Miscelánea Matemática **66** (2018) 5-21 SMM

Estimación empírica en sistemas de control de Markov

J. Adolfo Minjárez-Sosa
Departamento de Matemáticas
Universidad de Sonora
aminjare@gauss.mat.uson.mx

A la memoria de nuestro colega y amigo Juan González-Hernández

Resumen

Consideramos una clase de sistemas de control estocástico que evoluciona a tiempo discreto de acuerdo a una ecuación en diferencia de la forma $x_{t+1} = F(x_t, a_t, \xi_t)$, $t = 0, 1, \ldots$, donde x_t y a_t representan el estado y el control aplicado al tiempo t, y $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con distribución desconocida. En este artículo se presentan algunas ideas del método de estimación y control para aproximar el costo óptimo y la política óptima bajo el criterio de optimalidad de costo descontado.

1. Introducción

La Teoría de Control Óptimo trata con problemas de optimización de sistemas que evolucionan en el tiempo. Sus aplicaciones se presentan en situaciones donde es necesario controlar o manipular el comportamiento de sistemas dinámicos que aparecen en muchas áreas como por ejemplo, Economía, Finanzas, Biología e Ingeniería; y el objetivo es determinar las acciones o decisiones de control que debe tomar un controlador durante la evolución del sistema para obtener un menor costo o mayor ganancia en su operación. Estas acciones se determinan por medio de sucesiones de reglas de decisión o funciones a las que se le conocen como *políticas de control*, o simplemente políticas. El comportamiento de las políticas lo mide un funcional al cual se le conoce como Índice de Funcionamiento, y mediante este índice, el controlador puede conocer

qué política proporciona una mejor respuesta (menor costo o mayor ganancia) que otra. Al problema de buscar una política que minimice o maximice el índice de funcionamiento se le llama *Problema de Control Óptimo*. En este artículo nos centraremos en analizar el problema de minimizar un funcional de costo.

En muchos de los campos de aplicación surge la necesidad de considerar elementos aleatorios que representen alguna incertidumbre que influye en el comportamiento del sistema dinámico, si este es el caso diremos que tenemos un *Problema de Control Óptimo Estocástico*. Al conjunto de componentes que determinan la evolución del sistema dinámico y que definen el problema de control óptimo le llamaremos *Modelo de Control*.

La teoría de control tuvo sus origenes en el cálculo de variaciones, y fue hasta los años 50's cuando se le dió gran impulso al desarrollarse diferentes técnicas para resolver el problema de control. Una de ellas fué la Programación Dinámica propuesta por Bellman en el año de 1951 (véase, e.g., [2]), la cual toma gran importancia ya que se puede extender directamente al caso estocástico. A partir de ese momento el interés por el estudio de la teoría de control óptimo estocástico creció vertiginosamente, lo cual permitió explorar nuevas teorías y variantes del problema de control original que fueran más realistas o que se apegaran más a las condiciones de los problemas que surgían en las áreas de aplicación. Uno de ellos es el que a continuación presentamos.

Generalmente, en los problemas de aplicación, algunas de las componentes del modelo de control no son completamente conocidas por el controlador. Esto nos lleva a implementar esquemas que nos permitan ir aprendiendo o recolectando información acerca de las componentes desconocidas durante la evolución del sistema, y de esta manera poder elegir una decisión o un control con la mayor información posible. Si lo anterior es posible de realizar, decimos que tenemos un problema de control estocástico adaptado, para el cual debemos de diseñar políticas de control que minimicen un índice de funcionamiento.

Una clase de sistemas de control estocástico que nos puede llevar a un problema de control adaptado la constituyen aquellos sistemas cuya evolución a tiempo discreto se determina por medio de una ecuación en diferencia estocástica de la forma

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0 := \{0, 1, \dots\}, \tag{1}$$

donde F es una función conocida, x_t y a_t representan el estado del sistema y el control (o acción) elegido por el controlador al tiempo t, respectivamente. Además, $\{\xi_t\}$ es una sucesión observable de variables aleatorias independientes e identicamente distribuidas (i.i.d.) con distribución común θ.

Es claro que la evolución del sistema es aleatoria y el comportamiento probabilístico lo determina la distribución θ. Es decir, la distribución θ es fundamental para estudiar la dinámica del sistema (1). Sin embargo, en muchos problemas, por ejemplo en finanzas donde ξ_t representa la tasa de interés o en sistemas de inventarios donde ξ_t representa la demanda de cierto artículo, el suponer que θ es conocida podría resultar poco realista. Entonces, el problema que se nos presenta al suponer que θ es desconocida por el controlador lo podemos plantear como un problema de control adaptado. En efecto, como las variables aleatorias $\{\xi_t\}$ son observables, durante la evolución del sistema podemos ir recolectando información de tal forma que en cada etapa t sea posible construir un estimador $\theta_t = \theta_t(\xi_0, \xi_1, \ldots, \xi_{t-1})$ de θ. Bajo este escenario, las acciones se seleccionarán tomando θ_t como la distribución verdadera. Por lo tanto, entre más observaciones del proceso $\{\xi_t\}$ se tengan, se contará con mejor información acerca de la distribución desconocida θ. A este procedimiento de controlar el sistema se le conoce como *Proceso de Estimación y Control*, el cual fue propuesto, de manera independiente por Kurano y Mandl en [22, 23], respectivamente, y a la política resultante la llamaremos política adaptada.

Esta es la clase de sistemas de control estocástico y problemas de control óptimo que estudiaremos en el presente artículo.

