Capítulo 5: Ejercicios.

Ej. 5.1: Considera los diagramas de la derecha en la figura 5.1.

Diagrama arriba derecha: el valor ~~del estado~~ de los estados en las últimas dos filas crece con independencia de lo que esté mostrando el dealer.

Cuando el jugador tiene 21 puntos, ha ganado o bien ha empatado, ~~cuando el dealer muestra~~ pero nunca podrá haber perdido.

Si el dealer muestra un As, entonces ~~en la~~ boca abajo puede tener entre 1 y 10. Si tiene 10, habrá empate. En cualquier otro caso, el jugador gana. Si tiene una carta boca arriba que no es as ni 10, entonces habrá ganado el jugado. Si tiene boca arriba una carta con valor de 10, entonces ~~ganará so~~ habrá empate sólo si ~~debe~~ boca abajo hay un as. Como hay menos ases que cartas que valen 10, ~~la situación~~ el estado (21, A) es menos ventajoso ~~que el~~, para el jugador que el estado (21, 10).

Si el jugador tiene 20, entonces puede ganar, perder o empatar. ~~Si~~ En el estado (20, A), sólo podrá perder si la otra carta es 10, y ~~empatará si no es~~ sólo podrá empatar ~~(no hay x / x+x=20~~ si lo que hay boca abajo es un 9. En el resto casos, gana el jugador.

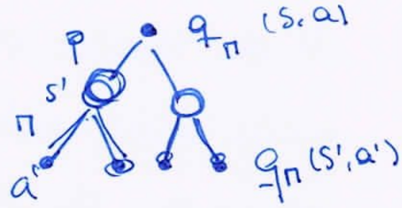
Si tenemos (20, ~~A~~10), entonces el empate sólo se da si hay otro 10 y perderá sólo si hay un as. Este estado es por tanto preferible a (20, A) para el jugador. ~~Tanto (20,10)~~ Además, (20, 10) < (21, 10) y (20, A) < (21, A).

Cuando el jugador tiene en total 19, entonces la política dicta que vuelva a pedir una carta. Esto aumenta drásticamente la probabilidad de pasarse y de que gane el dealer.

5.2. Supón que utilizamos every-visit MC en vez de first-visit MC en la tarea de blackjack. ¿Cambiará mucho la función?

En blackjack, los estados no se repiten. La primera vez que se visita un estado es todas las veces que se visita, por lo que en este caso los resultados serían los mismos.

5.3. ¿Backup diagram for MC estimation of q_π?



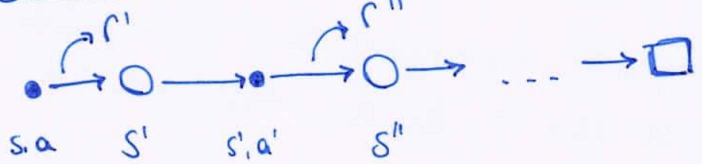
Para calcular q_π, usamos que:

$$q_\pi = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] =$$

$$= \mathbb{E}[R_{t+1} | S_t = s, A_t = a] + \gamma \cdot \mathbb{E}_\pi[\mathbb{E}_\pi[G_{t+2} | S_{t+1} = s', A_{t+1} = a']] =$$

$$= \mathbb{E}[R_{t+1} | S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[q_\pi(S_{t+1}, A_{t+1}) | S_t = s]$$

Por tanto, si generamos un episodio $S_0, A_0, R_1, S_1, A_1, R_2, \dots, R_T$, tenemos:



5.4. : The pseudocode for Monte Carlo ES is inefficient...

En la sección 2.4 habíamos visto que se puede escribir un promedio de forma incremental:

$$Q_n = \frac{\sum_i^n R_i}{\cancel{\sum_i^n} \; n} = \frac{1}{n}\sum_{i=1}^{n} R_i = \cancel{\frac{1}{n}R_1 + \frac{1}{n}\left[R_1 + \sum_{i=1}^{n} R_{i+1}\right]}$$

$$= \frac{1}{n}\left[R_n + \sum_{i=1}^{n-1} R_i\right] = \frac{1}{n} R_n + \frac{(n-1)}{n} Q_{n-1} =$$

$$= \cancel{\frac{1}{n}R_n + \frac{n}{n}(n-1)Q_{n-1}} = \cancel{\frac{1}{n}R_n + nQ_{n-1} - \frac{Q_{n-1}}{n}} =$$

$$= \cancel{nQ_{n-1} + \frac{1}{n}\left[R_n - Q_{n-1}\right]}$$

$$= Q_{n-1} + \frac{1}{n}\left[R_n - Q_{n-1}\right]$$

donde $R_n \equiv$ ~~Rewards~~ ~~Returns(S_t, A_t)~~ G_n en la i-ésima iteración.

