SAHUARUS
Revista Electrónica de Matemáticas,
Departamento de Matemáticas, Universidad de Sonora,
Vol. 1, No. 2, Septiembre 2016, pp. 73–81.

Modelos de control markovianos: una aplicación de estimación empírica al caso con descuento

Luz del Carmen Rosas Rosas
Jessica Liliana Leyva Domínguez

Departamento de Matemáticas
Universidad de Sonora
e-mail: lcrosas@gauss.mat.uson.mx
jessicaliliana.leyva@gmail.com

Resumen

En este trabajo se abordan modelos de control markovianos en tiempo discreto con espacios de estado y de control numerables, con costos acotados y distribución de perturbaciones aleatorias desconocida θ. Asumiendo perturbaciones observables y aplicando la distribución empírica como estimador de θ, el objetivo es construir políticas adaptadas, las cuales son asintóticamente óptimas en costo descontado.

1 Introducción

Los modelos de control markovianos constituyen una clase de modelos de control estocástico en tiempo discreto, cuya evolución en el tiempo la podemos describir de la siguiente manera. Si al tiempo $t = 0, 1, 2, \ldots$ el sistema se encuentra en el estado $x_t = x$, entonces el controlador elige una acción o control $a_t = a$ y ocurre lo siguiente: 1) se produce un costo c que depende tanto del estado como de la acción elegida; 2) el sistema se mueve a un nuevo estado $x_{t+1} = y$ de acuerdo a una ley de transición. Una vez ocurrido lo anterior, el proceso se repite. Luego, bajo tal escenario los costos de operación se acumulan durante la evolución del sistema, de manera que el propósito del controlador es encontrar una política de control que minimice el costo total acumulado definido por un índice de funcionamiento. En particular nos enfocaremos en el índice de costo total esperado α-descontado, el cual se introducirá en la Sección 2.

Una clase particular de modelos de control markovianos es aquella en la que la dinámica del sistema está modelada por medio de una ecuación en diferencias de la forma

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots,$$

donde $\{\xi_t\}$ es una sucesión de variables aleatorias (v.a's) independientes e idénticamente distribuidas (i.i.d.) con distribución común θ; entonces la ley de transicíon de este modelo de control está determinada por la función F junto con la distribución θ. Cabe señalar que bajo

este esquema regularmente se supone que θ es conocida por el controlador, lo cual en algunos casos resulta ser una hipótesis restrictiva. Por lo tanto, en este trabajo consideramos la situación en que θ es desconocida, de tal manera que el controlador deberá combinar métodos de estimación estadística (distribución empírica) con técnicas de optimización, (véase [6]). De hecho, se llama política adaptada a la que se obtiene de dicha combinación. Así que, el objetivo del presente trabajo es estudiar la optimalidad de políticas adaptadas bajo el criterio de costo descontado. Sin embargo, debido a las características propias del mencionado índice, la optimalidad de las políticas adaptadas se estudiará en un sentido asintótico, ya que bajo métodos de estimación y control no es posible en general garantizar la existencia de políticas óptimas (véase [3]).

A lo largo del trabajo usaremos la siguiente notación. $\mathbb{N}_0$, $\mathbb{N}$ y $\mathbb{R}$ representan, respectivamente, el conjunto de los números enteros no negativos, enteros y reales.

2 Preliminares

Modelo de Control. Consideremos un modelo de control markoviano en tiempo discreto:

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

donde $\mathbb{X}$, $\mathbb{A}$ y $\mathbb{S}$ son los espacios de estado, de control y de perturbaciones aleatorias, respectivamente, los cuales supondremos numerables, además, para cada $x \in \mathbb{X}$, $A(x) \subset \mathbb{A}$ es un conjunto finito no vacío que representa al espacio de controles admisibles cuando el sistema se encuentra en el estado x. Luego, la dinámica del sistema se define mediante la ecuación en diferencias (1), donde

$$F : \mathbb{X} \times \mathbb{A} \times \mathbb{S} \to \mathbb{X}$$

es una función dada y, específicamente, $\{\xi_t\}_{t\in\mathbb{N}_0}$ es una sucesión de v.a's i.i.d. definidas en un espacio de probabilidad común $(\Omega, \mathcal{F}, P)$, tal que toman valores en algún conjunto numerable $\mathbb{S}$, y con distribución común θ, la cual es desconocida por el controlador), es decir,

$$\theta(s) := P\{\xi_t = s\} \quad \forall\, t \in \mathbb{N}_0 \text{ y } s \in \mathbb{S}.$$

Finalmente, el costo por etapa es una función

$$c : \mathbb{K} \to \mathbb{R}$$

donde

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

es el espacio de pares de estado-acción admisibles.

