Ejercicio 1

[Programación] Para el proceso de Markov de recompensas (MRP) de la figura 1, calcular los valores de los estados de forma iterativa con los siguientes algoritmos y compare sus convergencias. Considere factor de descuento $\gamma < 0.9$.

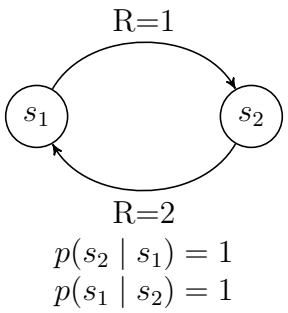

Figura 1: Grafo de transición de estados

- Actualizar todos los valores a la vez por iteración: $v_{k+1} = r + \gamma P v_k$, con $v_k\, r$ siendo los vectores y recompensas, respectivamente; y P la matriz de probabilidades de transición.
- Actualizar los valores de un estado por vex *(in place)*: $v_{k+1}(s') = r(s') + \gamma v_k(s)$, con $v_k(s)$ y $r(s')$ siendo los valores y recompensas correspondientes a los estados s y s', respectivamente.

Solución

Ejercicio 2

En el Ejemplo 4.1 (*GridWorld, Sutton&Barto, 2018*) [1], donde la política π es aleatoria y equiprobable:

- ¿Cuánto vale $q_\pi(11, down)$?
- ¿Cuánto vale $q_\pi(7, down)$?
  Nota: utilice $v(11) = -14$ (Fig. 4.1 del libro).

Justifique sus respuestas.

Solución

Ejercicio 3

En el Ejemplo 4.1 (*GridWorld, Sutton&Barto, 2018*) [1], suponga que se agrega un nuevo estado 15 del estado 13 y sus acciones: *left, up, right* y *down*, lleva al agente a los estados 12, 13, 14 y 15, respectivamente.

- Considere que las transiciones desde los estados originales no se cambian. ¿Cuánto vale $v_\pi(15)$ para la política π aleatoria y equiprobable? Utilice $v(12) = -22$, $v(13) = -20$, $v(14) = -14$ (Fig. 4.1 del libro)

Justifique su respuesta.

Solución

Ejercicio 4

En el Ejemplo 4.3 (*Gambler's problem, Sutton&Barto, 2018*) [1], la política óptima tiene una forma particular (ver figura 2) con máximo en 50. Es decir, cuando el jugador tiene \$50, le conviene apostarlo todo; sin embargo, cuando tiene \$51, le conviene apostar \$1. ¿Por qué sucede esto?

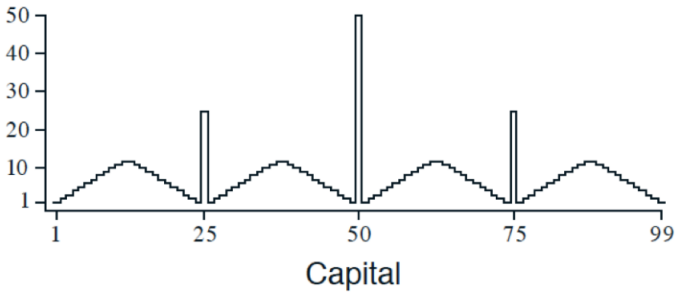

Figura 2: Ejemplo 4.3, Sutton&Barto, 2018

Solución

Ejercicio 5

[Programación] Implemente el Algoritmo de Iteración de Valores para el el Ejemplo 4.3 (*Gambler's problem, Sutton&Barto, 2018*) [1] para los siguientes casos:

- $P(CARA) = 0.25$.
- $P(CARA) = 0.55$.

Solución

Referencias

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.