Miscelánea Matemática **66** (2018) 37-56 SMM

Una introducción a los procesos de decisión markovianos con costo promedio

Óscar Vega Amaya
Departamento de Matemáticas
Universidad de Sonora
ovega@mat.uson.mx

1. Introducción

Los procesos de decisión o control markovianos modelan sistemas dinámicos estocásticos controlados, es decir, sistemas cuya evolución está sujeta a factores aleatorios y que puede modificarse por medio de la selección de ciertas variables de decisión o de control. Este tipo de modelos surgen en un sinnúmero de áreas de la ciencia y la ingeniería como finanzas [6, 40, 42], economía [27, 44], robótica [14], programas de salud [17, 30, 35], redes de comunicación [1, 16, 32, 56], aprovechamiento de recursos naturales (pesqueros [2, 46, 47]); acuíferos [28, 54, 55]; petroleros [5, 53]), sistemas de transporte [41], sistemas de producción e inventario [15], programas de mantenimiento y remplazo de equipo [26, 34], etc. Las referencias [10, 19, 20, 22, 50, 51, 52] ofrecen extensas listas de referencias sobre applicaciones de los procesos de decisión markovianos.

Fue Richard E. Bellman (1920-1984) quien abrió el campo de estudio de los procesos de decisión markovianos. Su interés por esta clase de procesos surgió durante las primeras visitas que hizo a Rand Corporation a finales de los 40's del siglo pasado, y en un corto plazo terminaron convirtiéndose en su principal proyecto de investigación. En su primer trabajo sobre el tema, publicado en 1952, Bellman [8] declara que está interesado en

> ... una clase de problemas matemáticos que surgen en situaciones en las que se requiere la ejecución de una sucesión finita o infinita de operaciones para alcanzar un resultado deseado.

Bellman se refirió a tales problemas como «programas dinámicos» y al enfoque que utilizó para resolverlos le llamó «programación dinámica». Posteriormente, en su trabajo de 1957, Bellman [9] usa el nombre

«proceso de decisión markoviano» para referirse a un programa dinámico cuyo mecanismo de evolución es aleatorio y satisface una propiedad «markoviana o de pérdida de memoria». Actualmente, la literatura sobre el tema también puede aparecer con alguno de los siguientes títulos: «procesos de decisión de Markov», «procesos de control de Markov», «procesos controlados de Markov», «programación dinámica estocástica», o simplemente como « programación dinámica», usándose este último nombre indistintamente para sistemas deterministas o estocásticos.

Bellman recibió por su invención y desarrollo de los procesos de decisión markovianos innumerables reconocimientos y premios [4, 13]. Además de su agenda de investigación, Bellman desarrolló una intensa actividad editorial: fundó las revistas de investigación «Mathematical Biosciences» y «Journal of Mathematical Analysis and Applications»; también fue editor fundador de la serie «Mathematics in Science and Engineering» de la editorial Academic Press. Las reseñas [4, 7, 12, 13, 29] dan cuenta del surgimiento de la programación dinámica y aportan elementos biográficos sobre Bellman, así como de sus distintos intereses matemáticos y científicos.

El núcleo de la teoría de los procesos de decisión markovianos es el problema de control óptimo, es decir, el de la selección de las variables de decisión que lleven a una evolución óptima del sistema estocástico con respecto a algún criterio de funcionamiento o desempeño. El estudio de tal problema descansa en la teoría de procesos estocásticos, particularmente, en la teoría de procesos de Markov–en tiempo discreto y continuo–y hecha mano de distintas ramas de las matemáticas como análisis real, teoría de integración abstracta de Lebesgue, topología, análisis funcional, teoría de matrices, programación lineal y no-lineal, entre otras.

El propósito del presente trabajo es dar una introducción breve y elemental a esta clase de procesos y para ello elegimos el problema de control óptimo en costo promedio. Presentaremos dos de las contribuciones centrales para este problema, la ecuación de optimalidad y el algoritmo de iteración de políticas, pero considerando únicamente modelos con espacios de estados y controles finitos. En la sección 4, se muestra que la validez de la ecuación de optimalidad garantiza la existencia de una política estacionaria óptima; por otra parte, el algoritmo de iteración de políticas, el cual se presenta en la sección 6, es un procedimiento iterativo para encontrar dicha política numéricamente.

2. El problema de control óptimo

Para plantear el problema de control óptimo se requiere especificar el modelo de decisión o control, la familia de políticas de control admisibles y el índice de funcionamiento que se desea optimizar. En esta sección se introducen brevemente estos elementos; el lector puede encontrar una descripción detallada en las referencias [10, 19, 20, 22].

Modelo de decisión. Un **modelo de decisión markoviano** consiste en la colección

$$\mathcal{M} = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C),$$

donde:

(a) $\mathbf{X}$ es el **espacio de estados**, el cual supondremos es un conjunto discreto (es decir, es un conjunto finito o numerable);

(b) $\mathbf{A}$ es el **espacio de controles**, el cual también supondremos es un conjunto discreto;

(c) Para cada $x \in \mathbf{X}$, el conjunto $A(x)$ representa al conjunto de los **controles admisibles** para el estado $x \in \mathbf{X}$. Al conjunto

$$\mathbb{K} := \{(x, a) \in \mathbf{X} \times \mathbf{A} : a \in A(x), x \in \mathbf{X}\}$$

se le llama conjunto de los **pares admisibles estado-control**. Un selector de $\mathbf{X}$ en $\mathbf{A}$ es una función $f : \mathbf{X} \to \mathbf{A}$ que satisface la restricción $f(x) \in A(x)$ para cada $x \in \mathbf{X}$. La clase de los selectores se denota por $\mathbb{F}$.

(d) Q, la **ley de evolución del sistema**, es una probabilidad de transición sobre $\mathbf{X}$ dado $\mathbb{K}$, esto es,

(i) $Q(y|x, a) \geq 0$ para todo $(x, a) \in \mathbb{K}, y \in \mathbf{X}$;

(ii) $\sum_{y \in \mathbf{X}} Q(y|x, a) = 1$ para todo $(x, a) \in A(x)$.

(e) C, la **función de costo por etapa**, es una función de $\mathbb{K}$ en $\mathbb{R}$.

Un modelo de decisión o control markoviano representa a un sistema estocástico que evoluciona en tiempo discreto de la siguiente manera: en el tiempo $n = 0$, el controlador observa el estado del sistema $x_0 = x \in \mathbf{X}$ y elige un control $a_0 = a \in A(x)$ con un costo de operación $C(x, a)$. Luego, el sistema transita a un nuevo estado $x_1 = y \in \mathbf{X}$ con probabilidad $Q(y|x, a)$, esto es,

$$Q(y|x, a) = \Pr[x_1 = y | x_0 = x, a_0 = a].$$

Una vez que se da la transición, el controlador elige otra vez un control $a_1 = b \in A(y)$ con un costo $C(y, b)$, y el sistema se mueve a un estado $x_2 = z$ con probabilidad $Q(z|y, b)$. Este proceso se repite una y otra vez hasta cierto tiempo $N \leq \infty$ conocido como **horizonte de planeación**.