Por tanto, haríamos, en vez de "append G to returns ...", lo siguiente:

- Mantenemos una cuenta del episodio en que estamos (n).

$$\left.\begin{aligned} n &\leftarrow n+1 \\ Q &\leftarrow Q + \frac{1}{n}\left[G - Q\right] \end{aligned}\right\}$$

Esto lo hacemos debajo del "if". Es decir, $n \leftarrow n+1$ no se ejecuta si no es la 1ª vez que (S_t, A_t) aparece en la secuencia!

Ejercicio 5.5. Sea un MDP:

$r = 1$; $\gamma = 1$.

What are the first-visit and every-visit estimators of the value of the non-terminal state?

Para el caso first-visit, tendríamos $V(s) = \mathbb{E}[\overset{10}{\cancel{G}}] = 10$.

Para el caso every-visit, tenemos $V(s) = \mathbb{E}[1, 2, 3, \dots, 10] = 5'5$

Ejercicio 5.6 ¿Cuál es la ec. análoga a 5.6 para $Q(s,a)$, supuestos dados los returns siguiendo la política b?

Tenemos $\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$

$\mathbb{E}_b[G_t | S_t, A_t] = q_b(s,a)$ / ¿$q_\pi(s,a)$?

Sabemos que $q_b(s,a) = \mathbb{E}[R_{t+1} + \gamma G_t(s) | S_t = s, A_t = a] =$

$= \sum_{s',r} p(s',r|s,a)[r + \gamma V_b(s')]$

y que $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')] =$

$= \sum_{s',r} p(s',r|s,a)[r + \gamma \rho_{t:T-1} V_b(s')] = \mathbb{E}[r + \gamma \rho$

$= \mathbb{E}[R_{t+1} + \gamma \rho_{t:T-1} V_b(S_{t+1}) | S_t = s, A_t = a]$

Ejercicio 5.6: ¿Ec. análoga a 5.6 para $Q(s,a)$?

En este caso, el return R_{t+1} no viene dado por la estrategia b, sino por la elección libre de "a". Por tanto, no debemos multiplicar a G_t por $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$, sino comenzar a usar ρ en $t+1$:

$$Q(s,a) \doteq \frac{\sum_{t\in\mathcal{T}(s,a)} \rho_{t+1:T(t)-1}\, G_t}{\sum_{t\in\mathcal{T}(s,a)} \rho_{t+1:T(t)-1}}$$

donde $\mathcal{T}(s,a)$ ahora contiene los sellos temporales en los que se dieron el par estado-acción (s,a). (Ref: Bryan Hayder).

Ejercicio 5.7. ¿Por qué creció y luego decreció el error en weighted-importance sampling aplicado a la figura 5.3?

Todas las trayectorias en las que existe alguna acción prohibida en π tienen $\rho=0$ y por tanto contribuyen con valor "0" a $V(s)$.

Con pocos episodios, ~~será más bajo~~ el número de trayectorias contribuyendo con ~~G~~ $\rho \cdot G \neq 0$ será bajo, y $V(s) \simeq 0$, que es cercano al valor real de $V(13)$.

Conforme aumenta el # de episodios, también hay un mayor de trayectorias con $\rho\neq 0$ y ~~V~~ $V(13)$ aumentará. En este punto aumenta el error ~~bias~~. Este crecimiento se detiene y revierte conforme la varianza comienza a decrecer.

5.8. Los resultados en el ~~ej~~ ejemplo 5.5 mostrados en la figura 5.4 usan first-visit MC. Supón que usaran every-visit MC. ¿Seguiría siendo la varianza infinita?

Tanto en este caso como en first-visit, nuestro objetivo es calcular el valor esperado del return escalado y al cuadrado.

~~Si Vestamos~~

El valor esperado de una variable aleatoria x puede obtenerse, o bien con la media muestral (es decir, generando muchos episodios, calculando G_t en cada episodio y ~~dividiendo en~~ luego sumando los g_t's obtenidos y dividiendo entre # de ~~ep~~ muestras). o bien ~~con~~ de forma exacta. Para obtenerla de forma exacta, es necesario multiplicar la probabilidad de cada posible valor de la variable aleatoria por dicho valor, y después sumar.