El estudio de los problemas de control se puede dividir, por ejemplo:

1. Según el tipo de espacios de estados:
   (a) espacio numerable;
   (b) espacio no numerable e.g., $\mathbb{R}$ o en general un espacio de Borel (i.e. subconjunto de Borel de un espacio métrico separable y completo);
2. Segun el tipo de indice de funcionamiento:
   (a) costo descontado;
   (b) costo promedio;
   (c) costo total.
3. Según la dinámica del sistema
   (a) tiempo discreto;
   (b) tiempo continuo.

Con el fin de ilustrar de una forma más amigable la teoría y presentar las ideas en términos generales sin entrar en formalismos, nuestra exposición se centrará en el estudio de sistemas de control estocástico a tiempo discreto de la forma (1) con espacios de estados numerable, bajo el criterio de optimalidad de costo descontado y cuando la distribución de las variables aleatorias $\{\xi_t\}$ es desconocida por el controlador. En particular proponemos un esquema de estimación y control usando la distribución empírica como estimador. Este problema, pero considerando espacios y escenarios más generales y bajo diferentes

criterios de optimalidad, ha sido ampliamente estudiado (véase, e.g., [5, 11, 12, 13, 14, 15, 17, 18, 24, 25, 26, 27] y sus referencias).

El artículo está estructurado de la siguiente manera. En la sección 2 se introduce la clase de modelos de control que estudiaremos. En la sección 3 se define el criterio de optimalidad de costo descontado así como el problema de control óptimo asociado. Después en la sección 4 presentamos el proceso de estimación y control con el cual se obtiene un algoritmo de aproximación a la función de costo óptimo y a la política óptima. Finalmente, concluimos con la sección 5 presentando otros criterios de optimalidad relacionados con el costo descontado donde se han implementado esquemas de estimación y control.

2. El modelo de control

Las componentes que describen un sistema de control estocástico de la forma (1) se agrupan en el siguiente arreglo al cual se le conoce como *modelo de control*:

$$\mathcal{M} := (X, A, \{A(x) \subset A | x \in X\}, S, F, \theta, c) \tag{2}$$

En nuestro caso, estas componentes satisfacen las siguientes condiciones. El espacio de estados X, el espacio de control A y el espacio de perturbaciones aleatorias S son conjuntos numerables. Para cada estado $x \in X$, $A(x)$ es un subconjunto finito no vacío de A que representa el conjunto de acciones admisibles cuando el sistema se encuentra en el estado x. Definimos el conjunto

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

de pares admisibles estado-acción el cual es un subconjunto del producto cartesiano de X y A. La función $F : X \times A \times S \to X$, como en (1), es una función conocida y representa la dinámica del sistema. Además, θ denota la función de probabilidad común de las variables aleatorias (v.a.) independientes e identicamente distribuidas (i.i.d.) ξ_t en (1), las cuales toman valores en el espacio S y están definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Es decir,

$$\theta(s) = P(\xi_t = s), \ t \in \mathbb{N}_0, \ s \in S. \tag{3}$$

Finalmente, el costo por etapa $c(x, a)$ es una función no negativa y acotada $c : \mathbb{K} \to \mathbb{R}$, i.e., existe una constante positiva M tal que

$$c(x, a) \leq M, \quad \forall (x, a) \in \mathbb{K}. \tag{4}$$

A lo largo del artículo, supondremos que el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ es fijo, y *a.s.* significa *casi seguramente con respecto a* P.

El modelo $\mathcal{M}$ colecciona todas las componentes que desciben el sistema estocástico controlado (1). En efecto, si al tiempo t el sistema se encuentra en el estado $x_t = x \in X$, el controlador elige un control admisible $a_t = a \in A(x)$. Entonces se genera un costo $c(x, a)$ y el sistema evoluciona al nuevo estado $x_{t+1} = y \in X$ cuya distribución condicional es determinada por la distribución θ de la siguiente manera

$$\begin{aligned} P_{x,y}(a) &: = P[x_{t+1} = y | x_t = x, a_t = a] \\ &= P[F(x_t, a_t, \xi_t) = y | x_t = x, a_t = a] = \sum_{k \in S_y} \theta(k), \end{aligned} \tag{5}$$

donde

$$S_y := \{s \in S : F(x, a, s) = y\}.$$

Las acciones se eligen de acuerdo a reglas de decisión las cuales son funciones que determinan el control en cada etapa. Estas reglas de decisión forman lo que llamaremos *políticas de control*, cuya definición presentamos a continuación.

Definimos el espacio de historias admisibles hasta el tiempo t como $\mathbb{H}_0 := X$ y $\mathbb{H}_t := (\mathbb{K} \times S)^t \times X$, $t \geq 1$. De esta manera, un elemento de $\mathbb{H}_t$ toma la forma

$$h_t = (x_0, a_0, \xi_0, \ldots, x_{t-1}, a_{t-1}, \xi_{t-1}, x_t).$$

Definimos además el conjunto

$$\mathbb{F} := \{f : X \to A; f(x) \in A(x)\}.$$

Una regla de decisión en la etapa t, es una función $g_t : \mathbb{H}_t \to \mathbb{A}$ tal que $g_t(h_t) \in A(x_t)$. Es decir, mediante la función g_t se eligen controles admisibles tomando en cuenta la historia del proceso: $a_t = g_t(h_t)$. Una regla de decisión es markoviana si existe $f_t \in \mathbb{F}$ tal que $g_t(h_t) = f_t(x_t)$. Es decir, una regla de decisión markovina elige los controles tomando en cuenta solo el estado actual del sistema: $a_t = f_t(x_t)$.