Si $N < \infty$ se dice que el problema de control tiene horizonte finito; en caso contrario, es un problema en horizonte infinito.

Una **historia admisible** del sistema controlado hasta el tiempo $n \in \mathbb{N}_0$ es un vector de la forma

$$h_n = (x_0, a_0, x_1, a_1, \ldots, x_{n-1}, a_{n-1}, x_n)$$

donde $(x_k, a_k) \in \mathbb{K}$ para $k = 1, \ldots, n-1$, y $x_n \in \mathbf{X}$. Al conjunto de las historias admisibles hasta el tiempo $n \in \mathbb{N}_0$ lo denotaremos como $\mathbf{H}_n$. Note que $\mathbf{H}_0 = \mathbf{X}$ y que $\mathbf{H}_n = \mathbb{K}^n \times \mathbf{X}$ para cada $n \in \mathbb{N}_0$.

La distribución del proceso estocástico controlado $\{(x_n, a_n)\}$ generado de esta manera satisface la siguiente propiedad de Markov o de pérdida de memoria:

$$\begin{aligned} Q(y|x,a) &= \Pr\left[x_{n+1} = y | x_n = x, a_n = a\right] \\ &= \Pr\left[x_{n+1} = y | h_n, x_n = x, a_n = a\right] \end{aligned} \tag{1}$$

para todo $h_n \in \mathbb{H}_n, (x,a) \in \mathbb{K}, y \in \mathbf{X}, n \in \mathbb{N}$. Esta propiedad se puede formular equivalentemente de la siguiente forma:

$$\begin{aligned} \sum_{y \in \mathbf{X}} v(y) Q(y|x,a) &= E[v(x_{n+1}) | x_n = x, a_n = a], \\ &= E[v(x_{n+1}) | h_n, x_n = x, a_n = a], \end{aligned} \tag{2}$$

para aquellas funciones $v : \mathbf{X} \to \mathbb{R}$ tales que la serie del lado izquierdo sea convergente.

En muchas aplicaciones la evolución del sistema controlado no se expresa directamente con probabilidades de transición sino como un sistema dinámico

$$x_{n+1} = F(x_n, a_n, w_n), \quad n \in \mathbb{N},$$

$$x_0 = x \in \mathbf{X},$$

donde F es una función de $\mathbb{K} \times \mathbb{W}$ a $\mathbf{X}$ y la sucesión $\{w_n\}$ esta formada por variables aleatorias independientes e idénticamente distribuidas que toman valores en un conjunto discreto $\mathbb{W}$.

La dinámica de estos sistemas también se puede representar en términos de probabilidades de transición. Para ver esto, denotemos por $\rho(\cdot)$ a la función de probabilidad (común) de las variables $\{w_n\}$, es decir, $\rho(w) := \Pr[w_n = w]$, $w \in \mathbb{W}, n \in \mathbb{N}$. Entonces, la ley de evolución del sistema está dada como

$$\begin{aligned} Q(y|x,a) &:= \Pr\left[x_{n+1} \in B | x_n = x, a_n = a\right], \\ &= \Pr\left[F(x, a, w_n) \in B\right], \\ &= \sum_{w \in W_{x,a}} \rho(w), \end{aligned}$$

donde $W_{x,a} := \{w \in \mathbb{W} : F(x,a,w) \in B\}$.

Para estos sistemas los costos de operación en un paso se expresan como $\widehat{C}(x_k, a_k, w_k)$ donde $\widehat{C}$ es una función de $\mathbb{K} \times \mathbb{W}$ a $\mathbb{R}$. En estos casos la función de costo por etapa se obtiene como

$$C(x,a) := \sum_{w \in \mathbb{W}} \widehat{C}(x,a,w)\rho(w), \ (x,a) \in \mathbb{K}.$$

Políticas de control admisibles. En términos generales, una política de control admisible $\pi = \{\pi_n\}$ es una sucesión de «reglas» para elegir controles admisibles en cada etapa de decisión y se clasifican como dependientes de la historia, aleatorizadas, deterministas, markovianas o estacionarias. Para simplificar la exposición nos restringiremos a la clase de las políticas markovianas deterministas, las cuales se introducen a continuación.

Una política de control markoviana (determinista) $\pi = \{f_n\}$ es una sucesión de selectores de $\mathbf{X}$ en $\mathbf{A}$. Si $f_n = f \in \mathbb{F}$ para todo $n \in \mathbb{N}_0$ diremos que la política π es estacionaria y la identificaremos con el selector f. Denotaremos por Π_m a la clase de las políticas markovianas e identificaremos a la clase de las políticas estacionarias con la clase de los selectores $\mathbb{F}$.

Para facilitar la presentación en la partes subsiguientes de este trabajo intoducimos la siguiente notación. Para cada $f \in \mathbb{F}$ y cada $x \in \mathbf{X}$ defina

$$Q_f(y|x) := Q(y|x, f(x)), \tag{3}$$

$$C_f(x) := C(x, f(x)). \tag{4}$$

En general, para una función $u : \mathbb{K} \to \mathbb{R}$ usaremos la notación

$$u_f(x) := u(x, f(x)), \quad x \in X, f \in \mathbb{F}.$$

Además denotaremos por Q_f a la matriz con coeficientes $\{Q_f(y|x)\}_{x,y \in \mathbf{X}}$.

Para cada política $\pi = \{f_n\} \in \Pi_m$, de la propiedad (1) se sigue que el proceso de estados $\{x_n\}$ es una cadena de Markov no-homogénea con matrices de transición en un paso $Q_{f_n}, n \in \mathbb{N}_0$. Si la política es estacionaria, es decir, $f_n = f \in \mathbb{F}$ para todo $n \in \mathbb{N}_0$, entonces $\{x_n\}$ es una cadena de Markov homogénea con matriz de transición en un paso Q_f.

Índice de funcionamiento. Suponga que el estado inicial del sistema es $x_0 = x \in \mathbf{X}$. La aplicación de una política de control $\pi = \{f_n\} \in \Pi_m$ genera el flujo de costos aleatorios

$$C(x_0, a_0), C(x_1, a_1), \ldots, C(x_n, a_n), \ldots$$

donde $a_n = f_n(x_n)$ para todo $n \in \mathbb{N}_0$. Estos costos se pueden acumular de distintas maneras para medir el funcionamiento de sistema, dando así

origen a distintos índices de funcionamiento. Usualmente, estos índices se definen en términos del flujo de costos esperados

$$E_x^\pi C(x_0, a_0), E_x^\pi C(x_1, a_1), \ldots, E_x^\pi C(x_n, a_n), \ldots$$

donde E_x^π denota la de esperanza matemática–o valor esperado–dado que el estado inicial es $x_0 = x$ y la política de control que se aplica es $\pi = \{f_n\}$. Los índices o criterios de funcionamiento más comunes son los siguientes:

