Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Industrial
IN79O - Modelos Estocásticos en Sistemas de Ingeniería
19 de noviembre de 2014

Pauta Control 3

Profesor: Raul Gouet.
Auxiliares: Ignacio Correa, Luis Fredes.

P1. (50 %) Armijo se levanta cada mañana y conecta su nuevo dispositivo *smart-health*, que mide el estado de ánimo y capacidad para trabajar ese día. Después de unos segundos el dispositivo entrega como resultado un número entero entre 1 y M, siendo 1 el nivel óptimo y M el peor. En ese momento Armijo va a trabajar si su nivel es óptimo, pero si su nivel es $i \in \{2, \dots, M-1\}$ puede decidir entre ir a trabajar, lo cual le reporta un ingreso esperado r_i, o bien ir a una jornada preventiva de descanso y relajo, que no le reporta ingreso pero lo deja completamente renovado, es decir, en nivel 1. Esta jornada además tiene un costo C_i (que depende de i). Por otra parte, cuando Armijo decide ir a trabajar, siendo $i \in \{1, \dots, M-1\}$ su nivel, entonces el nivel al día siguiente será $j \in \{i, i+1 \dots, M\}$, con probabilidad $p_{ij} \geq 0$. Es decir, el ánimo solo puede mantenerse o empeorar si no descansa. Si un día cualquiera Armijo despierta con nivel M entonces se genera automáticamente una llamada a un servicio de reanimación de urgencia, que lo lleva desde su domicilio hasta un centro donde debe permanecer dos días en recuperación intensiva, al cabo de los cuales sale completamente renovado. Este tratamiento tiene un costo total C_e, mucho mayor que los C_i, y durante este par de días tampoco genera ingresos. Suponga que $r_1 > r_2 > \cdots > r_{M-1}$, $C_2 < C_3 < \cdots < C_{M-1}$ y $\sum_{j=i}^{M} p_{ij} = 1$.

(**a**) (3 pts.) Modele la situación que enfrenta Armijo diariamente, como un proceso de decisión markoviano con horizonte finito N, especificando detalladamente todos los elementos del modelo, es decir, espacio de estados E, conjuntos de acciones A_i, probabilidades de transición y recompensas[1]. Suponga que la recompensa final es $f_i > 0$, como función del nivel $i < M$. La recompensa final cuando Armijo está en descanso o tratamiento es 0.

Solución: Modelaremos el problema como sigue: definimos el horizonte como $T = \{1, \dots, N\}$, el espacio de estados como $E = \{1, \dots, M\} \cup \{e\}$, las acciones serán $A_s = \{T, D\}$ (por «trabajar» y «descansar») cuando $s \neq 1, M, e$. Para las restantes, $A_1 = \{T\}$, $A_M = A_e = \{D\}$[2].

Las probabilidades de transición son

$$p(j \mid i, a) = \begin{cases} p_{ij} & \text{si } a = T, i \leq M-1, j \in \{i, \dots, M\}, \\ 1 & \text{si } i = M, j = e, a = D, \\ 1 & \text{si } i = e, j = 1, a = D, \\ 1 & \text{si } i \in \{2, \dots, M-1\}, j = 1, a = D, \\ 0 & \text{en otro caso.} \end{cases}$$

Las recompensas terminales son

$$r_N(i) = \begin{cases} f_i & \text{si } i < M, \\ 0 & \text{si } i = M \text{ o } i = e, \end{cases}$$

y para un tiempo $t < N$ son $r_t(1, a) = r_1$, $r_t(M, a) = -C_e$, $r_t(e, a) = 0$ y[3]

$$r_t(i, a) = \begin{cases} r_i & \text{si } a = T, \\ -C_i & \text{si } a = D. \end{cases}$$

(**b**) (3 pts.) Escriba las ecuaciones de optimalidad de Bellman en general. Luego adáptelas a la situación modelada en el apartado (a) y escriba explícitamente la solución $u_k^*(h_k)$ para $k = N$ y $k = N-1$.

[1] Se sugiere considerar un estado que represente la situación de la recuperación intensiva.
[2] Se tuvo en cuenta modelamientos distintos, como acciones vacías o dos estados especiales.
[3] Acá también se consideró el pagar «en dos cuotas», es decir, $\frac{C_e}{2}$ por día.

Solución: Las ecuaciones de optimalidad de Bellman generales son:

$$u_N^*(h_N) = r_N(i_N) \quad \forall i_N \in H$$
$$u_k^*(h_k) = \max_{a \in A_{i_k}} \left\{ r_k(i_k, a) + \sum_{j \in E} u_{k+1}(h_k, a, j) p(j \mid i_k, a) \right\},$$

en donde $h_k = (h_{k-1}, a_{k-1}, i_k) \in H_k$. Reemplazando en los datos de problema, tenemos que

$$u_N^*(h_N) = \begin{cases} f_i & \text{si } i < M, \\ 0 & \text{si } i = M \text{ o } i = e, \end{cases}$$
$$u_{N-1}^*(h_{N-1}) = \max_{a \in \{T,D\}} \left\{ r_{N-1}(i, a) + \sum_{j \in E} r_N^*(j) p(j \mid i, a) \right\}.$$

