UNIVERSIDAD DE SONORA

DIVISIÓN DE CIENCIAS EXACTAS Y NATURALES

DEPARTAMENTO DE MATEMÁTICAS

Programa de Posgrado en Matemáticas

Modelos de control estocástico con factor de descuento no constante

T E S I S

Que para obtener el título de:

Maestro en Ciencias (Matemáticas)

Presenta:

Jessica Liliana Leyva Domínguez

Director de tesis:
Dr. Jesús Adolfo Minjárez Sosa

Hermosillo, Sonora, México, Diciembre 2015

SINODALES

Dr. Fernando Luque Vásquez
Universidad de Sonora, Hermosillo, México

Dr. Evgueni Illich Gordienko
Universidad Autónoma Metropolitana, Del. Iztapalapa DF, México

Dra. Luz del Carmen Rosas Rosas
Universidad de Sonora, Hermosillo, México

Dr. Jesús Adolfo Minjárez Sosa
Universidad de Sonora, Hermosillo, México

Contenido

Introducción

Dentro de los criterios de optimalidad más analizados en el área de control estocástico se encuentra el criterio de costo descontado. Este criterio ha sido ampliamente estudiado bajo diferentes enfoques: programación dinámica ([3], [13], [14], [15], [24]), control adaptado ([12], [18]), control minimax [11]. Ver también ([23], [16], [20]) para otros enfoques.

En todos estos casos, generalmente el factor de descuento se supone constante durante la evolución del proceso, lo cual simplifica el análisis matemático. Sin embargo, desde el punto de vista de las aplicaciones, esta hipótesis podría resultar demasiado fuerte o poco realista. En efecto, en modelos de economía y finanzas (ver, por ejemplo [5], [26]), el factor de descuento regularmente es una función de la tasa de interés la cual a su vez puede depender, ya sea de la cantidad de capital, o bien de las acciones que toman ciertos inversionistas, y más aún, de un ruido aleatorio.

El objetivo del presente trabajo es estudiar Procesos de Control de Markov (PCM's) bajo el criterio de costo descontado donde el factor de descuento es no constante. Específicamente, consideramos PCM's con factor de descuento de la forma $\tilde{\alpha}(x_n, a_n, \xi_{n+1})$ donde x_n y a_n representan el estado y la acción al tiempo n, respectivamente, y $\{\xi_n\}$ es una sucesión de variables aleatorias que representan una perturbación aleatoria al tiempo n.

Este tipo de factor de descuento juega el siguiente papel en la evolución de sistema. En el estado inicial x_0, el controlador elige una acción a_0. Entonces se genera un costo $c(x_0, a_0)$ y el sistema se mueve a un estado x_1 de acuerdo a una ley de transición. Luego aparece una perturbación aleatoria ξ_1. Una vez que el sistema está en el estado x_1 el controlador selecciona una acción a_1 y se incurre

en un costo $\tilde{\alpha}(x_0, a_0, \xi_1)c(x_1, a_1)$. Después el sistema avanza a un nuevo estado x_2 y el proceso se repite. En general, considerando la historia de estados, controles y perturbaciones aleatorias, al tiempo $n \geq 1$, los costos toman la forma

$$\tilde{\alpha}(x_0, a_0, \xi_1) \cdots \tilde{\alpha}(x_{n-1}, a_{n-1}, \xi_n)c(x_n, a_n). \tag{1}$$

Es decir, los costos son descontados de acuerdo a un factor de descuento multiplicativo. El objetivo de este trabajo es estudiar el problema de control óptimo bajo el índice de funcionamiento definido mediante la acumulación de los costos (1). Es decir, encontrar una política óptima que minimice el índice correspondiente.

Para ese fin, supondremos que los espacios de estados y de control son numerables, y que el costo por etapa c es posiblemente no acotado.

Adicionalmente al problema de control óptimo anteriormente descrito, aquí estudiaremos el problema cuando las variables aleatorias $\{\xi_n\}$ son independientes e idénticamente distribuidas con distribución desconocida θ. En este caso implementaremos procedimientos de estimación estadística y control con el fin de definir las políticas de control. En particular, usaremos la distribución empírica para estimar θ. A la política resultante de un proceso de estimación y control se le llama política adaptada, y aquí estudiaremos su optimalidad en un sentido asintótico.

Aunque no es muy común, existen importantes trabajos que tratan PCM's con factor de descuento no constante. Por ejemplo, Hinderer en [19], Schäl en [25] estudiaron el problema con factor de descuento de la forma $\alpha(x, a)$ considerando costos acotados inferiormente. Estos trabajos se extendieron en [21] para costo no acotado, en tanto que, en [27] se considera el factor de descuento de la forma $\alpha(x)$ con costo no acotado.