- **costo total (esperado) en N etapas con descuento** $\alpha \in (0, 1]$:

$$V_{N,\alpha}(\pi, x) := \sum_{k=0}^{N-1} \alpha^k E_x^\pi [C(x_k, a_k) + \alpha^N C_T(x_N)], \tag{5}$$

  donde la función $C_T : \mathbf{X} \to \mathbb{R}$ representa un costo terminal, es decir, por parar el proceso de control.
- **costo total (esperado) en horizonte infinito:**

$$V(\pi, x) := \sum_{k=0}^{\infty} E_x^\pi C(x_k, a_k). \tag{6}$$

- **costo total (esperado) en horizonte infinito con descuento** $\alpha \in (0, 1)$**:**

$$V_\alpha(\pi, x) := \sum_{k=0}^{\infty} \alpha^k E_x^\pi C(x_k, a_k). \tag{7}$$

- **costo promedio (esperado) por etapa:**

$$J(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} E_x^\pi C(x_k, a_k) \tag{8}$$

Problema de control óptimo. Una vez que se ha especificado un índice de funcionamiento $H(\pi, x)$–ya sea alguno de los anteriores o cualquier otro–el problema de control óptimo consiste en encontrar una política $\pi^* \in \Pi_m$ que satisfaga la igualdad

$$H(\pi^*, x) = H^*(x) := \inf_{\pi\in\Pi_m} H(\pi, x) \quad \forall x \in \mathbf{X}.$$

De los índices anteriores, el índice en costo promedio (8) es el que plantea los mayores retos matemáticos para la solución del problema de control óptimo. De hecho, se considera que dicho problema no es un solo problema, sino una familia de problemas [3], y se tienen ejemplos en los que el problema de control óptimo no tienen solución, o si la tiene, la solución no es «buena»; las referencias [36, 38] proporcionan ejemplos que ilustran esta afirmación. En general, para garantizar que existen

soluciones para este problema se requiere que el modelo de control satisfaga una serie de propiedades que pueden considerarse restrictivas. Sin embargo, para modelos con espacios de estados y controles finitos, sin imponer alguna condición adicional, se ha demostrado la existencia de políticas estacionarias óptimas [3, Teo. 4.3].

3. Índice en costo promedio

El **costo (esperado) en n-etapas** generado por operar el sistema bajo una política $\pi \in \Pi_m$ dado que el estado inicial es $x_0 = x$ está dado por

$$J_n(\pi, x) := E_x^{\pi} \sum_{k=0}^{n-1} C(x_k, a_k), \ n \in \mathbb{N}. \tag{9}$$

El costo promedio por etapa se define como

$$J(\pi, x) = \lim_{n\to\infty} \sup \frac{1}{n} J_n(\pi, x). \tag{10}$$

A la función

$$J^*(x) := \inf_{\pi\in\Pi} J(\pi, x), \ x \in \mathbf{X}, \tag{11}$$

se le llama **función de valor óptimo en costo promedio**. Diremos que la política $\pi^* \in \Pi$ es **óptima en costo promedio** si

$$J^*(x) = J(\pi^*, x) \quad \forall x \in \mathbf{X}. \tag{12}$$

Como mencionamos previamente, el problema de control óptimo en costo promedio no siempre tiene solución. El siguiente ejemplo, tomado de [39, p. 142], muestra este hecho.

Consideremos el modelo de decisión especificado de la siguiene manera:

- $\mathbf{X} = \{1, -1, 2, -2, 3, -3, \cdots\}$;
- $\mathbf{A} = A(x) = \{1, 2\}$;
- $Q(x + 1|x, 1) = Q(-x|x, 2) = 1$;
- $Q(x|x, 1) = Q(x|x, 2) = 1$ si $x \leq -1$;
- $C(x, 1) = C(x, 2) = \begin{cases} 1 & \text{si } x \geq 1 \\ 1/|x| & \text{si } x \leq -1. \end{cases}$

En palabras: si $x_0 = x \leq -1$, el proceso permanecerá en dicho estado para siempre con un costo $1/|x|$ por cada etapa de decisión. Por otro lado, si $x_0 = x \geq 1$ y se elige el control $a_0 = 1$, el sistema transita al estado $x_1 = x + 1$ con un costo igual a 1; si se elige el control $a_0 = 2$ el sistema se mueve al estado $x_1 = -x$ con un costo $1/|x|$ y permanecerá en dicho estado para siempre con el mismo costo por cada etapa.

Es claro que $J^*(x) = 1/|x|$ si $x_0 = x \leq -1$. Mostraremos que $J^*(x) = 0$ si $x_0 = x \geq 1$. Para hacer esto, suponga que $x_0 = 1$ y considere la políticas estacionarias

$$f^{(k)}(x) := \begin{cases} 1 & \text{si} \quad 1 \leq x < k, \\ 2 & \text{si} \quad x \geq k, \end{cases}$$

definidas para $k \geq 1$. Después de realizar algunos cálculos se obtiene que

$$J_n(f^{(k)}, 1) = \begin{cases} n & \text{si} \quad n < k. \\ k + \dfrac{n-k}{k} & \text{si} \quad n \geq k. \end{cases}$$

Entonces,

$$0 \leq J^*(1) \leq J(f^{(n)}, 1) = \frac{1}{k}.$$

Puesto que n es arbitrario, concluimos que $J^*(1) = 0$. De forma análoga se demuestra que $J^*(x) = 0$ para todo $x \geq 1$.

Ahora considere una política arbitraria $\pi \in \Pi$, y observe que sólo una de las siguientes situaciones puede presentarse:

- π siempre elige la acción de control 1, en cuyo caso $J(\pi, 1) = 1$;
- π elige el control 2 en alguno de los tiempos de decisión; denotemos por $\widehat{n}$ al menor de estos tiempos. Entonces,

$$J(\pi, 1) \geq \frac{1}{\widehat{n}} > 0.$$

De la desigualdad anterior se concluye que **no existen** políticas óptimas en costo promedio para este modelo.

El ejemplo anterior hace evidente que si no se tienen propiedades de «recurrencia» adecuadas no se puede garantizar la existencia de políticas óptimas en costo promedio. Estas propiedades usualmente se garantizan haciendo algunos supuestos sobre el modelo de control y su elección dependen del enfoque o esquema que se use para estudiar el problema de control óptimo en costo promedio. A este se le puede abordar como límite de programas descontados, o plantearse como un problema de programación lineal o como un problema de optimización convexa; también se puede estudiar por medio del algoritmo de iteración de valores o del algoritmo iteración de políticas, o directamente por medio de la ecuación de optimalidad. Estos enfoques, si bien son distintos y en general requieren de supuestos diferentes, también pueden ser complementarios. De hecho, en este trabajo estudiaremos el problema de control en costo promedio por medio de la ecuación de optimalidad y el algoritmo de iteración de políticas.

