}} = \sum_{i=1}^{6} \mathbf{P}_c\{X_i\} X_i$$

  - Puede tener varianza alta, cuando $\frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}} \gg$

- Muestreo por importancia pesado:

$$\sum_{i=1}^{T} X_t$$

- Muestreo por importancia pesado:

$$\sum_{i=1}^{T} X_t \frac{\frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}{\sum_{i=1}^{T} \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}$$

- Muestreo por importancia pesado:

$$\sum_{i=1}^{T} X_t \frac{\frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}{\sum_{i=1}^{T} \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}$$

  - Estimador sesgado.

- Muestreo por importancia pesado:

$$\sum_{i=1}^{T} X_t \frac{\frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}{\sum_{i=1}^{T} \frac{\mathbf{P}_c\{X_t\}}{\mathbf{P}_u\{X_t\}}}$$

  - Estimador sesgado.
  - Menor varianza (coeficientes siempre <1).

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

  $$\pi(A_t \mid S_t)$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1})$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \ldots p(S_T \mid S_{T-1}, A_{T-1})$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \ldots p(S_T \mid S_{T-1}, A_{T-1})$$
$$= \prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \ldots p(S_T \mid S_{T-1}, A_{T-1})$$
$$= \prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)$$

- Importance sampling ratio:

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \ldots p(S_T \mid S_{T-1}, A_{T-1})$$
$$= \prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)$$

- Importance sampling ratio:

$$\rho_{t:T-1} = \frac{\prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)}{\prod_{k=1}^{T-1} b(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)}$$

Importance Sampling

- Probabilidad de observar secuencia $A_t, S_{t+1}, A_{t+1}, \ldots S_T$ a partir de S_t bajo política π:

$$\pi(A_t \mid S_t)p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \ldots p(S_T \mid S_{T-1}, A_{T-1})$$
$$= \prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)$$

- Importance sampling ratio:

$$\rho_{t:T-1} = \frac{\prod_{k=1}^{T-1} \pi(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)}{\prod_{k=1}^{T-1} b(A_k \mid S_k)p(S_{k+1} \mid S_k, A_k)} = \frac{\prod_{k=1}^{T-1} \pi(A_k \mid S_k)}{\prod_{k=1}^{T-1} b(A_k \mid S_k)}$$

Aplicando Importance Sampling

Aplicando Importance Sampling

- Queremos estimar $v_\pi(s) = \mathbb{E}_\pi \{G_t \mid S_t = s\}$, pero muestras G_t se generan con b.

Aplicando Importance Sampling

- Queremos estimar $v_\pi(s) = \mathbb{E}_\pi \{G_t \mid S_t = s\}$, pero muestras G_t se generan con b.
- Muestreo por importancia simple:

$$V(s) = \frac{1}{|\mathcal{T}(s)|} \sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t$$

Aplicando Importance Sampling

- Queremos estimar $v_\pi(s) = \mathbb{E}_\pi \{G_t \mid S_t = s\}$, pero muestras G_t se generan con b.
- Muestreo por importancia simple:

$$V(s) = \frac{1}{|\mathcal{T}(s)|} \sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t$$

- Muestreo por importancia pesado:

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Muestreo por importancia simple o pesado?

Muestreo por importancia simple o pesado?

- Simple: estimativo no sesgado, pero varianza puede ser infinita.

Muestreo por importancia simple o pesado?