Por otro lado, en [6], [7], [8], [9], [10] han sido analizados criterios de optimalidad con factor de descuento aleatorio considerando varios enfoques.

A diferencia de esta tesis, todas estas referencias consideran espacio de estados y de control generales. Sin embargo, nuestro trabajo podría considerarse

una combinación de ellos, al considerar un factor de descuento que depende del estado, control y ruido aleatorio.

El trabajo está estructurado de la siguiente manera.

En el Capítulo 1 se introducen los PCM's considerando un factor de descuento de la forma $\tilde{\alpha}(x,a)$. Además, se presenta una descripción del modelo de control correspondiente, así como del problema de control óptimo.

En el Capítulo 2 se analiza la existencia de políticas óptimas, y se presentan algoritmos iterativos de aproximación a la función de valor óptimo considerando un factor de descuento (x,a)-dependiente.

Finalmente en el Capítulo 3 analizamos el caso general cuando el factor de descuento es de la forma $\alpha(x,a,\xi)$, donde las variables aleatorias ξ tienen distribución desconocida.

Concluimos el trabajo con un Apéndice donde se incluyen algunos resultados que usaremos a lo largo del trabajo.

Capítulo 1

Procesos de control de Markov con costos descontados

1.1 Introducción

El propósito de este capítulo es introducir el problema de control óptimo (PCO) con respecto al criterio de optimalidad de costo con factor de descuento (x,a)-dependiente. Para la formulación del PCO es necesario describir tres elementos: un modelo de control, un conjunto de políticas admisibles y un índice de funcionamiento que mide el comportamiento del sistema cuando se usan diferentes políticas.

1.2 Modelo de control

1.2.1 Descripción del modelo

Definición 1.1 *Un modelo de control markoviano (MCM) con factor de descuento que depende del estado y control es un arreglo de la forma*

$$\mathcal{M} = (\mathbb{X}, \mathbb{A}, \{A(x) : x \in \mathbb{X}\}, P, \alpha, c) \tag{1.1}$$

donde

- $\mathbb{X}$ *representa el espacio de estados.*
- $\mathbb{A}$ *representa el espacio de controles o acciones.*

En este trabajo supondremos que $\mathbb{X}$ y $\mathbb{A}$ son conjuntos numerables.

- *Para cada $x \in \mathbb{X}$, $A(x) \subset \mathbb{A}$ es un conjunto no vacío, cuyos elementos representan las acciones admisibles cuando el sistema se encuentra en el estado $x \in \mathbb{X}$.*

- *P representa la ley de transición entre los estados, es decir, P es una probabilidad condicional definida en el espacio de estados $\mathbb{X}$. Reescribir bien*

Sea $\mathbb{K} := \{(x,a) : x \in \mathbb{X}, a \in A(x)\}$ el conjunto de pares estado-acción admisibles, el cual es un subconjunto del espacio $\mathbb{X} \times \mathbb{A}$.

- *$\alpha : \mathbb{K} \to (0,1)$ es una función que representa el factor de descuento,*

- *$c : \mathbb{K} \to \mathbb{R}$ es una función no negativa que representa el costo por etapa.*

1.2.2 Interpretación

El modelo $\mathcal{M}$ representa un sistema dinámico que evoluciona de la siguiente manera. En el estado inicial $x_0 \in \mathbb{X}$, el controlador selecciona una acción o control $a_0 \in \mathbb{A}(x_0)$ y se genera un costo $c(x_0, a_0)$. Entonces el sistema se mueve a un nuevo estado $x_1 \in \mathbb{X}$ de acuerdo a la ley de transición

$$P[x_1 = y \mid x_0, a_0] := P_{x_0,y}(a_0)$$

Una vez que el sistema se encuentra en el estado x_1 el controlador elige una acción $a_1 \in A(x_1)$ y se genera un costo descontado $\alpha(x_0,a_0)c(x_1,a_1)$. Después el sistema se mueve a un estado x_2, y el proceso se repite. En general, en la etapa $n \geq 1$, cuando el sistema se encuentra en el estado $x_n \in \mathbb{X}$ y se elige un control $a_n \in A(x_n)$, sucede lo siguiente: 1) se genera un costo descontado de la forma

$$\alpha(x_0,a_0)\ldots\alpha(x_{n-1},a_{n-1})c(x_n,a_n); \tag{1.2}$$

2) el sistema avanza a un nuevo estado $x_{n+1} \in \mathbb{X}$ de acuerdo a la ley de transición

$$P[x_{n+1} = y \mid x_n, a_n] := P_{x_n,y}(a_n);$$

y el proceso se repite.