Ahora nos colocaremos en casos, dependiendo del estado en el cual se está en el instante $N-1$:

- Si $i = M$, entonces sólo hay una acción posible, descansar, y sólo se puede ir a e. Por lo tanto
$$u_{N-1}^*(M) = r_{N-1}(M, D) + \sum_{j \in E} r_N^*(j) p(j \mid M, D) = -C_e + r_N(e) = -C_e,$$
- Si $i = e$, entonces sólo hay una acción posible, descansar, y sólo se puede ir a 1. Por lo tanto
$$u_{N-1}^*(e) = r_{N-1}(e, D) + \sum_{j \in E} r_N^*(j) p(j \mid e, D) = 0 + r_N(1) = f_1,$$
- Si $i = 1$, entonces sólo hay una acción posible, trabajar. Por lo tanto
$$u_{N-1}^*(1) = r_{N-1}(1, T) + \sum_{j \in E} r_N^*(j) p(j \mid 1, D) = r_1 + \sum_{j \neq e} p_{1j} f_j,$$
- Si $i \in \{2, \ldots, M-1\}$, hay que diferenciar las dos acciones. Por lo tanto
$$u_{N-1}^*(i) = \max \left\{ r_{N-1}(i, T) + \sum_{j \in E} r_N^*(j) p(j \mid i, T), r_{N-1}(i, D) + \sum_{j \in E} r_N^*(j) p(j \mid i, D) \right\}$$
$$= \max \left\{ r_i + \sum_{\substack{j \leq i \\ j \neq e}} p_{ij} f_j, -C_i + f_1 \right\}.$$

(c) (1 pt.) Evalúe numéricamente $u_{N-1}^*(h_{N-1})$ a partir de los siguientes datos: $M = 3$, $C_2 = 1$, $C_e = 10$, $r_1 = 5$. $r_2 = 4$, $f_1 = 2$, $f_2 = 1$, $p_{11} = 0{,}9$, $p_{12} = 0{,}1$, $p_{22} = 0{,}8$, $p_{23} = 0{,}2$.

Solución: Aquí basta evaluar usando las expresiones antes encontradas:

- Si $i = M$, entonces $u_{N-1}^*(M) = -C_e = -10$,
- Si $i = e$, entonces $u_{N-1}^*(e) = f_1 = 2$,
- Si $i = 1$, entonces
$$u_{N-1}^*(1) = r_1 + \sum_{j \neq e} p_{1j} f_j = 5 + 0{,}9 f_1 + 0{,}1 f_2 = 5 + 0{,}9 \cdot 2 + 0{,}1 \cdot 1 = 6{,}9$$
- Si $i \in \{2, \ldots, M-1\}$ entonces
$$u_{N-1}^*(i) = \max \left\{ r_i + \sum_{\substack{j \leq i \\ j \neq e}} p_{ij} f_j, -C_i + f_1 \right\} = \max \{4 + p_{22} f_2, -1 + 2\} = \max\{4 + 0{,}8, 1\} = 4{,}8.$$

P2. (50 %) Pedro y Juan (primos lejanos de Armijo) juegan el clásico juego de monedas coincidentes, que se describe así: inicialmente cada uno tiene n monedas equilibradas y en la primera vuelta ambos lanzan una de sus monedas. Si al caer ambas monedas muestran lo mismo, entonces Pedro conserva su moneda y gana la moneda de Juan. En caso contrario Juan conserva su moneda y gana la moneda de Pedro. Se sabe que Juan está dispuesto a jugar hasta que uno de los dos se arruine pero Pedro tiene el privilegio de detener el juego cuando le plazca, antes de que alguno de los dos se arruine. Es decir, el juego debe parar si alguien se arruina o bien, si Pedro lo decide. Cuando el juego se detiene, cada jugador conserva las monedas que tenga en ese momento. Note que mientras se juega, la recompensa esperada en cada jugada es 0 pero al detenerse el juego, ésta es el número de monedas que tiene en ese momento.

(**a**) (4 pts.) Describa el problema de decisión markoviana que enfrenta Pedro y especifique todos los elementos que intervienen, es decir, espacios, probabilidades de transición, recompensas, etc.
Solución:
Definimos el horizonte como $T = \{1, \dots\}$, osea un horizonte infinito, el que representa cada lanzamiento de moneda. El espacio de estados como $S = \{0, \dots, 2n\} \cup \{\Delta\}$ que guarda la fortuna de pedro cada vez que lanza la moneda considerando el caso especial donde el juego esta parado, para esto se utiliza el estado sumidero. Las acciones serán $A_s = \{C, P\}$ (por «continuar» y «parar») cuando $s \neq 0, 2n, \Delta$, $A_s = \{P\}$ para $s = 0, 2n$ y $A_\Delta = \{\delta\}$ que lo obliga a quedarse en este estado de sumidero [4].