Definición 2.1. a) Una política de control es una sucesión de reglas de decisión $\pi = \{g_t\}$. Si las reglas de decisión son markovianas diremos que π es una política markoviana.

(b) Una política markoviana $\pi = \{f_t\}$ es estacionaria si existe $f \in \mathbb{F}$ tal que $f_t = f \ \forall t \in \mathbb{N}_0$. En este caso la política toma la forma $\pi = \{f, f, \ldots\} := \{f\}$, de tal manera que $a_t = f(x_t)$.

Denotaremos por Π al conjunto de todas las políticas, por Π_M al conjunto de políticas de Markov, y por Π_S al conjunto de políticas estacionarias. Por lo tanto $\Pi_S \subset \Pi_M \subset \Pi$.

Observación 2.1. *Consideremos el sistema de control estocástico (1). Observe que bajo una política markoviana $\pi = \{f_t\}$, el proceso $\{x_t\}$ se*

genera mediante la ecuación

$$x_{t+1} = F(x_t, f_t(x_t), \xi_t), \quad t \in \mathbb{N}_0.$$

A partir de este hecho, es fácil mostrar que $\{x_t\}$ *es una cadena de Markov con probabilidades de transición dadas por (5) (véase, e.g., [20]). En el caso de una política estacionaria* $\pi = \{f\}$*, el proceso* $\{x_t\}$ *que se genera es una cadena de Markov homogenea en el tiempo.*

2.1 Ejemplos

A continuación presentamos algunos ejemplos clásicos de sistemas de control estocástico de la forma (1).

Sistemas cash-balance. El problema consiste en controlar el nivel de dinero en efectivo de una firma (banco, cajero, etc.) para satisfacer la demanda de efectivo de los clientes (véase e.g., [19]). Las variables que intervienen en en el sistema son:

$x_t =$ dinero en efectivo que se tiene disponible al tiempo t, representando el estado del sistema;

$a_t =$ variable de control que representa la cantidad de dinero que se decide retirar $-a_t$ (si $a_t < 0$), o la cantidad de dinero que se decide suministrar a_t (si $a_t > 0$). El dinero que quede disponible después de tomar ésta decisión es para satisfacer la demanda de efectivo al tiempo t;

$\xi_n =$ demanda de efectivo durante el período t. Una demanda positiva significa que se retira dinero, mientras una demanda negativa significa depósito.

Entonces, el proceso cash-balance $\{x_t\}$ evoluciona de acuerdo a una ecuación en diferencia estocástica de la forma

$$x_{t+1} = x_t + a_t + \xi_t.$$

Sistemas de producción-inventario. El problema en un sistema de inventario es controlar la cantidad de artículos que se ordena o produce para poner en existencia con el fin de satisfacer la demanda. En este caso las variables son:

$x_t =$ cantidad de artículos en el inventario al principio del período t;

$a_t =$ cantidad de artículos que se decide ordenar o producir al inicio del período t para satisfacer la demanda;

$\xi_t =$ demanda del artículo durante el período t.

Por lo tanto, la evolución en el tiempo del proceso de inventario $\{x_t\}$ es:

$$x_{t+1} = (x_t + a_t - \xi_t)^+ := \text{máx}\{0, x_t + a_t - \xi_t\}. \tag{6}$$

En general, ecuaciones de la forma

$$x_{t+1} = x_t \pm a_t \pm \xi_t,$$

ya sea que consideran la parte positiva como en (6) o no, modelan los llamados *sistemas de almacenamiento* dentro de los cuales se encuentran, además de los ejemplos anteriores, sistemas de espera (colas), modelos de teoría de riesgo en seguros, sistema de control de presas y algunos modelos de telecomunicaciones (véase, e.g., [28, 31]).

Existen otras familias de sistemas de control estocástico clásicos como por ejemplo los *procesos autorregresivos* usados en el área de economía los cuales tienen la forma

$$x_{t+1} = \rho(a_t)x_t + \xi_t,$$

donde ρ es una función apropiada. También se encuentran sistemas con dinámicas lineales que aparecen en muchas áreas de la ingeniería:

$$x_{t+1} = \beta x_t + \gamma a_t + \xi_t.$$

Otra variedad de ejemplos se pueden encontrar en [3, 16, 32].

3. El problema de control óptimo

Es claro que el comportamiento del sistema depende de la condición inicial $x_0 = x \in X$ y de la política de control que se aplica para determinar los controles en cada etapa. Su evolución en el tiempo está determinada por las probabilidades de transición (5). Por lo tanto, el *índice de funcionamiento* o *función objetivo* que mide el rendimiento de una política de control será un funcional de costo definido por medio de un valor esperado. La forma específica de este funcional de costo esperado o índice de funcionamiento depende de las condiciones del problema y de los objetivos del controlador. Entre los índices mas comunes tenemos los siguientes.

- Costo total esperado con horizonte finito $N < \infty$, cuando se aplica la política $\pi \in \Pi$ y el estado inicial es $x_0 = x \in X$:

$$J_N(\pi, x) := E\left[\sum_{t=0}^{N-1} c(x_t, a_t) + C_N(x_N)\right], \tag{7}$$

  donde C_N es una función de costo terminal.
- Costo descontado total esperado cuando se aplica la política $\pi \in \Pi$ y el estado inicial es $x_0 = x \in X$:

$$V(\pi, x) := E\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t), \tag{8}$$

  donde a $\alpha \in (0, 1)$ se le conoce como el factor de descuento.