1.3 Políticas de control admisibles

Definición 1.2 *Dado un MCM definimos el espacio de historias admisibles hasta la n-ésima etapa como:*

$$\mathbb{H}_0 := \mathbb{X}$$
$$\mathbb{H}_n := \mathbb{K}^n \times \mathbb{X}, \;\; n \in \mathbb{N}.$$

Entonces, un elemento de $\mathbb{H}_n$ es un vector (o historia) de la forma

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

con $(x_k, a_k) \in \mathbb{K}$ para $k = 0, 1, \ldots, n-1$ y $x_n \in \mathbb{X}$.

Una regla de decisión es un procedimiento para elegir un control (acción) en cada etapa, el cual puede depender, ya sea, de la historia o bien únicamente del estado actual del sistema.

Además las reglas de decisión pueden ser deterministas o aleatorizadas.

Específicamente, una regla de decisión determinista dependiente de la historia, es una función $f_n : \mathbb{H}_n \to \mathbb{A}$ tal que $f_n(h_n) \in A(x_n)$. Mientras que, si f_n depende de h_n sólamente a través de x_n, diremos que f_n es una regla de decisión markoviana, y en cuyo caso podemos decir que una regla de este tipo es una función $f_n : \mathbb{X} \to \mathbb{A}$ tal que $f_n(x) \in A(x)$.

Por otro lado, una regla de decisión aleatorizada π_n determina una distribución de probabilidad sobre el conjunto de acciones $\mathbb{A}$. En particular, una regla de decisión aleatorizada markoviana es una función $\pi_n : \mathbb{X} \to \mathbb{P}(\mathbb{A})$; y es dependiente de la historia si $\pi_n : \mathbb{H}_n \to \mathbb{P}(\mathbb{A})$. Es decir, en este caso cada acción se elige de acuerdo a una distribución de probabilidad:

$$a_n \sim \pi_n(\cdot \mid x_n)$$

para el caso markoviano, y

$$a_n \sim \pi_n(\cdot \mid h_n)$$

para el caso en que dependa de la historia.

Una regla de decisión determinista puede ser considerada como una regla de decisión aleatorizada en la cual la correspondiente distribución de probabilidades esta concentrada en algún $a \in A(x)$, es decir, $\pi_n(a \mid x_n) = 1$ para $a = f_n(x_n)$ o $\pi_n(a \mid h_n) = 1$ para $a = f_n(h_n)$.

Definición 1.3 *Una política de control (o simplemente política) es una sucesión de reglas de decisión* $\pi = \{\pi_0, \pi_1, \ldots, \pi_N\}$, $N \leq \infty$.

Denotaremos por Π al conjunto de todas las políticas.

Dependiendo de la clase de reglas de decisión, las políticas se clasifican como aleatorizadas o deterministas, y estas a su vez como markovianas o dependientes de la historia. Por ejemplo, introduciendo el conjunto

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

entonces, una política determinista markoviana toma la forma $\pi = \{f_n\}$, donde $f_n \in \mathbb{F}$. Cuando para alguna función $f \in \mathbb{F}$, $f_n = f$, $\forall n \in \mathbb{N}_0$, entonces diremos que π es estacionaria. En este caso se tiene que $\pi = \{f, f, \ldots\}$, así que denotaremos por f a la política π.

1.4 Procesos de control de Markov

Sea $\mathcal{M}$ un modelo de control. Consideremos el espacio muestral $\Omega := (\mathbb{X} \times \mathbb{A})^\infty = \mathbb{X} \times \mathbb{A} \times \mathbb{X} \times \mathbb{A} \times \ldots$. Un elemento genérico $\omega \in \Omega$, representa una trayectoria de la forma

$$\omega = (x_0, a_0, x_1, a_1, \ldots)$$

donde $x_n \in \mathbb{X}$ y $a_n \in \mathbb{A}$, $n = 0, 1, \ldots$.

Para cada $\pi \in \Pi$ y cada estado inicial $x \in \mathbb{X}$ existe una probabilidad P_x^π definida en los subconjuntos de Ω tal que

$$\begin{aligned}
&P_x^\pi [x_0 = x] = 1 \\
&P_x^\pi [a_n = a \mid h_n] = \pi_n(a \mid h_n), \;\; a \in \mathbb{A} \\
&P_x^\pi [x_{n+1} = y \mid h_n, a_n] = P_{x_n,y}(a_n), \;\; y \in \mathbb{X}.
\end{aligned}$$

Al proceso estocástico $(\Omega, P_x^{\pi}, \{x_t\})$ se le conoce como proceso de control de Markov o proceso de decisión de Markov a tiempo discreto.