~~Si~~ Dado que conocemos la probabilidad de cada valor porque conocemos la dinámica del MDP, calculamos $\mathbb{E}_b[X^2]$ como

$\sum_i p_b(x_i) \cdot x_i^2$, donde x_i^2 es la variable aleatoria y $p_b(x_i)$ es la probabilidad de obtener x_i si se sigue la dinámica del MDP junto con la estrategia b:

$$\cancel{\mathbb{E}_b[X^2] = \sum \ldots} \sum b(s)$$

$$P_b(x_i) = P_b(\text{trayectoria}) \; t$$

Dada una trayectoria de longitud T, su probabilidad se obtiene multiplicando probs. de transición $p(s'|s,a)$ a los estados de la trayectoria y prob. de elección de las acciones ejecutadas en la muestra:

$$Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t, A_{t:T-1} \sim b\} = \prod_{k=t}^{T-1} b(A_k|S_k)\, p(S_{k+1}|S_k, A_k)$$

En el caso de every-time visit, si estamos en una trayectoria de longitud T, necesariamente habremos pasado T-1 veces por "s" (el único estado del problema). Por tanto, $\mathcal{T}(s)=\{1,2,\dots,T-1\}$ y tenemos

$$V(s) \doteq \frac{\sum_{t=1}^{T-1} \rho_{t:T(t)-1}\, G_t}{T-1} \qquad T \geq 2$$

Donde $G_t = G_0 = 1 \ \forall t$ y $T(t) = T \ \forall t$, con lo que V(s) queda:

$$V(s) \doteq \frac{\sum_{t=1}^{T-1} \rho_{t:T-1}\, G_0}{T-1} = \frac{\sum_{t=1}^{T-1} \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}\, G_0}{T-1}$$

Para obtener $\mathbb{E}_b[V(s)^2]$, habrá que multiplicar estos valores por la probabilidad de cada trayectoria. Dada una trayectoria de longitud T, decimos que

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \dots, S_T \mid S_t, A_{t:T-1} \sim b\} = \prod_{k=t}^{T-1} b(A_k|S_k)\, p(S_{k+1}|S_k, A_k)$$

Por tanto, tendremos que $\mathbb{E}_b[V(s)]^2 =$

$$= \sum_{T=2}^{\infty} \underbrace{\prod_{k=1}^{T-1} b(A_k|S_k)\, p(S_{k+1}|S_k, A_k)}_{\text{prob. trayectoria de longitud } T} \cdot \frac{1}{(T-1)^2} \left(\sum_{t=1}^{T-1} \rho_{t:T-1} \right)^2 \quad (*)$$

$t=1$ $\quad t=2$

$\circ \longrightarrow \square$

Ej.: $T=2$ $(k=1)$

$$b(A_1|S_1)\, p(S_2|S_1, A_1) \cdot \left(\frac{1}{1}\right)^2 \cdot \rho^2 = \frac{1}{2} \cdot 0.1 \cdot \left(\frac{1}{0.5}\right)^2$$

$T=3$ $(k=2)$

$$\cancel{b(A_1|S_1)\, p(S_2|S_1, A_1) \cdot b(A_2|S_2)\, p(S_3|S_2, A_2)} \cdot \left(\frac{1}{2}\right)^2 \cdot \left[\left(\frac{1}{0.5} + \frac{1}{0.5} \right)^2 \right] = \dots$$

$\frac{1}{0.5^2} = \cancel{2}\, 2^2 = 4$

$\ldots \left(\frac{1}{2}\right)^2 0.1 \cdot 0.9 \cdot \left(\frac{1}{2}\right)^2 \left[2 \cdot \frac{1}{0.5^2}\right]$

~~Este resultado~~ Nótese que $\rho_{t:T-1} = \prod_{t}^{T-1} 2 = 2^{(T-t)}$

(*) Manipulando la expresión anterior (*) queda:

$$\sum_{T=2}^{\infty} \prod_{k=1}^{T-1} \frac{1}{2} p(S_{k+1} | S_k, A_k) \left(\frac{1}{T-1}\right)^2 \left(\sum_{t=1}^{T-1} 2^{T-t}\right)^2 \ldots$$

... cuyo límite en el infinito, es también infinito por el término exponencial?

Ejercicio 5.11. En el algoritmo, cabría esperar ver el ratio

$\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ pero en su lugar aparece $\frac{1}{b(A_t|S_t)}$ ¿Por qué?