En la sección 4 veremos que la validez de la ecuación de optimalidad– es decir, la existencia de soluciones a dicha ecuación–garantiza bajo ciertas condiciones la existencia de una política estacionaria óptima. En la

sección 5 estudiaremos el algoritmo de iteración de políticas introducido por Ronald A. Howard. Este algoritmo calcula en cada etapa de ejecución una política estacionaria con un costo promedio igual o menor que el costo generado por la política calculada en el paso anterior; entonces, si se excluyeran la aparición de ciclos, podría esperarse que dicho algoritmo finalice identificando a una política estacionaria con el menor costo posible entre todas las políticas estacionarias. En la sección 5, bajo el supuesto de que se satisface una condición de «irreducibilidad» y considerando modelos con espacios de estado y controles finitos, mostraremos que esto ocurre en un número finito de pasos y que además se cumple la ecuación de optimalidad, lo cual a su vez garantiza que la política identificada por el algoritmo es óptima.

4. Ecuación de optimalidad en costo promedio

La **ecuación de optimalidad** o **ecuación de Bellman** (en costo promedio) correspondiente el modelo de decisión $\mathcal{M} =(\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C)$ es la ecuación funcional

$$\rho^* + h^*(x) = \inf_{a \in A(x)} [C(x,a) + \sum_{y \in \mathbf{X}} h^*(y) Q(y|x,a)] \quad \forall x \in \mathbf{X}, \tag{13}$$

donde la «incógnita» es el par formado por el escalar $\rho^* \in \mathbb{R}$ y la función $h^* : \mathbf{X} \to \mathbb{R}$. Si este par existe, diremos que la ecuación se cumple o que es válida, y diremos que (ρ^*, h^*) es una solución de la ecuación de optimalidad. Si adicionalmente existe un selector $f^* \in \mathbb{F}$ tal que $f^*(x)$ alcanza el ínfimo en (13) para cada $x \in \mathbf{X}$, es decir,

$$\rho^* + h^*(x) = C(x, f^*(x)) + \sum_{y \in \mathbf{X}} h^*(y) Q(y|x, f^*(x)) \quad \forall x \in \mathbf{X},$$

diremos que (ρ^*, h^*, f^*) es una **terna canónica.** Observe que si usamos la notación (3)-(4), la igualdad anterior se expresa como

$$\rho^* + h^* = C_{f^*} + Q_{f^*} h^*.$$

El siguiente teorema muestra la importancia de la ecuación de optimalidad para la solución del problema de control en costo promedio.

Teorema 4.1. *Si existe una solución (ρ^*, h^*) de la ecuación de optimalidad tal que*

$$\lim_{n \to \infty} \frac{1}{n} E_x^\pi h^*(x_n) = 0 \quad \forall x \in \mathbf{X}, \pi \in \Pi, \tag{14}$$

entonces

$$J^*(x) \geq \rho^* \quad \forall x \in \mathbf{X}.$$

Si adicionalmente existe f^* tal que (ρ^*, h^*, f^*) es una terna canónica, entonces ρ^* es el costo promedio óptimo y f^* es una política estacionaria óptima, es decir,

$$J^*(x) = J(f^*, x) = \rho^* \quad \forall x \in \mathbf{X}.$$

Demostración. Fijemos $x \in \mathbf{X}$ y $\pi = \{f_n\} \in \Pi_m$. Puesto que (ρ^*, h^*) es una solución de la ecuación de optimalidad, tenemos que

$$\rho^* + h^*(x) \leq C(x, a) + \sum_{y \in \mathbf{X}} h^*(y) Q(y|x, a) \quad \forall (x, a) \in \mathbb{K}.$$

Entonces, usando la propiedad (2), resulta que

$$\rho^* + h^*(x_k) \leq C(x_k, a_k) + E_x^\pi[h^*(x_{k+1})|x_k] \quad \forall x \in \mathbf{X}, k \in \mathbb{N}_0,$$

de donde se sigue la desigualdad

$$n\rho^* \leq \sum_{k=0}^{n-1} C(x_k, a_k) + \sum_{k=0}^{n-1} [E_x^\pi[h^*(x_{k+1})|x_k] - h^*(x_k)].$$

Ahora observe que

$$\begin{aligned} E_x^\pi \sum_{k=0}^{n-1} [E_x^\pi[h^*(x_{k+1})|x_k] - h^*(x_k)] &= \sum_{k=0}^{n-1} [E_x^\pi h^*(x_{k+1}) - E_x^\pi h^*(x_k)] \\ &= E_x^\pi h^*(x_n) - E_x^\pi h^*(x_0) \\ &= E_x^\pi h^*(x_n) - h^*(x), \end{aligned}$$

lo cual combinado con la desigualdad anterior implica que

$$n\rho^* \leq E_x^\pi \sum_{k=0}^{n-1} C(x_k, a_k) + E_x^\pi h^*(x_n) - h^*(x) \quad \forall n \in \mathbb{N}.$$

Entonces, dividiendo ambos lados en la desigualdad anterior por n y tomando límite cuando n tiende a infinito, de la condición (14) se obtiene

$$J(\pi, x) \geq \rho^*.$$

Puesto que $\pi \in \Pi$ y $x \in \mathbf{X}$ son arbitrarios, se concluye que

$$J(x) \geq \rho^* \quad \forall x \in \mathbf{X}. \qquad \square$$

5. Distribuciones invariantes y ecuación de Poisson

Para presentar el algoritmo de iteración de políticas requerimos de notación adicional, así como de algunos conceptos y resultados de la teoría de cadenas de Markov.

Recuerde que para cada política $f \in \mathbb{F}$ el proceso de estados $\{x_n\}$ es una cadena de Markov homogénea con espacio de estados $\mathbf{X}$ y probabilidades de transición Q_f. Denotatemos por $Q_f^{(n)}$ a la matriz de probabilidades de transición en n pasos, es decir, a la matriz con coeficientes

$$Q_f^{(n)}(y|x) := P_x^f[x_n = y], \quad x, y \in \mathbf{X}.$$

Es bien sabido que las que las matrices de probabilidades de transición satisfacen las ecuaciones de Chapman-Kolmogorov:

$$\begin{aligned} Q_f^{(n+m)}(y|x) &= \sum_{z\in\mathbf{X}} Q^{(n)}(y|z)Q^{(m)}(z|x) \\ &= \sum_{z\in\mathbf{X}} Q^{(m)}(y|z)Q^{(n)}(z|x), \end{aligned}$$

para todo $n, m \in \mathbb{N}_0$, donde $Q_f^{(0)}$ denota la matriz identidad. Puesto que $Q_f^{(1)} = Q_f$, estas ecuaciones implican que

$$Q_f^{(n)} = Q_f^n \ \ \forall n \in \mathbb{N},$$

donde Q_f^n denota la n-ésima potencia de la matriz Q_f.