Observación 1.4 *Si $v : \mathbb{X} \to \mathbb{R}$ es una función de x_{t+1}, entonces*

$$E_x^{\pi}[v(x_{t+1}) \mid h_t, a_t] = \sum_{y \in \mathbb{X}} v(y) P_{x_t,y}(a_t).$$

1.5 Problema de control óptimo

En general, un índice de funcionamiento (o criterio de optimalidad) consiste en una función que de alguna manera, mide el comportamiento del sistema al utilizar diferentes políticas de control dado el estado inicial.

Con el fin de medir el rendimiento de las políticas de control, usaremos como índice de funcionamiento el costo total esperado descontado con factor de descuento que depende del estado-acción.

Sea

$$\Gamma_n := \prod_{k=0}^{n-1} \alpha(x_k, a_k), \quad \text{si } n \in \mathbb{N}$$
$$\Gamma_0 := 1.$$

Entonces, de acuerdo a la relación (1.2), se tiene la siguiente definición.

Definición 1.5 *Sean $\pi \in \Pi$ y $x_0 = x \in \mathbb{X}$. Se define el costo total esperado descontado con factor de descuento (x,a)-dependiente como:*

$$V(x, \pi) := E_x^{\pi}\left[\sum_{n=0}^{\infty} \Gamma_n c(x_n, a_n)\right]. \tag{1.3}$$

Entonces, el problema de control óptimo (PCO) asociado con el modelo de control $\mathcal{M}$ consiste en encontrar una política $\pi^* \in \Pi$ tal que $V(x, \pi^*) = V^*(x)$ para toda $x \in \mathbb{X}$, donde

$$V^*(x) := \inf_{\pi \in \Pi} V(x, \pi), \quad x \in \mathbb{X}, \tag{1.4}$$

es la función de valor óptimo. En este caso π^* es una política óptima.

Capítulo 2

Existencia de Políticas Óptimas

2.1 Introducción

En este capítulo analizaremos el PCO asociado al MCM (1.1) bajo el índice de funcionamiento con factor de descuento (x,a)-dependiente. Además, introduciremos un conjunto de hipótesis que garantizará la existencia de políticas óptimas estacionarias.

Finalmente presentamos los algoritmos de iteración de valores e iteración de políticas para aproximar a la función de valor óptimo.

2.2 Criterio de optimalidad de costo descontado

Recordemos de la Definición 1.5, que dados $x \in \mathbb{X}$ y $\pi \in \Pi$,

$$V(x,\pi) := E_x^{\pi}\left[\sum_{n=0}^{\infty} \Gamma_n c(x_n,a_n)\right]$$

define el costo total esperado descontado con factor de descuento (x,a)-dependiente cuando se usa la política π dado el estado inicial $x_0 = x$. En cuyo caso, obsérvese que usando (1.4) el PCO respectivo consiste en encontrar una política $\pi^* \in \Pi$ que

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

2.3 Condiciones

En esta sección introduciremos condiciones de acotamiento y compacidad sobre el modelo de control, las cuales garantizarán la existencia de políticas óptimas.

Hipótesis 2.1 *(a) Existe una función $W : \mathbb{X} \to \mathbb{R}$ y una constante positiva L tal que*

$$|c(x,a)| \leq LW(x) \quad \forall\, (x,a) \in \mathbb{K}. \tag{2.1}$$

(b) La función α es uniformemente acotada,

$$\alpha^* := \sup_{(x,a)\in\mathbb{K}} \alpha(x,a) < 1.$$

(c) Existe una constante β tal que $1 \leq \beta \leq \frac{1}{\alpha^*}$, y

$$\sum_{y\in\mathbb{X}} W(y)P_{x,y}(a) \leq \beta W(x), \quad \forall\, (x,a) \in \mathbb{K}. \tag{2.2}$$

(d) Para cada $x \in \mathbb{X}$, $A(x)$ es un conjunto finito.

Observación 2.2 *Denotemos como $B_W(\mathbb{X})$ al espacio de todas las funciones $u : \mathbb{X} \to \mathbb{R}$ con W-norma finita definida como sigue:*

$$\|u\|_W := \sup_{x\in\mathbb{X}} \frac{|u(x)|}{W(x)} < \infty. \tag{2.3}$$

2.4 Ecuación de optimalidad

A continuación introduciremos un elemento, el cual es clave para caracterizar y obtener políticas óptimas.