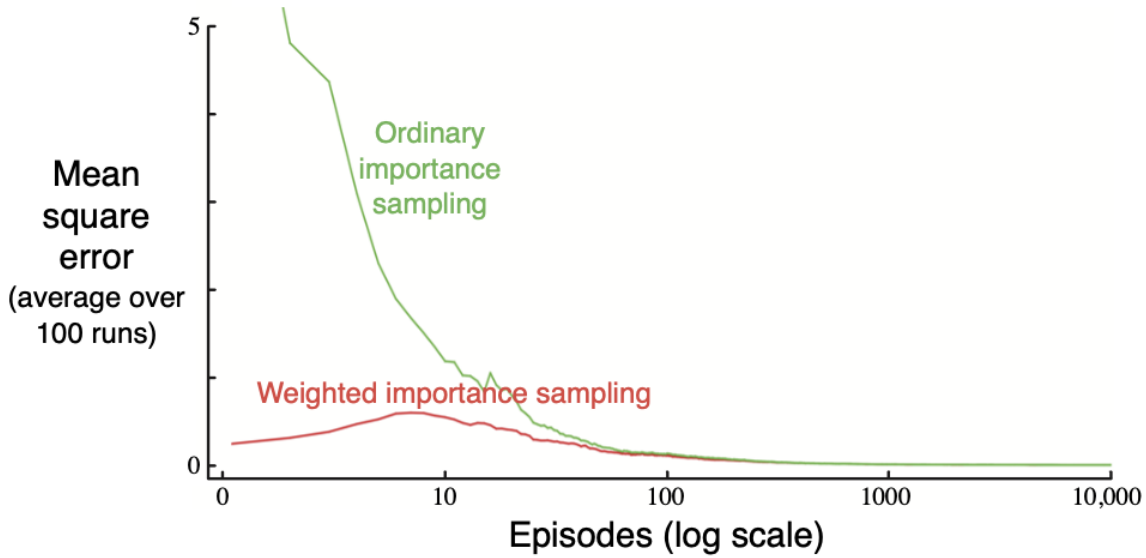
- Simple: estimativo no sesgado, pero varianza puede ser infinita.
- Pesado: sesgado, pero sesgo y varianza tienden a cero asimptóticamente.

Muestreo por importancia simple o pesado?

- Simple: estimativo no sesgado, pero varianza puede ser infinita.
- Pesado: sesgado, pero sesgo y varianza tienden a cero asimptóticamente.
- En la práctica pesado tiene menor varianza y es preferible.

Muestreo por importancia simple o pesado?

- Simple: estimativo no sesgado, pero varianza puede ser infinita.
- Pesado: sesgado, pero sesgo y varianza tienden a cero asimptóticamente.
- En la práctica pesado tiene menor varianza y es preferible.



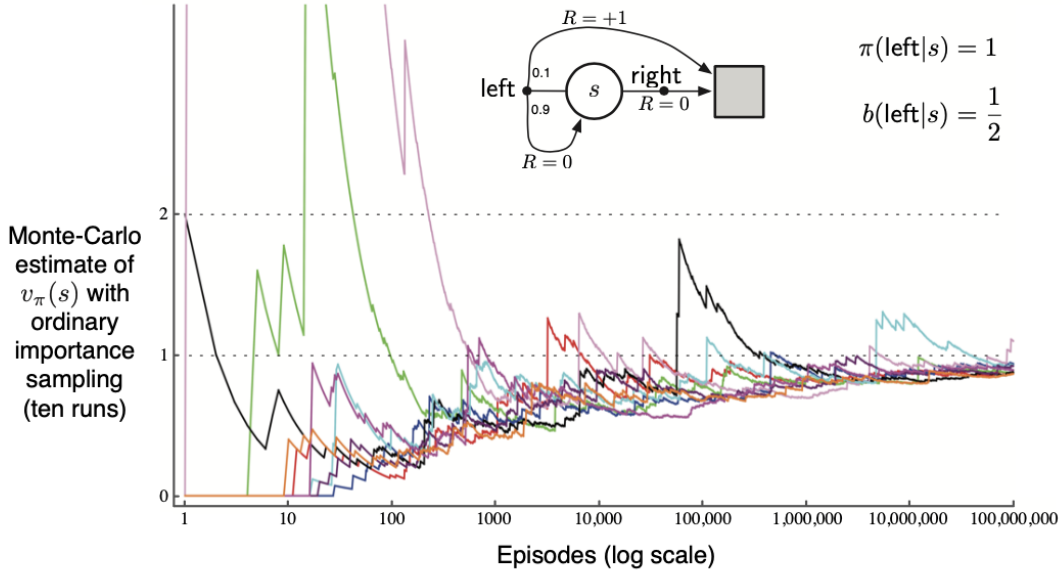


Figure 5.4: Ordinary importance sampling produces surprisingly unstable estimates on the one-state MDP shown inset (Example 5.5). The correct estimate here is 1 ($\gamma = 1$), and, even though this is the expected value of a sample return (after importance sampling), the variance of the samples is infinite, and the estimates do not converge to this value. These results are for off-policy first-visit MC.