- Costo promedio por etapa cuando se aplica la política $\pi \in \Pi$ y el estado inicial es $x_0 = x \in X$:

$$J(\pi, x) := \lim_{n\to\infty} \frac{1}{n} E \sum_{t=0}^{n-1} c(x_t, a_t). \tag{9}$$

El índice costo total (7) es el más natural para fines prácticos ya que regularmente el número de etapas en las que se quiere controlar el sistema es finito; sin embargo, para N muy grande resulta inconveniente desde el punto de vista computacional. En estos casos se recurre a los índices con horizonte infinito (8) y (9). Dentro de estos últimos, el costo descontado refleja el comportamiento en las primeras etapas de operación, es decir, las decisiones tomadas en las primeras etapas influyen más en el costo total. Esto se debe a que el factor de descuento α es un número entre 0 y 1, de tal forma que después de transcurrido un determinado número de etapas el factor α^t se desvanece a cero. A partir de este hecho, sus aplicaciones se encuentran principalmente en problemas donde V tiene una interpretación monetaria, y en este caso el factor de descuento toma la forma $\alpha = \frac{1}{1+i}$ donde i es la tasa de interés. Entonces, el término α^t representa el descuento para obtener el valor presente t períodos después.

Por otro lado, a diferencia del costo descontado, el costo promedio (9) mide el comportamiento asintótico del sistema, y es en este sentido que sus aplicaciones se encuentran en problemas donde es necesario llevar a cabo un análisis de estabilidad.

En este trabajo solo nos centraremos en estudiar el índice de costo descontado (8). Bajo este contexto, el *problema de control óptimo* (PCO) puede establecerse de la siguiente manera:

Dado un modelo de control $\mathcal{M}$ (véase (2)) y una familia de políticas Π, el PCO para el criterio de optimalidad de costo descontado consiste en encontrar una política $\pi^* \in \Pi$ tal que

$$V(\pi^*, x) = \min_{\pi\in\Pi} V(\pi, x), \quad \forall x \in X.$$

A la política π^* se le llama política óptima. Además, a la función

$$V^*(x) := \min_{\pi\in\Pi} V(\pi, x)$$

le llamaremos función de valor óptimo.

Como podemos observar, la solución al PCO no es trivial; minimizar un funcional de costo definido como el valor esperado de una serie de funciones, sobre un conjunto de sucesiones de funciones (políticas), lo hace un problema difícil. Existen varias técnicas de solución del PCO como por ejemplo, Programación Dinámica, Programación Lineal, Principio del Máximo y Programación Convexa. Dentro del contexto de

estimación y control la *Programación Dinámica* (PD) es la más conveniente debido a que su formulación es mediante algoritmos iterativos. En términos generales, la idea de la PD es descomponer el problema original en sub-problemas de optimización correspondientes a cada etapa y que resultan más sencillos de resolver, y luego obtener algoritmos recursivos y combinarlos con técnicas de solución de ecuaciones funcionales estudiadas en los textos de Análisis Funcional. Para fijar ideas, sea $B(X)$ el espacio lineal normado de todas las funciones acotadas $v : X \to \mathbb{R}$ con la norma

$$||v|| := \sup_{x \in X} |v(x)|. \tag{10}$$

Se sabe (véase [21]) que $B(X)$ es un espacio de Banach, i.e., un espacio vectorial, normado y completo. Para $u \in B(X)$ definimos el operador

$$\begin{aligned} Tu(x) &: = \min_{a \in A(x)} \{c(x,a) + \alpha E\left[u(F(x,a,\xi))\right]\} \\ &= \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{s \in S} u(F(x,a,s))\theta(s) \right\}. \end{aligned}$$

Por cálculos directos es fácil mostrar que el operador T satisface las siguientes propiedades:

T1 $Tu \in B(X)$ si $u \in B(X)$;

T2 T es un operador de contracción módulo $\alpha \in (0,1)$, i.e., para cada par de funciones $u, v \in B(X)$,

$$||Tu - Tv|| \leq \alpha ||u - v||.$$

Definamos el siguiente proceso iterativo. Para una función arbitraria $v \in B(X)$,

$$\begin{aligned} v_0 &= v; \\ v_n(x) &= Tv_{n-1}(x) \\ &= \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{s \in S} v_{n-1}(F(x,a,s))\theta(s) \right\}, \quad n \geq 1. \end{aligned} \tag{11}$$

A partir del hecho de que $B(X)$ es un espacio de Banach, las propiedades T1 y T2 implican el siguiente resultado (el cual es conocido como el *Teorema de Punto Fijo de Banach (véase, e.g., [21])).*

Lema 3.1. *(a) Existe una única función $v^* \in B(X)$ que es punto fijo del operador T, i.e., $Tv^*(x) = v^*(x)$, $x \in X$.*

(b) Para cualquier función $v_0 \in B(X)$,

$$||v_n - v^*|| \leq \alpha^n ||v_0 - v^*||.$$

En particular, para $v_0 = 0$ *tenemos*

$$||v_n - v^*|| \leq \alpha^n ||v^*||.$$

Este resultado es la base para proponer un método de solución del PCO así como un procedimiento de aproximación. Específicamente tenemos el siguiente resultado.