~~Estamos considerando único~~

Porque hacemos a $\pi(s)$ una política determinista. Concretamente, $\pi(\arg\max_a Q(S_t, a) | S_t) = 1.0$ y si $A_t \neq \pi(S_t)$, entonces $\pi(A_t|S_t) = 0$. Por tanto, para las acciones que maximizan $Q(S_t, a)$ (que son ~~a las que se llega cuando~~ con las que se llega a la línea de código que actualiza a W), $\pi(a|S_t) = 1$, de ahí la actualización

$$W \leftarrow W \frac{1}{b(A_t|S_t)}$$

Si se escoge una acción A_t que ~~no figura en $\pi(S_t)$~~ imposible en π, los cálculos se frenan y el episodio no cuenta (aunque sí sirve para ~~calcular Q_t~~ actualizar $Q(S_t, A_t)$)

Ej. 5.9: Modifica el algoritmo de first-visit MC para evaluación de estrategia usando la implementación incremental de la sección 2.4.

Offline:

Input: an arbitrary target policy π

Initialize:

$V(s) \in \mathbb{R}$ arbitrarily, for all $s \in \mathcal{S}$.

$n(s) = 0 \ \forall s \in \mathcal{S}$.

Loop forever (for each episode):

Generate episode following b: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \dots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$W \leftarrow 1$

Loop for each step of episode: $t = T-1, \dots, 0$:

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

$W \leftarrow \frac{\pi(A_t|S_t)}{b(A_t|S_t)} W$

Unless S_t appears in $S_0, \dots, S_{t-1}$:

$n(S_t) \leftarrow n(S_t) + 1$

$V(S_t) \leftarrow V(S_t) + \frac{W}{n(S_t)} [G - V(S_t)]$

Online:

Input: π (a predecir).

Initialize:

$V(s) \in \mathbb{R}$ arbitrarily for all $s \in \mathcal{S}$

$n(s) = 0 \ \forall s \in \mathcal{S}$

Loop forever:

Generate episode

$G \leftarrow 0$

for each step in episode: $t = T-1, \dots, 0$:

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

Unless S_t in $S_0, \dots, S_{t-1}$:

→

$n(S_t) \leftarrow n(S_t) + 1$

$$V(S_t) \leftarrow V(S_t) - \frac{1}{n(S_t)} [G - V(S_t)]$$

Ej. 5.10: Deriva la regla 5.8 de 5.7.

Sabemos del capítulo 2 que dado un valor $\cancel{G = \frac{\sum R_i}{n}}$

$$E[G_t | S_t] = E[\underbrace{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}}_{G_t} | S_t] \simeq \sum_{i=1}^{N} \frac{G_i}{N}$$

donde $G_i \equiv$ cada uno de los returns obtenidos al empezar en S_t en la muestra y $N \equiv$ nº veces en que se ha observado S_t en la muestra.

Por tanto, tenemos: $V_{n+1} = \frac{1}{n} \sum_{i=1}^{n} G_i = \frac{1}{n}\left(G_n + \sum_{i=1}^{n-1} G_i\right) =$

$$= \frac{1}{n}\left(G_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} G_i\right) = \frac{1}{n}\left(G_n + (n-1)V_n\right) =$$

$$= \frac{1}{n}\left(G_n + nV_n - V_n\right) = V_n + \frac{1}{n}[G_n - V_n]$$

En 5.7, lo que tenemos es $V_{n+1} = \dfrac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}$

Sabemos por tanto que

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \rightarrow \sum_{k=1}^{n-1} W_k G_k = \sum_{k=1}^{n-1} W_k V_n \quad (*)$$

Además, $\sum_{k=1}^{n} W_k G_k = \sum_{k=1}^{n-1} W_k G_k + W_n G_n$

Sustituyendo (*) en esta última igualdad:

$$\sum_{k=1}^{n} W_k G_k = \sum_{k=1}^{n-1} W_k V_n + W_n G_n \quad (a)$$

También sabemos que $\sum_{k=1}^{n-1} W_k = \sum_{k=1}^{n} W_k - W_n$ (**)

Sustituyendo (**) en (a):

$$\sum_{k=1}^{n} W_k G_k = V_n \cdot \left[\sum_{k=1}^{n} W_k - W_n \right] + W_n G_n \quad (b)$$

Manipulando:

$$\sum_{k=1}^{n} W_k G_k = \sum_{k=1}^{n} W_k V_n + W_n [G_n - V_n] \quad (c)$$

Y ahora sustituimos (c) en el ~~deno~~ numerador de 5.7:

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

$$= V_n + \frac{W_n}{\sum_{k=1}^{n} W_k} [G_n - V_n] = V_n + \frac{W_n}{C_n} [G_n - V_n] \quad (5.8)$$

donde $C_n = \sum_{k=1}^{n} W_k$