Interpretación. A diferencia de un modelo de control estándar (θ conocida), $\mathcal{M}$ representa un sistema dinámico que evoluciona en el tiempo de la siguiente manera: en la etapa t el sistema se encuentra en el estado $x_t = x \in \mathbb{X}$ y el controlador usa la distribución

empírica para obtener un estimador θ_t de la distribución desconocida θ definido de la siguiente manera:

$$\theta_t(k) := \frac{1}{t}\sum_{j=0}^{t-1}\delta_k(\xi_j),$$

donde

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

Luego el controlador combina este proceso con la historia del sistema para seleccionar un control adaptado al estimador, de modo que

$$a = a_t(\theta_t) \in A(x).$$

Entonces se genera un costo $c(x, a)$ y el sistema avanza a un nuevo estado

$$x_{t+1} = x^{'} \in \mathbb{X}$$

de acuerdo a la ley de probabilidad dada por

$$\begin{aligned} P_{x,x^{'}}(a) : &= P[x_{t+1} = x^{'} \mid x_t = x, a_t = a] \\ &= \sum_{k \in S_F} \theta(k). \end{aligned}$$

con

$$S_F := \left\{ s \in \mathbb{S} : F(x, a, s) = x^{'} \right\}.$$

Y una vez que la transición se presenta el proceso se repite.

Políticas de control. Para cada $t \in \mathbb{N}_0$ definimos el espacio de historias admisibles hasta el tiempo t por

$$\mathbb{H}_0 := \mathbb{X}$$
$$\mathbb{H}_0 := (\mathbb{K} \times \mathbb{S})^t \times \mathbb{X}, \quad t \in \mathbb{N}.$$

Un elemento de $\mathbb{H}_t$ es un vector o historia de la forma

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

con $(x_j, a_j) \in \mathbb{K}$ y $\xi_j \in \mathbb{S}$ para $j = 0, 1, \ldots, t-1$.

Una regla de decisión es un procedimiento para elegir un control en una etapa, la cual puede depender, ya sea de la historia hasta la etapa actual, o bien solamente del estado del sistema en dicho estado. De hecho, una regla de decisión dependiente de la historia es una función

$$f_t : \mathbb{H}_t \to \mathbb{A}$$

tal que $f_t(h_t) \in A(x_t)$. Más aún, si f_t depende de h_t únicamente a través de x_t, decimos que f_t es una regla de decisión markoviana, y en cuyo caso tendremos

$$f_t : \mathbb{X} \to \mathbb{A},$$

tal que $f_t(x) \in A(x)$.

Definición 2.1 *Una política de control admisible (o simplemente una política) es una sucesión*

$$\pi = \{f_0, f_1, \ldots\}$$

de reglas de decisión. Si las f_t son markovianas se dice que π es una política markoviana, y en caso de que $f_t = f$ para alguna $f : \mathbb{X} \to \mathbb{A}$, diremos que π es una política estacionaria.

Denotaremos por Π al espacio de todas las políticas, e identificaremos al conjunto de políticas estacionarias con el conjunto

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

Criterio de optimalidad. Ahora introducimos el llamado índice de funcionamiento, el cual consiste en una función que de alguna manera mide el comportamiento del sistema cuando se utilizan diferentes políticas de control. En particular, en este trabajo consideramos el índice de costo descontado, el cual definimos a continuación.

Definición 2.2 *Para cada $x \in \mathbb{X}$, $\pi \in \Pi$ y $\alpha \in (0,1)$, se define el costo total esperado α-descontado como*

$$V_\alpha(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right],$$

donde α es el factor de descuento y E_x^π denota la esperanza condicional cuando se usa la política π y el estado inicial es $x_0 = x$ (véase [2]).