Teorema 3.1. *a) La función de valor* $V^*(x)$ *es la única función en* $B(X)$ *que satisface la ecuación*

$$V^*(x) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{s \in S} V^* \left[F(x,a,s) \right] \theta(s) \right\}, \quad x \in X, \quad (12)$$

i.e., V^* *es el único punto fijo del operador* T, $TV^* = V^*$.

b) Existe $f^* \in \mathbb{F}$ *tal que* $f^*(x) \in A(x)$ *minimiza el lado derecho de (12), i.e.,*

$$V^*(x) = c(x, f^*(x)) + \alpha \sum_{s \in S} V^* \left[F(x, f^*(x), s) \right] \theta(s), \quad x \in X. \quad (13)$$

c) La política estacionaria $\pi^* = \{f^*\} \in \Pi_S$ *es óptima para el PCO, i.e.,*

$$V^*(x) = V(\pi^*, x), \quad x \in X.$$

d) Para cada $n \in \mathbb{N}$, $\|v_n - V^*\| \leq \dfrac{\alpha^n M}{1-\alpha}$. *Por lo tanto, cuando* $n \to \infty$,

$$v_n(x) \to V^*(x), \quad x \in X.$$

A la ecuación (12) se le conoce como *Ecuación de Optimalidad* (EO) o *Ecuación de Programación Dinámica*.

La demostración de la parte (a) del teorema 3.1 consiste en mostrar que $v^* = V^*$ (ver lema 3.1). La parte (b) se sigue del hecho de que el conjunto $A(x)$ es finito, mientras que para obtener la parte (c) se deben aplicar argumentos de programación dinámica. Para demostrar la parte (d), observemos que para cualquier política $\pi \in \Pi$ y estado inicial $x_0 = x \in X$,

$$V(\pi, x) := E\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \leq M \sum_{t=0}^{\infty} \alpha^t = \frac{M}{1-\alpha}.$$

Por lo tanto

$$V^*(x) = \min_{\pi \in \Pi} V(\pi, x) \leq \frac{M}{1-\alpha},$$

lo cual implica

$$\|V^*\| \leq \frac{M}{1-\alpha}.$$

Entonces la parte (d) se sigue del lema 3.1(b), del teorema 3.1(a) y de la desigualdad anterior.

A la ecuación (11) se le conoce como Algoritmo de Iteración de Valores, y proporciona un esquema de aproximación a la función de valor óptimo con una tasa de convergencia geométrica dada por el teorema 3.1(d).

4. Estimación y control

Consideremos el modelo de control (2) asociado al sistema de control estocástico (1):

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0.$$

Ahora supondremos que $\{\xi_t\}$ es una sucesión observable de v.a. i.i.d. cuya distribución θ es *desconocida* por el controlador. Observemos que la solución al PCO dada en el teorema 3.1 no es accesible para el controlador ya que la ecuación de optimalidad (12) depende fuertemente de la función de probabilidad desconocida θ. El objetivo ahora es introducir un procedimiento que combine métodos de estimación estadística para θ y procesos de control para aproximar a la función de valor óptimo y a la política óptima. En particular usaremos la distribución empírica como estimador, la cual se define como:

$$\theta_t(k) = \frac{1}{t} \sum_{j=0}^{t-1} I_k(\xi_j), \tag{14}$$

donde

$$I_k(\xi_j) = \begin{cases} 1, & \text{si } \xi_j = k, \\ \\ 0, & \text{si } \xi_j \neq k. \end{cases}$$

En términos generales, el proceso de estimación y control consiste en lo siguiente. En la etapa t, cuando el proceso se encuentra en el estado $x_t = x \in X$, antes de elegir el control a_t y con el conocimiento de la historia observada $\bar{\xi}_t := (\xi_0, \xi_1, \ldots, \xi_{t-1})$, el controlador obtiene una estimación $\theta_t(\bar{\xi}_t)$ de θ por medio de (14). Enseguida elige el control $a_t = a_t(\theta_t) = a \in A(x_t)$ y sucede lo siguiente: (a) se genera un costo $c(x, a)$, y (b) el proceso avanza a un nuevo estado $x_{t+1} = y \in X$ de acuerdo a la probabilidad de transición

$$P_{x,y}(a) := P[x_{t+1} = y | x_t = x, a_t = a] = \sum_{k \in S_y} \theta(k),$$

donde

$$S_y := \{s \in S : F(x, a, s) = y\}.$$

Una vez en el estado y, el proceso se repite una y otra vez. Los costos se acumulan de acuerdo al índice de costo descontado.

Observe que para cada $k \in S$, por la ley fuerte de los grandes números,

$$\theta_t(k) \to \theta(k) \text{ a.s. cuando } t \to \infty. \tag{15}$$

Además, para cualquier función acotada v,

$$\sum_{k\in S} v(k)\theta_t(k) = \frac{1}{t}\sum_{j=0}^{t-1} v(\xi_j) \to \sum_{k\in S} v(k)\theta(k) \text{ a.s., cuando } t \to \infty, \tag{16}$$

lo cual implica (debido a que V^* es una función acotada)

$$\sum_{k\in S} V^*\left[F(x,a,k)\right]\theta_t(k) \to \sum_{k\in S} V^*\left[F(x,a,k)\right]\theta(k) \text{ a.s., cuando } t \to \infty, \tag{17}$$

para cada $(x,a) \in \mathbb{K}$. De hecho, la convergencia en (17) es uniforme en $(x,a) \in \mathbb{K}$, i.e., cuando $t \to \infty$.

$$\eta_t := \sup_{(x,a)\in\mathbb{K}} \left| \sum_{k\in S} V^*\left[F(x,a,k)\right]\theta_t(k) - \sum_{k\in S} V^*\left[F(x,a,k)\right]\theta(k) \right| \to 0 \text{ a.s.} \tag{18}$$

Estas propiedades de la distribución empírica son bien conocidas (véase e.g., [4, 6, 29] para contextos más generales).