Para cada función $v : \mathbf{X} \to \mathbb{R}$ defina

$$Q_f^n v(x) := \sum_{y\in\mathbf{X}} v(y)Q_f^n(y|x), \ x \in \mathbf{X}, \, n \in \mathbb{N},$$

siempre y cuando la serie sea convergente. Observe que

$$Q_f^n v(x) = E_x^f v(x_n) \ \forall n \in \mathbb{N}_0.$$

Además, para cada distribución de probabilidad μ sobre $\mathbf{X}$ y cada función $v : \mathbf{X} \to \mathbb{R}$ defina

$$\mu(v) := \sum_{y\in\mathbf{X}} \mu(y)v(y),$$

siempre y cuando la serie sea convergente. Recuerde que μ es una distribución de probabilidad sobre $\mathbf{X}$ si $\mu(x) \geq 0$ para todo $x \in \mathbf{X}$ y se cumple que $\sum_{x\in\mathbf{X}} \mu(x) = 1$. Note que $Q_f v$ es una función de $\mathbf{X}$ en $\mathbb{R}$ y que $\mu(v)$ es una constante.

Uno de los temas centrales de la teoría de cadenas de Markov es el relativo al comportamiento asintótico (a largo plazo) de las cadenas [23]. Dicho comportamiento esta ligado a la existencia de distribuciones de probabilidad invariantes. Diremos que una distribución de probabilidad μ sobre $\mathbf{X}$ es una **distribución de probabilidad invariante** (o de equilibrio) para la matriz de transición Q_f si satisface la ecuación

$$\mu(y) = \sum_{x\in\mathbf{X}} Q_f(y|x)\mu(x) \ \ \forall x \in \mathbf{X}. \tag{15}$$

Puede mostrarse directamente que μ es una distribución de probabilidad invariante para Q_f si y sólo si

$$\mu(v) = \sum_{x\in\mathbf{X}} Q_f^n v(x)\mu(x) \quad \forall n \in \mathbb{N}, \tag{16}$$

para toda función acotada $v : \mathbf{X} \to \mathbb{R}$.

El siguiente teorema reune algunos resultados de la teoría de cadenas de Markov con espacios finitos. La demostración de estos resultados puede consultarse en [45, Teo. 3.3.2, 3.3.3] y [3, Teo. 4.1]. El concepto central en este teorema es el de irreducibilidad: a la matriz de probabilidades de transición Q_f se le llama **irreducible** si para cada par de estados $x, y \in \mathbf{X}$ existe un número natural $m = m_{x,y}$ tal que

$$Q_f^m(y|x) > 0.$$

Teorema 5.1. *Suponga que el espacio de estados es finito. Si la matriz Q_f correspondiente al selector $f \in \mathbb{F}$ es irreducible, entonces:*

(a) *la matriz Q_f tiene una única distribución de probabilidad invariante μ_f; además, $\mu_f(x) > 0$ para todo $x \in \mathbf{X}$;*

(b) *para cada función $v : \mathbf{X} \to\mathbb{R}$ se cumple que*

$$\begin{aligned}\mu_f(v) &= \lim_{n\to\infty} \frac{1}{n} J_n(f, x) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Q_f^n v(x) \quad \forall x \in \mathbf{X},\end{aligned}$$

En particular, se cumple que

$$J(f, x) = \mu_f(C_f) \quad \forall x \in \mathbf{X}.$$

(c) *Sea $z \in \mathbf{X}$ un estado fijo arbitrario. Existe una única función $h_f : \mathbf{X} \to\mathbb{R}$ tal que $h_f(z) = 0$ y una constante ρ_f que satisfacen la* ***ecuación de Poisson***

$$\rho_f + h_f = C_f + Q_f h_f. \tag{17}$$

En la siguiente observación se muestra que la ecuación de Poisson proporciona información más precisa que el teorema 5.1(b) sobre el comportamiento asintótico de las cadenas de Markov.

Observación 5.1. *Suponga que se satisfacen las condiciones del teorema 5.1.*

(a) Iterando la ecuación de Poisson (17) se obtiene las igualdades

$$\begin{aligned}n\rho + h_f(x) &= \sum_{k=0}^{n-1} Q^k C_f(x) + Q_f^{n+1} h_f(x) \\ &= +E_x^f h_f(x_n),\end{aligned}$$

para todo $x \in \mathbf{X}$ *y* $n \in \mathbb{N}_0$. *Reordenando términos, se obtiene*

$$J(f,x) = n\rho + h_f(x) - E_x^f h_f(x_n).$$

Entonces, dividiendo por n *y tomando límite resulta que*

$$\rho_f = \lim_{n\to\infty} \frac{1}{n} J(f,x) = J(f,x) = \mu_f(C_f) \quad \forall x \in \mathbf{X},$$

puesto que la función h_f *es acotada.*

(b) Para tener una idea más concreta de la naturaleza de la función h_f *supongamos adicionalmente que la matriz* Q_f *es aperiódica. En este caso puede mostrarse que la función*

$$h'_f(x) := \sum_{n=0}^{\infty} E_x^f[C_f(x_n) - \rho_f], \quad x \in \mathbf{X},$$

es la única solución de la ecuación de Poisson que satisface la condición $\mu_f(h_f) = 0$. *Entonces, puede verificarse directamente que*

$$h_f(x) = h'_f(x) - h'_f(z) \quad \forall x \in \mathbf{X}.$$

Observación 5.2. *En general, el cálculo explícito del costo promedio* $J(f,x)$ *usando directamente su definición es prácticamente imposible. No obstante, si el espacio de estados es finito y la matriz* Q_f *es irreducible, el teorema 5.1(a)-(b) ofrece una alternativa numéricamente fáctible si la cardinalidad del espacio de estados es «moderada». Observe que si* $\mathbf{X} = \{1, 2, \ldots, s\}$, *la condición (15) que define a una distribución de probabilidad invariante es un sistema de* s *ecuaciones lineales con* s *incógnitas, a saber,* $\mu(1), \mu(2), \ldots, \mu(s)$. *Si sustituimos una de estas ecuaciones con la ecuación*

$$\mu(1) + \cdots + \mu(s) = 1,$$

se obtiene un sistema no-homogéneo que puede resolverse con el método de eliminación de Gauss-Jordan. Las referencia [45] ofrece una presentación introductoria a los métodos númericos para el computo de distribuciones de probabilidad invariantes y la referenica [43] da un tratamiento exhaustivo y matemáticamente más avanzado que la anterior.

Así, en el supuesto de que $\mathbf{X}$ *es finito y la matriz* Q_f *es irreducible, para calcular el costo promedio*

$$J(f,x) = \lim_{n\to\infty} \frac{1}{n} J(f,x),$$

podemos proceder a calcular primero la distribución de probabilidad invariante μ_f *resolviendo un sistema lineal y luego calcular la cantidad*

$$\mu_f(C_f) = \sum_{k=1}^{s} \mu(k) C_f(k),$$

que por el teorema 5.1(b), es igual a $J(f,x)$ *para todo* $x \in \mathbf{X}$.