Problema de control óptimo. Debido a que los costos de operación se acumulan durante la evolución del sistema de acuerdo al índice de costo descontado, entonces, el problema de control óptimo consiste en encontrar una política π^* tal que minimice el costo total esperado α-descontado previamente introducido, es decir,

$$V_\alpha(\pi^*, x) = \inf_{\pi \in \Pi} V_\alpha(\pi, x) =: V^*(x), \quad x \in \mathbb{X}.$$

A la función obtenida se le conoce como función de valor óptimo, mientras que la política π^* es llamada política óptima.

Condiciones. A continuación introducimos las condiciones que asumiremos sobre el modelo $\mathcal{M}$ en lo que resta de este trabajo.

Hipótesis 2.3 *(a) Para cada $x \in \mathbb{X}$, $A(x)$ es un conjunto finito.*

(b) Existe una constante M tal que

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

No es difícil demostrar que la parte (b) en la Hipótesis 2.3 implica que el índice $V_\alpha(\pi, x)$ está acotado (véase: [5], p.10). Este hecho nos permitirá establecer los principales resultados de optimalidad descontada apoyándonos en la teoría de funciones sobre espacios lineales normados. Para lo anterior, denotemos por $B(\mathbb{X})$ al espacio lineal normado de todas las funciones acotadas $v : \mathbb{X} \to \mathbb{R}$ con norma

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

Observaciones 2.1 *Como consecuencia directa de la Hipótesis 2.3 (b) se tiene que:*

a) $B(\mathbb{X})$ es una espacio de Banach;

b) $V^* \in B(\mathbb{X})$.

Ecuación de Optimalidad. Introducimos ahora un elemento que es la clave para caracterizar y obtener políticas óptimas.

Definición 2.4 *Diremos que una función V es solución de la ecuación de optimalidad α-descontada (α-EO) siempre que*

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

donde $\theta(k) := P[\xi = k]$.

A continuación incluimos un resultado bien conocido (véase: [3] y [5]), el cual garantiza que la función de valor óptimo V^* resuelve la α-EO, y a su vez, la existencia de una política estacionaria tal que optimiza dicha ecuación.

Teorema 2.5 *(a) La función V^* es la única función acotada que satisface la α-EO, es decir*

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

(b) Existe $f \in \mathbb{F}$ tal que minimiza la ecuación anterior, esto es,

$$V(x) = c(x,f) + \alpha \sum_{k\in\mathbb{S}} V\left[F(x,f,k)\right]\theta(k), \quad x \in \mathbb{X},$$

Y además, la política $\pi = \{f\}$ es óptima.

3 Construcción de políticas adaptadas

Estimación y control. Como establecimos inicialmente, en este trabajo consideramos a θ desconocida, de manera que el controlador deberá combinar métodos de estimación estadística (distribución empírica) con técnicas de optimización adecuadas, con el propósito de construir una *política adaptada*, que es como se le conoce a la política obtenida de la citada combinación. Específicamente, para construir esa clase de políticas definimos (véase: [3], [4] y [6]) la sucesión de funciones $\{V_t\}_{t=0}^{\infty}$ en $B(\mathbb{X})$ como sigue: $V_0 := 0$, mientras que para cada $t \in \mathbb{N}$, mediante la ecuación recursiva dada por

$$V_t(x) := \min_{a\in A(x)} \left\{ c(x,a) + \alpha \sum_{k\in\mathbb{S}} V_{t-1}\left[F(x,a,k)\right]\theta_t(k) \right\}, \quad x \in \mathbb{X}.$$

De lo cual, no es difícil demostrar (véase: [4] y [6]) que bajo la Hipótesis 2.3,

$$\|V_t - V^*\| \to 0 \quad a.s. \;\; cuando \quad t \to \infty,$$

donde *a.s.* significa convergencia casi segura respecto a la medida de probabilidad P; y además que, para cada $t \in \mathbb{N}$ existe $f_t = f_t^{\theta_t} \in \mathbb{F}$ tal que minimiza el lado derecho de la espresión recursiva previamente introducida, es decir,

$$V_t(x) = c(x, f_t) + \alpha \sum_{k \in \mathbb{S}} V_{t-1}\left[F(x, f_t, k)\right] \theta_t(k), \quad x \in \mathbb{X}.$$

Definimos ahora la política adaptada $\hat{\pi} := \{\hat{\pi}_t\}$ como

$$\hat{\pi}_t(h_t) = \hat{\pi}_t(h_t; \theta_t) := f_t(x_t) \;\; t \in \mathbb{N},$$

y $\hat{\pi}_0$ alguna acción fija. No obstante, respecto a la optimalidad de la política $\hat{\pi}$, nótese que de la Definición 2.2 se deduce que el costo total esperado α-descontado depende fuertemente de las acciones seleccionadas durante las primeras etapas, que es cuando la información respecto a la distribución θ es aún deficiente para el estimador, razón por la cual no es posible en general obtener una política óptima, y en tales circunstancias estudiaremos el concepto de optimalidad de una política dada en un sentido asintótico de acuerdo a lo siguiente.