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = \frac{V_n}{}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = \frac{V_n \sum_{k=1}^{n-1} W_k}{}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = \frac{V_n \sum_{k=1}^{n-1} W_k + W_n G_n}{}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = \frac{V_n \sum_{k=1}^{n-1} W_k + W_n G_n}{\sum_{k=1}^{n-1} W_k + W_n}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$\begin{aligned} V_{n+1} &= \frac{V_n \sum_{k=1}^{n-1} W_k + W_n G_n}{\sum_{k=1}^{n-1} W_k + W_n} \\ &= \frac{V_n(\sum_{k=1}^{n-1} W_k + W_n) + W_n G_n - V_n W_n}{\sum_{k=1}^{n-1} W_k + W_n} \end{aligned}$$

Promedio ponderado incremental

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$\begin{aligned} V_{n+1} &= \frac{V_n \sum_{k=1}^{n-1} W_k + W_n G_n}{\sum_{k=1}^{n-1} W_k + W_n} \\ &= \frac{V_n(\sum_{k=1}^{n-1} W_k + W_n) + W_n G_n - V_n W_n}{\sum_{k=1}^{n-1} W_k + W_n} \\ &= V_n + \frac{W_n}{C_n}\left[G_n - V_n\right] \end{aligned}$$

con $C_{n+1} = C_n + W_{n+1}$, y $C_0 = 0$

Estimación MC de q off-policy

Input: Política objetivo π

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s, a) \in \mathbb{R}$,

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
    $b \leftarrow$ política con cubrimiento de π.

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$
if $W = 0$ **then**

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$
if $W = 0$ **then** break
end if

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
$b \leftarrow$ política con cubrimiento de π.
Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
$G \leftarrow 0$, $W \leftarrow 1$
for $t = T-1, T-2, \ldots 0$ **do**
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$
if $W = 0$ **then** break
end if
end for

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$
    if $W = 0$ **then** break
    end if
  end for
until ∞

Estimación MC de q off-policy

Input: Política objetivo π
Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T$,
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $W \leftarrow W \frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$
    if $W = 0$ **then** break
    end if
  end for
until ∞

Control MC off-policy para estimar π^*

Incialice $Q(s, a) \in \mathbb{R}$,

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
    $b \leftarrow$ política con cubrimiento de π.

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

$\pi(s) \leftarrow \arg\max_a Q(s,a)$

repeat

$b \leftarrow$ política con cubrimiento de π.

Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    if $A_t \neq \pi(S_t)$ **then**

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    if $A_t \neq \pi(S_t)$ **then** break
    end if
    $W \leftarrow W \frac{1}{b(A_t \mid S_t)}$

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    if $A_t \neq \pi(S_t)$ **then** break
    end if
    $W \leftarrow W \frac{1}{b(A_t \mid S_t)}$
  end for

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    if $A_t \neq \pi(S_t)$ **then** break
    end if
    $W \leftarrow W \frac{1}{b(A_t \mid S_t)}$
  end for
until ∞

Control MC off-policy para estimar π^*

Incialice $Q(s,a) \in \mathbb{R}$, $C(s,a) \leftarrow 0\ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$
repeat
  $b \leftarrow$ política con cubrimiento de π.
  Episodio b : $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
  $G \leftarrow 0$, $W \leftarrow 1$
  for $t = T-1, T-2, \ldots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    if $A_t \neq \pi(S_t)$ **then** break
    end if
    $W \leftarrow W \frac{1}{b(A_t \mid S_t)}$
  end for
until ∞