Las relaciones (15)-(18) proporcionan diferentes propiedades del proceso de estimación empírico de la distribución θ, de las cuales se deduce que entre más observaciones se tengan de la v.a. ξ_t mejor es la estimación. Sin embargo, el hecho de que el índice de costo descontado le dé más importancia a las decisiones tomadas en las primeras estapas, precisamente donde el método de estimación proporciona una información pobre respecto a la distribución desconocida θ, implica que la política resultante de este proceso no necesariamente sea óptima. Por lo tanto la optimalidad de políticas que combinan estimación estadística y control se estudia en un sentido asintótico, cuya definición es motivada por el siguiente hecho.

Sea $\Phi : \mathbb{K} \to \mathbb{R}$ la función definida como

$$\Phi(x,a) := c(x,a) + \alpha \sum_{s\in S} V_\alpha^*\left[F(x,a,s)\right]\theta(s) - V^*(x). \tag{19}$$

Observemos que del teorema 3.1 (b), si $f^* \in \mathbb{F}$ satisface (13), entonces $\Phi(x, f^*(x)) = 0$, y por el teorema 3.1 (c) $\pi = \{f^*\} \in \Pi_S$ es una política óptima. De aquí, la optimalidad asintótica puede definirse de la siguiente manera (véase, e.g., [30]).

Definición 4.1. Una política de control $\pi = \{g_t\} \in \Pi$ es asintóticamente óptima para el modelo de control $\mathcal{M}$ si, para $x \in X$,

$$\lim_{t\to\infty} E\left[\Phi(x_t, g_t(h_t)\right] = 0.$$

4.1 Políticas asintóticamente óptimas

A continuación introducimos un esquema de construcción de una política asintóticamente óptima. Este esquema consiste en combinar el algoritmo de iteración de valores (11) con el método de estimación empírica. Específicamente, sea $\{V_t\}$ la sucesión de funciones en $B(X)$ definida como $V_0 = 0$, y para $t \geq 1$

$$V_t(x) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{s \in S} V_{t-1}(F(x,a,s))\theta_t(s) \right\}. \tag{20}$$

Es fácil mostrar, por inducción, que

$$V_t(x) \leq \frac{M}{1-\alpha}, \quad t \in \mathbb{N}_0, x \in X.$$

Además, como $A(x)$ es finito, tenemos que, para cada $t \in \mathbb{N}$, existe $\bar{f}_t = \bar{f}_t^{\theta_t} \in \mathbb{F}$ tal que

$$V_t(x) = c(x, \bar{f}_t(x)) + \alpha \sum_{s \in S} V_{t-1}(F(x, \bar{f}_t(x), s))\theta_t(s) \quad a.s. \tag{21}$$

Las reglas de decisión $\bar{f}_t$ provienen de la combinación del método de estimación empírica y el proceso de minimización. La política markoviana que determinan estas reglas de decisión la denotaremos como $\bar{\pi} = \left\{\bar{f}_t\right\} \in \Pi_M$, donde $f_0 \in \mathbb{F}$ es arbitraria. El objetivo, por lo tanto, es mostrar que V_t «aproxima» a la función de valor V^* y que $\bar{\pi}$ es una política asintóticamente óptima. Esto lo establecemos en el siguiente resultado.

Teorema 4.1. *(a)* $\|V_t - V^*\| \to 0$ *a.s. cuando* $t \to \infty$.

b) $E\left[\Phi(x_t, \bar{f}_t(x_t)\right] \to 0$ *cuando* $t \to \infty$*, i.e.,* $\bar{\pi}$ *es una política asintóticamente óptima.*

Demostración. De las ecuaciones (12) y (20), para cada $x \in X$ y $t \in \mathbb{N}$

$$\begin{aligned} |V^*(x) - V_t(x)| \quad \leq \quad & \sup_{a \in A(x)} \left| \alpha \sum_{s \in S} V^*\left[F(x,a,s)\right]\theta(s) \right. \\ & \left. -\alpha \sum_{s \in S} V_{t-1}(F(x,a,s))\theta_t(s) \right|. \end{aligned}$$

Sumando y restando $\alpha \sum_{s\in S} V^*\left[F(x,a,s)\right]\theta_t(s)$ y agrupando términos obtenemos

$$\begin{aligned}
&|V^*(x) - V_t(x)| \\
\leq\ & \sup_{a\in A(x)}\left|\alpha\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta(s) - \alpha\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta_t(s)\right| \\
&+ \sup_{a\in A(x)}\left|\alpha\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta_t(s) - \alpha\sum_{s\in S} V_{t-1}(F(x,a,s))\theta_t(s)\right| \\
\leq\ & \eta_t + \alpha\left\|V^* - V_{t-1}\right\| \quad \text{(por (18))}
\end{aligned}$$

De aquí

$$\left\|V^* - V_t\right\| \leq \eta_t + \alpha\left\|V^* - V_{t-1}\right\| \quad a.s. \tag{22}$$