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

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

Obsérvese que, por el Teorema 2.5 (a) junto con la definición de V^*, de la ecuación previa se tiene que la α-EO es equivalente a la expresión

$$\min_{a \in A(x)} \Phi(x, a) = 0,$$

de donde se obtiene que, en particular, una política $\pi = \{f^*\}$ es α-óptima si, y solo si

$$\Phi(x, f) = 0 \;\; \forall x \in \mathbb{X}.$$

En consecuencia, podemos definir la optimalidad asintótica de la siguiente manera.

Definición 3.1 *Diremos que una política $\pi \in \Pi$ es* ***asintóticamente óptima descontada*** *para el modelo $\mathcal{M}$ si para cada $x \in \mathbb{X}$,*

$$E_x^{\pi}\left[\Phi(x_t, a_t)\right] \to 0 \;\; cuando \;\; t \to \infty.$$

4 Resultado principal

Ahora introduciremos el teorema que contiene el resultado principal de este trabajo.

Teorema 4.1 *Bajo la Hipótesis 2.3, la política $\hat{\pi} = \{f_t\}$ es asintóticamente óptima descontada para el modelo $\mathcal{M}$.*

El esquema de la demostración de este resultado es como sigue.

- Primero se demuestra que la Hipótesis 2.3(a) implica que para cada $x \in \mathbb{X}$ y $\pi \in \Pi$ se satisface

$$\sup_{(x,a)\in\mathbb{K}} |\Phi(x,a) - \Phi_t(x,a)| \to 0 \quad P-a.s \text{ cuando } t \to \infty,$$

donde para cada $t \in \mathbb{N}$ y $(x,a) \in \mathbb{K}$:

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

- Además, se usa el hecho de que la familia

$$\mathcal{V} := \{V^*\left[F\left(x,a,\cdot\right)\right] : (x,a) \in \mathbb{K}\}$$

de funciones

$$V^* : \mathbb{S} \to \mathbb{R}$$

es uniformemente acotada debido a la Hipótesis 2.3 (b) y, como además $\mathbb{S}$ es numerable, entonces $\mathcal{V}$ es una clase Glivencko-Cantelli (véase: [1]), es decir,

$$\eta_t \to 0 \quad P-a.s \text{ cuando } t \to \infty,$$

donde

$$\eta_t \ := \ = \sup_{(x,a)\in\mathbb{K}} \left| \sum_{k\in\mathbb{S}} V^*\left[F\left(x,a,k\right)\right]\theta_{t-1}\left(k\right) - \sum_{k\in\mathbb{S}} V^*\left[F\left(x,a,k\right)\right]\theta\left(k\right) \right|.$$

Referencias

[1] Billingsley P. *Convergence of Probability Measures*; 1a. edic., Editorial John Wiley & Sons, Inc.; New York, U.S.A. (1968).

[2] Dynkin E.B., Yushkevich A.A. *Controlled Markov Processes*; Springer-Verlag, New York. (1979).

[3] Hernández-Lerma O. *Adaptive Markov Control Processes*; 2a. edic., Vol. 79. Editorial Springer-Verlag; New York, U.S.A. (1989).

[4] Hilgert N.,Minjárez-Sosa J.A. *Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria*; Math. Meth. Oper. Res., 63, 3:443-460 (2006).

[5] Leyva-Domínguez J.L. *Estimación empírica en modelos de control markovianos descontados*; Tesis de Licenciatura (Lic. Matem.), Universidad de Sonora, Departamento de Matemáticas. (2013).

[6] Minjárez-Sosa J.A. *Approximation and estimation in Markov control processes under discounted criterion*; Kybernetika, 6, 40:681-690. (2004).