Observación 5.3. *Una segunda alternativa para el cálculo de $J(f,x)$ la da el teorema 5.1(b)-(c) a través de la ecuación de Poisson tomando, por ejemplo, $h_f(1) = 0$. Esto da origen al siguiente sistema de s ecuaciones lineales con s incógnitas:*

$$\rho_f + h(i) = C_f(i) + \sum_{j=1}^{s} h_f(j)Q(j|i), \quad j = 1, \ldots, s,$$

donde las incógnitas son las cantidades $h_f(2), \ldots, h_f(s)$ y ρ_f.

La ecuación de Poisson se estudia ampliamente en las referencias [23] y [31]; la segunda estudia cadenas con espacio de estados numerables y la primera con cadenas con espacios no-numerables.

6. Algoritmo de iteración de políticas

Supongamos que tanto el espacio de estados como el de controles son finitos; note que en este caso $\mathbb{F}$ tiene una cantidad finita de elementos. Si su cardinalidad es mayor que uno, el conjunto de políticas markovianas Π_M es no-numerable. Este hecho, nos advierte que la busqueda «directa» de una política óptima será con toda seguridad infructuosa. Ahora, podría parecernos natural que restrinjamos la búsqueda a la clase de las políticas estacionarias $\mathbb{F}$ puesto que esta es finita. De hecho, esto último garantiza que existe al menos una política estacionaria f^* que es óptima entre las políticas estacionarias, es decir,

$$\rho_{f^*} = \inf_{f \in \mathbb{F}} \rho_f.$$

Para encontrár dicha política podría intentarse una búsqueda exhaustiva en $\mathbb{F}$ basada en los algoritmos presentados en las observaciones 5.2 y 5.3; sin embargo, como pasa en la mayoría de los problemas de optimización, esta búsqueda no es factible numéricamente si $\mathbb{F}$ no es de cardinalidad pequeña. El algoritmo de iteración de políticas resuelve este problema de forma elegante y numéricamente eficiente. De hecho, para modelos finitos, este algoritmo calcula en un cantidad finita de pasos una política f^* que es óptima en la clase de todas las políticas, es decir,

$$J(\pi, x) \geq J(f^*, x) = \rho_{f^*} \quad \forall x \in \mathbf{X}, \pi \in \Pi_m.$$

El siguiente resultado es esencial para el algoritmo de iteración de políticas.

Lema 6.1. *Suponga que el espacio de estados $\mathbf{X}$ es finito y que Q_f es irreducible. Si existe una constante ρ y una función $h : \mathbf{X} \to \mathbb{R}$ tal que*

$$\rho + h \geq C_f + Q_f h, \tag{18}$$

entonces $\rho \geq \rho_f = J(f,x)$ *para todo* $x \in \mathbf{X}$. *Si la desigualdad es estricta para algún estado, entonces* $\rho > \rho_f$.

Demostración. Este resultado se puede probar iterando la desigualdad de manera análoga a lo hecho en la observación 5.1. Una demostración alternativa se obtiene usando la distribución invariante μ_f de Q_f de la siguiente manera. Primero observe que de la hipótesis se tiene la desigualdad

$$\sum_{k=1}^{s}[\rho + h(k)]\mu_f(k) \geq \sum_{k=1}^{s}[C_f(k) + Q_f h(k)]\mu_f(k),$$

la cual combinada con (16) implica que

$$\begin{aligned}\rho + \mu_f(h) &\geq \mu_f(C_f) + \sum_{k=1}^{s} Q_f h(k)\mu_f(k) \\ &\geq \rho_f + \mu_f(h),\end{aligned}$$

de donde se concluye que $\rho \geq \rho_f$.

Si la desigualdad (18) es estricta para algún estado, entonces

$$\rho + \mu_f(h) > \rho_f + \mu_f(h),$$

puesto que $\mu_f(i) > 0$ para todo $i \in \mathbf{X}$ (teorema 5.1(a)). Por lo tanto, $\rho > \rho_f$. □

A continuación se describe el algoritmo de iteración de políticas. En el teorema 6.1 se prueba que dicho algortimo identifica una política estacionaria óptima en un número finito de pasos.

Algoritmo de iteración de políticas.:

Paso 0 (inicio): tome $n = 0$, fije un estado $z \in X$ y elija una política estacionaria $f_n \in \mathbb{F}$.

Paso 1 (evaluación): encuentre la solución $\rho_n := \rho_{f_n}$ y $h_n := h_{f_n}$ de la ecuación de Poisson

$$\begin{aligned}\rho_n + h_n &= C_{f_n} + Q_{f_n} h_n, \\ h_n(z) &= 0.\end{aligned}$$

Paso 2 (mejoramiento): encuentre un selector $f_{n+1} \in \mathbb{F}$ tal que

$$C_{f_{n+1}}(i) + Q_{f_{n+1}} h_n(i) = \min_{a \in A(i)}[C(i,a) + \sum_{k=1}^{s} h_n(k)Q(k|i,a)] \quad \forall i \in \mathbf{X}.$$

(a): Si $f_{n+1}(i) \neq f_n(i)$ para algun $i \in \mathbf{X}$ vaya al paso 1 y repita el procedimiento.

(b): Si $f_{n+1}(i) = f_n(i)$ para todo $i \in \mathbf{X}$, pare el algoritmo con la política f_n.

Teorema 6.1. *Suponga que los espacios de estados y controles son finitos y que las matrices de transición $Q_f, f \in \mathbb{F}$, son irreducibles. Entonces:*

(a) *el algoritmo de iteración de políticas para en un número finito de pasos;*

(b) *si el algoritmo se detiene en la n-ésima etapa, entonces la constante $\rho^* := \rho_n = \rho_{f_n}$, la función $h^* := h_n = h_{f_n}$ y la política $f^* = f_n$ forman una terna canónica, es decir, satisfacen las ecuaciones*

$$\rho^* + h^*(i) = \min_{a \in A(i)} [C(i,a) + \sum_{k=1}^{s} h^*(k) Q(k|i,a)], \tag{19}$$

$$= C_{f^*}(i) + Q_{f^*} h^*(i), \tag{20}$$

para todo $i \in \mathbf{X}$.

(c) *la política f^* es óptima y ρ^* es el costo promedio óptimo, es decir,*

$$\rho^* = \rho_{f^*} = J(f^*, i) \leq J(\pi, i) \quad \forall i \in \mathbf{X}, \pi \in \Pi_m.$$

Demostración. Supongamos que el algoritmo no para, es decir, que para cada $n \in \mathbb{N}$ existe un estado $i_n \in \mathbf{X}$ tal que $f_n(i_n) \neq f_{n+1}(i_n)$. Entonces,

$$\begin{aligned} \rho_n + h_n(i) &= C_{f_n}(i) + Q_{f_n} h_n(i) \\ &\geq C_{f_{n+1}}(i) + Q_{f_{n+1}} h_n(i) \end{aligned}$$

para todo $i \in \mathbf{X}$, con desigualdad estricta para $i = i_n$. Por el lema 6.1 se tiene que

$$\rho_0 > \rho_1 > \cdots > \rho_n > \cdots > \inf_{f \in \mathbb{F}} \rho_f > -\infty,$$

de donde se sigue que la sucesión $\{\rho_n\}$ es convergente. Por otra parte, puesto que el conjunto

$$\{\rho_f : f \in \mathbb{F}\},$$

es finito, la convergencia de la sucesión implica que existe $m \in \mathbb{N}$ ta que $\rho_k = \rho_{k+1}$ para todo $k \geq m$, contradiciendo que la sucesión $\{\rho_n\}$ es estrictamente decreciente. Por lo tanto, el algoritmo se detiene en un número finito de pasos.