Sea $\lambda := \lim\sup_{t\to\infty}\left\|V^* - V_t\right\| < \infty$. Entonces, tomando límite superior en ambos lados de la desigualdad (22), de (18) tenemos que $\lambda \leq \alpha\lambda$, lo cual, como $\alpha \in (0,1)$, implica que $\lambda = 0$. Por lo tanto

$$\lim_{t\to\infty}\left\|V^* - V_t\right\| = 0 \quad a.s.$$

(b) Para cada $t \in \mathbb{N}$, definimos la función

$$\Phi_t(x,a) := c(x,a) + \alpha\sum_{s\in S} V_{t-1}\left[F(x,a,s)\right]\theta_t(s) - V_t(x). \tag{23}$$

Observe que por la definición de la polítca $\bar{\pi}$ (ver (21))

$$\Phi_t(x, \bar{f}_t(x)) = 0, \quad x \in X, t \in \mathbb{N}.$$

Entonces

$$\begin{aligned}
&\Phi(x_t, \bar{f}_t(x_t)) = \left|\Phi(x_t, \bar{f}_t(x_t)) - \Phi_t(x_t, \bar{f}_t(x_t))\right| \leq |V^*(x_t) - V_t(x_t)| \\
&+ \sup_{a\in A(x_t)}\left|\alpha\sum_{s\in S} V^*\left[F(x_t,a,s)\right]\theta(s) - \alpha\sum_{s\in S} V_{t-1}(F(x_t,a,s))\theta_t(s)\right| \\
&\leq \left\|V^* - V_t\right\| + \sup_{(x,a)\in\mathbb{K}}\left|\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta(s) - \sum_{s\in S} V_{t-1}(F(x,a,s))\theta_t(s)\right|.
\end{aligned}$$

Ahora, sumando y restando $\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta_t(s)$

$$\begin{aligned}
\Phi\left(x_t, \bar{f}_t(x_t)\right) &\leq \left\|V^* - V_t\right\| \\
&+ \sup_{(x,a)\in\mathbb{K}}\left|\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta(s) - \sum_{s\in S} V^*(F(x,a,s))\theta_t(s)\right| \\
&+ \sup_{(x,a)\in\mathbb{K}}\left|\sum_{s\in S} V^*\left[F(x,a,s)\right]\theta_t(s) - \sum_{s\in S} V_{t-1}(F(x,a,s))\theta_t(s)\right| \\
&\leq \left\|V^* - V_t\right\| + \left\|V^* - V_{t-1}\right\| + \eta_t.
\end{aligned}$$

Entonces, de (18) y la parte (a) del teorema tenemos que

$$\Phi\left(x_t, \bar{f}_t(x_t)\right) \to 0 \text{ a.s., cuando } t \to \infty. \tag{24}$$

Finalmente, como Φ es una función acotada, la convergencia casi segura en (24) implica la convergencia en media (véase, e.g., [1])

$$E\Phi\left(x_t, \bar{f}_t(x_t)\right) \to 0, \text{ cuando } t \to \infty,$$

lo cual demuestra el teorema. □

5. Otros criterios de costo descontado

Concluimos el artículo exponiendo algunos modelos no usuales, pero interesantes por sus aplicaciones, relacionados con el criterio de costo descontado, donde se han implementado esquemas de estimación y control.

El criterio de costo descontado es el índice de funcionamiento más estudiado en la teoría de control óptimo, ya sea por lo atractivo y fácil en términos matemáticos y/o por su natural interpretación en modelos de economía y finanzas. En ambos casos, regularmente se supone que el factor de descuento permanece constante durante la evolución del sistema, y esto es precisamente lo que simplifica su análisis matemático. Sin embargo, desde el punto de vista de las aplicaciones, que α sea constante puede ser demasiado restrictivo o poco realista en algunos casos. En efecto, en ciertos modelos financieros, los factores de descuento son regularmente funciones de las tasas de interés, las cuales a su vez son inciertas. Tal incertidumbre se puede deber a la cantidad de dinero circulando y/o decisiones de ciertas empresas líderes en el mercado y/o factores aleatorios externos cuya distribución es imposible de conocer. Por tanto, un factor de descuento constante difícilmente modelaría esta situación. A continuación presentamos dos índices de funcionamiento con factor de descuento no constante que sirven para modelar problemas como los anteriores.

Modelos con factores aleatorios y exponencialmente descontados.

Estos modelos tratan con sistemas que evolucionan de acuerdo a las ecuaciones en diferencia

$$\begin{aligned} x_{t+1} &= F(x_t, \alpha_t, a_t, \xi_t), \\ \alpha_{t+1} &= G(\alpha_t, \eta_t), \end{aligned}$$

donde F y G son funciones conocidas, x_t, α_t y a_t son el estado, el factor de descuento y el control al tiempo t, respectivamente. Además $\{\xi_t\}$ y

$\{\eta_t\}$ son sucesiones de v.a. i.i.d. con distribuciones desconocidas θ^ξ y θ^η, respectivamente.

El costo descontado cuando se usa la política $\pi \in \Pi$, dado el estado inicial $x_0 = x$ y el factor de descuento inicial $\alpha_0 = \alpha$, se define como

$$V(\pi, x, \alpha) := E\left[\sum_{t=0}^{\infty} \exp(-S_t) c(x_t, a_t)\right], \tag{25}$$

donde $S_t = \sum_{i=0}^{t-1} \alpha_i$ si $t \geq 1$, $S_0 = 0$. El problema de estimación y control de esta clase de sistemas ha sido estudiado en [7, 8, 9, 10].

Modelos con factores aleatorios que dependen del estado y control.