Las probabilidades de transición son

$$p(j \mid i, a) = \begin{cases} \frac{1}{2} & \text{si } i \in \{1, \dots, 2n-1\},\ j \in \{i-1, i+1\},\ a = C \\ 1 & \text{si } i \in \{0, \dots, 2n\}, j = \Delta, a = P, \\ 1 & \text{si } i = \Delta, j = \Delta, a = \delta, \\ 0 & \text{en otro caso.} \end{cases}$$

Las recompensas para un tiempo t son $r_t(i, P) = i \quad \forall i \in \{0, \dots, 2n\}$, $r_t(i, c) = 0 \quad \forall i \in \{1, \dots, 2n-1\}$ y $r_t(\Delta, \delta) = 0$.

$$r_t(i, a) = \begin{cases} i & \text{si } i \in \{0, \dots, 2n\},\ a = P \\ 0 & \text{si } i \in \{1, \dots, 2n-1\},\ a = C, \\ 0 & \text{si } i = \Delta, a = \delta, \end{cases}$$

(**b**) (3 pts.) Suponga que se pone como horizonte un número N de jugadas, es decir, si ninguno se ha arruinado ni Pedro ha detenido el juego, entonces en la jugada N todo se acaba y como recompensa terminal, cada jugador conserva sus monedas. Muestre que las ecuaciones de Bellman se escriben como $u_N^*(i) = i$. $u_k^*(0) = 0$, $u_k^*(2n) = 2n$ y

$$u_k^*(i) = \text{máx} \left\{ i, \frac{u_{k+1}^*(i-1) + u_{k+1}^*(i+1)}{2} \right\},\ \forall 0 < i < 2n,\ 1 \leq k \leq N-1$$

Concluya que $u_k^*(i) = i$ para $i \in \{0, \dots, n\}$, $k \in \{1, \dots, N\}$. Intuitivamente, ¿en qué momento piensa usted que Pedro debería detener el juego (comente)? Tome en cuenta que hay total simetría de los jugadores en las probabilidades de ganar cada lanzamiento.
Solución:
En esta parte hay que cambiar el problema a uno de horizonte finito, así $T = \{1, \dots, N\}$. Además de esto hay que agregar que las recompensas terminales son

$$r_N(i) = i \quad \forall i \in \{0, \dots, 2n\} \quad \text{y} \quad r_N(\Delta) = 0$$

Luego tenemos que si en algún paso llegamos a 0 ó $2n$ debemos parar y luego nuestras ecuaciones de Bellman nos dicen que en estos casos extremos optimizamos sobre un conjunto de solo un elemento, por lo que $u_k^*(0) = 0$ y $u_k^*(2n) = 2n$. Además de esto nos dice que en el instante N no tomamos

[4] Se tuvo en cuenta modelamientos distintos, como acciones vacías o estados con dos coordenadas.

decisiones por lo que $u_N^*(i) = r_N(i) = i$ y $u_N^*(\Delta) = r_N(\Delta) = 0$. Además de esto las ecuaciones establecen que para $k \in \{1, \ldots, N-1\}$

$$\begin{aligned} u_k^*(i) &= \max_{a \in A_i} \left\{ r_k(i,a) + \sum_{j \in S} p_k(j|i,a) u_{k+1}^*(j) \right\} \\ &= \max \left\{ r_k(i,P) + \sum_{j \in S} p_k(j|i,P) u_{k+1}^*(j), r_t(i,C) + \sum_{j \in S} p_k(j|i,C) u_{k+1}^*(j) \right\} \\ &= \max \left\{ i, \frac{1}{2} u_{k+1}^*(i-1) + \frac{1}{2} u_{k+1}^*(i+1) \right\} \end{aligned}$$

Demostremos que $u_k^*(i) = i$ Notemos que dado los valores de u_N^* tenemos que gracias a la formula recién obtenida

$$\begin{aligned} u_{N-1}^*(i) &= \max \left\{ i, \frac{1}{2} u_N^*(i-1) + \frac{1}{2} u_N^*(i+1) \right\} \\ &= \max \left\{ i, \frac{1}{2}(i-1) + \frac{1}{2}(i+1) \right\} \\ &= \max \left\{ i, i \right\} \\ &= i \end{aligned}$$

Mediante inducción retrograda supongamos que la propiedad se cumple para $t = k$, probaremos que en base a esto se cumple para $t = k-1$, luego

$$\begin{aligned} u_{k-1}^*(i) &= \max \left\{ i, \frac{1}{2} u_k^*(i-1) + \frac{1}{2} u_k^*(i+1) \right\} \\ &= \max \left\{ i, \frac{1}{2}(i-1) + \frac{1}{2}(i+1) \right\} \\ &= \max \left\{ i, i \right\} \\ &= i \end{aligned}$$

Finalmente se tiene que $u_1^*(n) = n$, luego es intuitivo que como en su primer turno se encuentra en n y deteniéndose de inmediato consigue la cantidad asociada a un pago óptimo, entonces lo natural es que se retire de inmediato (una política óptima); de todos modos, se consideraron buenas otras respuestas intuitivas justificadas en base a los resultados obtenidos.