(b) Suponga que el algoritmo para en el n-ésimo paso, es decir, $f_{n+1} = f_n$. Entonces,

$$\begin{aligned} \rho_n + h_n(i) &= C_{f_n}(i) + Q_{f_n} h_n(i) \\ &= C_{f_{n+1}}(i) + Q_{f_{n+1}} h_n(i) \\ &= \min_{a \in A(i)} [C(i,a) + \sum_{k=1}^{s} h_n(k) Q(k|i,a)], \end{aligned}$$

para todo $i \in \mathbf{X}$, lo cual prueba el resultado deseado.

(c) Esta afirmación se sigue de la parte (b) de este teorema y del teorema 4.1 puesto que h^* es una función acotada. □

7. Comentarios finales

En este trabajo se presentaron dos de las principales aportaciones en el estudio del problema de control óptimo en costo promedio: la ecuación de optimalidad y el algoritmo de iteración de políticas (o algoritmo de Howard). Este algoritmo proporciona un procedimiento para encontrar una política estacionaria óptima–en principio, en la clase de las políticas estacionarias–resolviendo ciertos sistemas de ecuaciones lineales. Desde su introducción, este algoritmo se ha estudiado ampliamente y se han logrado aportaciones importantes; consulte [38] para una discusión detallada y exhaustiva al respecto. Sin embargo, la mayor parte de los trabajos se restringen a modelos con espacios de estados discretos (es decir, finitos o infinitos numerables). Para modelos con espacios de estados no-numerables los únicos trabajos que estudian este algoritmo, según nuestro conocimiento, son las referencias [11, 21, 33], y los resultados que en ellos se presentan no son completamente satisfactorios.

Por otra parte, tenemos a la ecuación funcional (19) conocida como ecuación de optimalidad (o ecuación de Bellman) en costo promedio. En la referencia [3] se da el crédito a C. Derman (1966) del reconocimiento de la importancia de esta ecuación para probar la existencia de políticas estacionarias óptimas; por otra parte, Puterman [38, p. 429] señala que esta ecuación ya había aparecido, implícitamente o en casos particulares, en trabajos previos de D. Blackwell (1962), D. Iglehart (1963) y H. M. Taylor (1965). No obstante, en 1957, Bellman [9] ya había estudiado el problema de control en costo promedio y mostrado la relación de la ecuación de optimalidad con la existencia de políticas estacionarias óptimas.

La ecuación de optimalidad también es importante porque hace posible el estudio de otros índices de funcionamiento complementarios al índice en costo promedio; entre tales índices se encuentran los criterios sensibles al horizonte de planeacion [24], el índice en costo promedio por trayectorias —es decir, sin tomar valor esperado en (8)— y de minimización de varianza [25].

Además del algoritmo de iteración de políticas existen otros enfoques o algoritmos para obtener soluciones de la ecuación de optimalidad en costo promedio como el algoritmo de iteración de valores [18], el enfoque de programación lineal [22], el enfoque de punto fijo [48] y el enfoque de aproximaciones por problemas descontados [49].

Los procesos de decisión markovianos es una rama matemática consolidada con un cuerpo teórico bien establecido [6, 15, 19, 20, 22]. Sin embargo, con su historia de sesenta años, es actualmente un campo de investigación muy activo debido a la gran variedad de problemas matemáticos que plantea y a la diversidad de sus aplicaciones reales y potenciales.

Las referencias [10, 37, 39, 45] ofrecen introducciones elementales a los procesos de decisión markovianos apropiadas para estudiantes de licenciatura, mientras que las monografías [6, 19, 20, 22] dan un tratamiento avanzado dirigido a estudiantes de posgrado e investigadores.

Bibliografía

[1] E. Altman, «Applications of Markov decision processes in communication networks», en *Handbook of Markov decision processes*, Springer, 2002, 489–536.

[2] R. A. Androkovich y K. R. Stollery, «A stochastic dynamic programming model of bycatch control in fisheries», *Marine Resource Economics*, vol. 9, núm. 1, 1994, 19–30.

[3] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh y S. I. Marcus, «Discrete-time controlled Markov processes with average cost criterion: a survey», *SIAM Journal on Control and Optimization*, vol. 31, núm. 2, 1993, 282–344.

[4] A. A. Assad, «Richard E. Bellman», en *Profiles in Operations Research*, Springer, 2011, 415–445.

[5] Y. Bai, P. Zhou, D. Zhou, F. Meng y K. Ju, «Desirable policies of a strategic petroleum reserve in coping with disruption risk: A Markov decision process approach», *Computers & Operations Research*, vol. 66, 2016, 58–66.

[6] N. Bäuerle y U. Rieder, *Markov decision processes with applications to finance*, Springer,, 2011.

[7] R. Bellman y E. Lee, «History and development of dynamic programming», *IEEE Control Systems Magazine*, vol. 4, núm. 4, 1984, 24–28.

[8] R. Bellman, «On the theory of dynamic programming», *Proceedings of the National Academy of Sciences*, vol. 38, núm. 8, 1952, 716–719.

[9] ———, «A markovian decision process», *Journal of Mathematics and Mechanics*, 1957, 679–684.

[10] D. P. Bertsekas, *Dynamic programming: Deterministic and stochastic models*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[11] O. L. d. V. Costa y F. Dufour, «Average control of Markov decision processes with Feller transition probabilities and general action spaces», *Journal of Mathematical Analysis and Applications*, vol. 396, núm. 1, 2012, 58–69.

[12] S. Dreyfus, «Richard bellman on the birth of dynamic programming», *Operations Research*, vol. 50, núm. 1, 2002, 48–51.

[13] ———, «Richard ernest bellman», *International Transactions in Operational Research*, vol. 10, núm. 5, 2003, 543–545.

[14] M. Fakoor, A. Kosari y M. Jafarzadeh, «Humanoid robot path planning with fuzzy Markov decision processes», *Journal of Applied Research and Technology*, vol. 14, núm. 5, 2016, 300–310.

[15] E. A. Feinberg, «Optimality conditions for inventory control», *Tutorials in operations research. Optimization challenges in complex, networked, and risky systems, INFORMS, Cantonsville, MD*, 2016, 14–44.

[16] K. Gatsis, A. Ribeiro y G. J. Pappas, «Optimal power management in wireless control systems», *IEEE Transactions on Automatic Control*, vol. 59, núm. 6, 2014, 1495–1510.

