Ejercicio 1

Considere un proceso de decisión de Markov (MDP) con dos estados: uno terminal y otro no-terminal. Existe sólo una acción que lleva del estado no-terminal al estado terminal con probabilidad $1-\rho$ y del estado no-terminal a si mismo con probabilidad ρ. La recompensa es $+1$ en todas las transiciones y el factor de descuento es $\gamma = 1$. Suponga que observa un episodio con 10 iteraciones y un retorno de 10.

- ¿Cuáles son las estimaciones Monte Carlo de primer-visita y de cada-visita del valor del estado no-terminal basadas en ese episodio?
- Compare los valores de estado obtenidos con el teórico si $\rho = 0.9$. Saque conclusiones.

Solución

Ejercicio 2

Demostrar que la política ε - *Greedy*, definida de la siguiente manera

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A(s)|} & si \quad a = a^* \\ \\ \frac{\varepsilon}{|A(s)|} & si \quad a \neq a^* \end{cases} \tag{1}$$

es una distribución de probabilidad válida, donde $|A(s)|$ es el número de acciones para el estado s, $\varepsilon < 1$ es un número positivo pequeño, y a^* es la acción óptima (decisión greedy) para el estado s. ¿Hay que pedir alguna condición sobre ε?

Solución

Ejercicio 3

Se dice que una política π es ε - *soft* si $\pi(a \mid s) \geq \frac{\varepsilon}{|A(s)|}$ $\forall a \neq a^*$ y $\forall s$. Demostrar que la política ε - *Greedy* $\pi'(a \mid s) \geq \frac{\varepsilon}{|A(s)|}$ (definida en el Ejercicio 1) es igual o mejor que cualquier política ε - *soft*, es decir $v_{\pi'}(s) \geq v_\pi(s)$ $\forall s$.

Solución

Ejercicio 4

Dada una trayectoria de acciones y estados $A_t,\ S_t,\ A_{t+1},\ S_{t+1}, \ldots, A_T,\ S_T$, en un proceso de decisión Markoviano (MDP) bajo la política $\pi(a \mid s)$, demostrar que la probabilidad conjunta de esa trayectoria se puede escribir como:

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

Nota: considere $P[S_t] = 1$.

Solución

Ejercicio 5

Usando el resultado del Ejercicio 4, demostrar que el *importance-sampling ratio* correspondiente a la aplicación del método *off-policy* es:

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

Solución

Ejercicio 6

Demostrar la fórmula de la implementación incremental de un promedio ponderado. Es decir, el promedio ponderado

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

puede calcularse incrementalmente con la fórmula:

$$V_{n+1} = V_n + \frac{W_n}{C_n} \left[G_n - V_n \right],$$

con $C_{n+1} = C_n + W_{n+1}$

Solución
