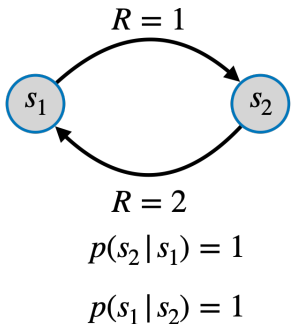
Guía 3: Programación Dinámica

Ejercicio 1

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



a) Actualizar todos los valores a la vez por iteración: $\mathbf{v}_{k+1} = \mathbf{r} + \gamma \mathbf{P}\mathbf{v}_k$, con $\mathbf{v}_k$, $\mathbf{r}$ siendo los vectores de valores y recompensas, respectivamente; y $\mathbf{P}$ la matriz de probabilidades de transición.
b) Actualizar el valor de un estado por vez (*in place*): $\mathrm{v}_{k+1}(s') = \mathrm{r}(\mathrm{s}') + \gamma \mathrm{v}_k(s)$, con $\mathrm{v}_k(s)$ y $\mathrm{r}(s')$ siendo los valores y recompensas correspondientes a los estados s y s', respectivamente.

Ejercicio 2

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

a) ¿Cuánto vale $q_\pi(11, down)$?
b) ¿Cuánto vale $q_\pi(7, down)$?
Nota: Utilice $\mathrm{v}(11) = -14$ (Fig. 4.1 del libro)

Justifique sus respuestas

Ejercicio 3

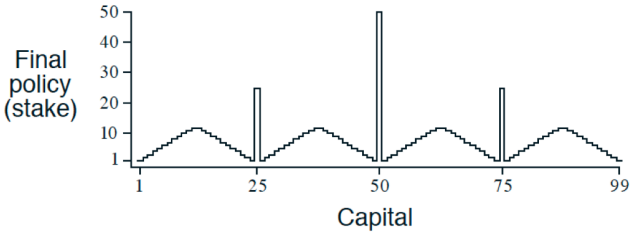
En el Ejemplo 4.1 (*GridWorld*, Sutton&Barto, 2018), suponga que se agrega un nuevo estado 15 debajo del estado 13 y sus acciones*: left*, *up*, *right* y *down*, lleva al agente a los estados 12, 13, 14 y 15, respectivamente.

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

Justifique su respuesta

Ejercicio 4

En el Ejemplo 4.3 (Gambler's problem, Sutton&Barto, 2018) la política óptima tiene una forma particular (ver figura) 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?



Ejercicio 5

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

a) P(CARA) = 0.25
b) P(CARA) = 0.55

Compare los resultados y saque conclusiones.