[17] Y. Gocgun, B. W. Bresnahan, A. Ghate y M. L. Gunn, «A Markov decision process approach to multi-category patient scheduling in a diagnostic facility», *Artificial intelligence in medicine*, vol. 53, núm. 2, 2011, 73–81.
[18] E. Gordienko y O. Hernández-Lerma, «Average cost Markov control processes with weighted norms: value iteration», *Applicationes Mathematicae*, vol. 23, núm. 2, 1995, 219–237.
[19] O. Hernandez-Lerma, «Adaptive Control of Markov Processes», 1989.
[20] O. Hernández-Lerma y J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30, Springer Science & Business Media, 1996.
[21] ———, «Policy iteration for average cost Markov control processes on Borel spaces», *Acta Applicandae Mathematica*, vol. 47, núm. 2, 1997, 125–154.
[22] ———, *Further topics on discrete-time Markov control processes*, vol. 42, Springer Science & Business Media, 1999.
[23] ———, *Markov chains and invariant probabilities*, vol. 211, Birkhäuser, 2003.
[24] O. Hernández-Lerma y O. Vega-Amaya, «Infinite-horizon Markov control processes with undiscounted cost criteria: from average to overtaking optimality», *Applicationes Mathematicae*, vol. 25, núm. 2, 1998, 153–178.
[25] O. Hernández-Lerma, O. Vega-Amaya y G. Carrasco, «Sample-path optimality and variance-minimization of average cost Markov control processes», *SIAM Journal on Control and Optimization*, vol. 38, núm. 1, 1999, 79–93.
[26] C.-I. Hsu, H.-C. Li, S.-M. Liu y C.-C. Chao, «Aircraft replacement scheduling: a dynamic programming approach», *Transportation research part E: logistics and transportation review*, vol. 47, núm. 1, 2011, 41–60.
[27] A. Jaśkiewicz y A. S. Nowak, «Discounted dynamic programming with unbounded returns: application to economic models», *Journal of Mathematical Analysis and Applications*, vol. 378, núm. 2, 2011, 450–462.
[28] B. F. Lamond y A. Boukhtouta, «Water reservoir applications of Markov decision processes», en *Handbook of Markov decision processes*, Springer, 2002, 537–558.
[29] A. Lew, «Richard Bellman's contributions to computer science», *Journal of Mathematical Analysis and Applications*, vol. 119, núm. 1-2, 1986, 90–96.
[30] P. Magni, S. Quaglini, M. Marchetti y G. Barosi, «Deciding when to intervene: a Markov decision process approach», *International Journal of Medical Informatics*, vol. 60, núm. 3, 2000, 237–253.
[31] A. M. Makowski y A. Shwartz, «The Poisson equation for countable Markov chains: probabilistic methods and interpretations», en *Handbook of Markov decision processes*, Springer, 2002, 269–303.
[32] A. R. Mesquita, J. P. Hespanha y G. N. Nair, «Redundant data transmission in control/estimation over lossy networks», *Automatica*, vol. 48, núm. 8, 2012, 1612–1620.
[33] S. P. Meyn, «The policy iteration algorithm for average reward Markov decision processes with general state space», *IEEE Transactions on Automatic Control*, vol. 42, núm. 12, 1997, 1663–1680.
[34] K. S. Moghaddam y J. S. Usher, «Preventive maintenance and replacement scheduling for repairable and maintainable systems using dynamic programming», *Computers & Industrial Engineering*, vol. 60, núm. 4, 2011, 654–665.
[35] L. G. N. Nunes, S. V. de Carvalho y R. d. C. M. Rodrigues, «Markov decision process applied to the control of hospital elective admissions», *Artificial intelligence in medicine*, vol. 47, núm. 2, 2009, 159–171.
[36] A. Piunovskiy, *Examples in Markov decision processes*, vol. 2, World Scientific, 2013.
[37] W. B. Powell, *Approximate dynamic programming: Solving the curses of dimensionality*, vol. 703, John Wiley & Sons, 2007.
[38] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 1994.

[39] S. M. Ross, *Applied probability models with optimization applications*, Dover Publications Inc., 1993.
[40] M. Schäl, «Markov decision processes in finance and dynamic options», en *Handbook of Markov decision processes*, Springer, 2002, 461–487.
[41] H. P. Simao, J. Day, A. P. George, T. Gifford, J. Nienow y W. B. Powell, «An approximate dynamic programming algorithm for large-scale fleet management: A case application», *Transportation Science*, vol. 43, núm. 2, 2009, 178–197.
[42] G. Stensland y D. Tjøstheim, «Optimal investments using empirical dynamic programming with application to natural resources», *Journal of Business*, 1989, 99–120.
[43] W. J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.
[44] N. L. Stokey, *Recursive methods in economic dynamics*, Harvard University Press, 1989.
[45] H. C. Tijms, *A first course in stochastic models*, John Wiley and sons, 2003.
[46] D. Van Dijk, R. Haijema, E. M. Hendrix, R. A. Groeneveld y E. C. van Ierland, «Fluctuating quota and management costs under multiannual adjustment of fish quota», *Ecological modelling*, vol. 265, 2013, 230–238.
[47] D. van Dijk, E. M. Hendrix, R. Haijema, R. A. Groeneveld y E. C. van Ierland, «On solving a bi-level stochastic dynamic programming model for analyzing fisheries policies: Fishermen behavior and optimal fish quota», *Ecological modelling*, vol. 272, 2014, 68–75.
[48] O. Vega-Amaya, «The average cost optimality equation: a fixed point approach», *Bol. Soc. Mat. Mexicana*, vol. 9, núm. 1, 2003, 185–195.
[49] Ó. Vega-Amaya, «On the vanishing discount factor approach for Markov decision processes with weakly continuous transition probabilities», *Journal of Mathematical Analysis and Applications*, vol. 426, núm. 2, 2015, 978–985.
[50] D. J. White, «Real applications of Markov decision processes», *Interfaces*, vol. 15, núm. 6, 1985, 73–83.
[51] ———, «Further real applications of Markov decision processes», *Interfaces*, vol. 18, núm. 5, 1988, 55–61.
[52] ———, «A survey of applications of Markov decision processes», *Journal of the Operational Research Society*, vol. 44, núm. 11, 1993, 1073–1096.
[53] G. Wu, Y. Fan, L.-C. Liu y Y.-M. Wei, «An empirical analysis of the dynamic programming model of stockpile acquisition strategies for china's strategic petroleum reserve», *Energy Policy*, vol. 36, núm. 4, 2008, 1470–1478.
[54] S. Yakowitz, «Dynamic programming applications in water resources», *Water resources research*, vol. 18, núm. 4, 1982, 673–696.
[55] W. W.-G. Yeh, «Reservoir management and operations models: A state-of-the-art review», *Water resources research*, vol. 21, núm. 12, 1985, 1797–1818.
[56] H. Zhang y W. X. Zheng, «Sensor power scheduling: Tradeoff between state estimation quality and power cost», en *Automation (YAC), 2017 32nd Youth Academic Annual Conference of Chinese Association of*, IEEE, 2017, 922–927.