En este caso se considera que el factor de descuento es una función de la forma

$$\tilde{\alpha}(x_t, a_t, \chi_{t+1}),$$

donde x_t y a_t son el estado y el control al tiempo t, respectivamente, y $\{\chi_t\}$ es una sucesión de v.a. i.i.d. con distribución desconocida. Entonces, el costo descontado cuando se aplica la política $\pi \in \Pi$ y el estado inicial es $x_0 = x$ tiene la forma

$$V(\pi, x) := E\left[\sum_{t=0}^{\infty} \tilde{\Gamma}_t c(x_t, a_t)\right], \tag{26}$$

donde

$$\tilde{\Gamma}_t = \prod_{k=0}^{t-1} \tilde{\alpha}(x_k, a_k, \xi_{k+1}) \quad \text{si} \quad t \in \mathbb{N}, \text{ y } \tilde{\Gamma}_0 = 1.$$

El problema de estimación y control para esta clase de sistemas se estudió en [27].

Bibliografía

[1] R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] R. Bellman, *Dynamic programming*, Princeton University Press, Princeton New Jersey, 1957.

[3] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, N.J., 1987.

[4] P. Billingsley y F. Topsoe, «Uniformity in weak convergence», *Z. Wahrsch. Verw. Geb.*, vol. 7, 1967, 1–16.

[5] R. Cavazos-Cadena, «Nonparametric adaptive control of discounted stochastic systems with compact state space», *J. Optim. Theory Appl.*, vol. 65, 1990, 191–207.

[6] P. Gaenssler y W. Stute, «Empirical processes: a survey for i.i.d. random variables», *Ann. Probab.*, vol. 7, 1979, 193–243.

[7] J. González-Hernández, R. R. López-Martínez y J. A. Minjárez-Sosa, «Adaptive policies for stochastic systems under a randomized discounted criterion», *Bol. Soc. Mat. Mex.*, vol. 14, 2008, 149–163.

[8] _____, «Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion», *Kybernetika*, vol. 45, 2009, 737–754.
[9] J. González-Hernández, R. R. López-Martínez, J. A. Minjárez-Sosa y J. R. Gabriel-Arguelles, «Constrained Markov control processes with randomized discounted cost criteria: occupation measures and extremal points», *Risk and Decision Analysis*, vol. 4, 2013, 163–176.
[10] _____, «Constrained Markov control processes with randomized discounted rate: infinite linear programming approach», *Optimal Control, Applications and Methods*, vol. 35, 2013, 575–591.
[11] E. I. Gordienko, «Adaptive strategies for certain classes of controlled Markov processes», *Theory Probab. Appl.*, vol. 29, 1985, 504–518.
[12] E. I. Gordienko y J. A. Minjárez-Sosa, «Adaptive control for discrete-time Markov processes with unbounded costs: average criterion», *ZOR- Math. Methods of Oper. Res.*, vol. 48, 1998, 37–55.
[13] _____, «Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion», *Kybernetika*, vol. 34, 1998, 217–234.
[14] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer, New York, 1989.
[15] O. Hernández-Lerma y R. Cavazos-Cadena, « Density estimation and adaptive control of Markov processes: average and discounted criteria», *Acta Appl. Math.*, vol. 20, 1990, 285–307.
[16] O. Hernández-Lerma y J. B. Lasserre, *Discrete-time Markov control processes: Basic optimality criteria*, Springer, New York, 1996.
[17] N. Hilgert y J. A. Minjárez-Sosa, «Adaptive policies for time-varying stochastic systems under discounted criterion», *Math. Methods Oper. Res..*, vol. 54, 2001, 491–505.
[18] _____, «Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria», *Math. Methods Oper. Res.*, vol. 63, 2006, 443–460.
[19] A. Hordjik y A. A. Yushkevich, «Blackwell optimality in the class of all policies in markov decision chains with borel state space an unbounded rewards», *Math. Methods Oper. Res.*, vol. 50, 1999, 421–448.
[20] J. G. Kemeny y J. L. Snell, *Finite Markov Chains*, Springer-Verlag, New York, 1976.
[21] E. Kreyszig, *Introductory functional analysis with applications*, Wiley, USA, 1978.
[22] M. Kurano, «Discrete-time markovian decision processes with an unknown parameter-average return criterion», *J. Oper. Res. Soc. Japan*, vol. 15, 1972, 67–76.
[23] P. Mandl, «Estimation and control in Markov chains», *Adv. Appl. Probab.*, vol. 6, 1974, 40–60.
[24] J. A. Minjárez-Sosa, «Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under average criterion», *Appl. Math. (Warsaw)*, vol. 26, 1999, 267–280.
[25] _____, «Approximation and estimation in Markov control processes under discounted criterion», *Kybernetika*, vol. 40, 2004, 681–690.
[26] _____, «Empirical estimation in average Markov control processes», *Applied Mathematics Letters*, vol. 21, 2008, 459–464.
[27] _____, «Markov control models with unknown random state-action-dependent discount factors», *TOP*, vol. 23, 2015, 743–772.
[28] N. U. Prabhu, *Stochastic Storage Processes*, Springer, New York, 1980.
[29] R. R. Rao, «Relations between weak and uniform convergence of measures with applications», *Ann. Math. Statistics*, vol. 33, 1962, 659–680.
[30] M. Schäl, «Estimation and control in discounted stochastic dynamic programming», *Stochastics*, vol. 20, 1987, 51–71.
[31] M. J. Sobel, «Reservoir management models», *Water Resources Research*, vol. 11, 1975, 767–778.
[32] N. L. Stockey, *Recursive Methods in Economic Dynamics*, Harvard University Press, Cambridge, MA, 1989.