Definición 2.3 *Diremos que una función $u : \mathbb{X} \to \mathbb{R}$ es una solución de la ecuación de optimalidad (EO) con factor de descuento (x,a)-dependiente si*

$$u(x) = \min_{a\in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y\in X} u(y)P_{x,y}(a) \right\} \quad \forall x \in \mathbb{X}. \tag{2.4}$$

Nuestro objetivo es demostrar que la función de valor óptimo V^* es una solución de la EO, y partiendo de aquí demostrar la existencia de políticas óptimas. Para tal fin definamos los siguientes operadores.

Para cada función $u \in \mathbb{X}$ y $(x,a) \in \mathbb{K}$,

$$Tu(x) := \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \right\} \quad \forall x \in \mathbb{X}, \tag{2.5}$$

y para cada $f \in \mathbb{F}$,

$$T_f u(x) := c(x,f) + \alpha(x,f) \sum_{y \in \mathbb{X}} u(y) P_{x,y}(f) \quad x \in \mathbb{X}. \tag{2.6}$$

Bajo la Hipótesis 2.1, para cada $Tu \in B_W(\mathbb{X})$, $u \in B_W(\mathbb{X})$ y además, este operador es de contracción, tal como lo establece el siguiente resultado.

Lema 2.4 *Bajo la Hipótesis 2.1, T es un operador de contracción módulo $\alpha^*\beta$ sobre $B_W(\mathbb{X})$, ésto es, para cada par de funciones $u, v \in B_W(\mathbb{X})$:*

$$||Tu - Tv||_W \leq \alpha^* \beta \, \|u - v\|_W .$$

Demostración. Primero mostraremos que $Tu \in B_W(\mathbb{X})$ para cada $u \in B_W(\mathbb{X})$. Para esto, nótese que por (2.5), para cada función $u \in B_W(\mathbb{X})$ y $x \in \mathbb{X}$,

$$Tu(x) = \inf_{a \in A(x)} v(x,a),$$

donde

$$v(x,a) := c(x,a) + \alpha(x,a) \sum_{y \in X} u(y) P_{x,y}(a), \quad a \in A(x).$$

Por otra parte, por la Hipótesis 2.1 se tiene

$$\begin{aligned}
|v(x,a)| &= \left| c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \right| \\
&\leq |c(x,a)| + \alpha^* \sum_{y \in \mathbb{X}} |u(y)| P_{x,y}(a) \\
&\leq LW(x) + \alpha^* \sum_{y \in \mathbb{X}} \|u\|_W W(y) P_{x,y}(a) \\
&= LW(x) + \alpha^* \|u\|_W \sum_{y \in \mathbb{X}} W(y) P_{x,y}(a) \\
&\leq LW(x) + \alpha^* \beta \, \|u\|_W W(x) = MW(x),
\end{aligned}$$

donde $M := L + \alpha^* \beta \, \|u\|_W$. Por lo tanto, $Tu \in B_W(\mathbb{X})$.

Ahora para cada $u, v \in B_W(\mathbb{X})$ y $x \in \mathbb{X}$ se cumple

$$\begin{aligned}
|Tu(x) - Tv(x)| &= \left| \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \right\} \right. \\
&\quad \left. - \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \right\} \right| \\
&\leq \max_{a \in A(x)} \left\{ \alpha(x,a) \sum_{y \in \mathbb{X}} |u(y) - v(y)| P_{x,y}(a) \right\} \\
&\leq \alpha^* \max_{a \in A(x)} \left\{ \sum_{y \in \mathbb{X}} |u(y) - v(y)| P_{x,y}(a) \right\} \\
&\leq \alpha^* \max_{a \in A(x)} \sum_{y \in \mathbb{X}} ||u - v||_W W(y) P_{x,y}(a).
\end{aligned}$$

Es decir

$$|Tu(x) - Tv(x)| \leq \alpha^* \beta ||u - v||_W W(x),$$

de donde

$$||Tu - Tv||_W := \sup_{x \in \mathbb{X}} \frac{|Tu(x) - Tv(x)|}{W(x)} \leq \alpha^* \beta \, \|u - v\|_W .$$

Por lo tanto, como $\alpha^* \beta < 1$ tenemos que $T : B_W(\mathbb{X}) \to B_W(\mathbb{X})$ es un operador de contracción con módulo $\alpha^* \beta$.

■

Una consecuencia del Lemma 2.4 y el teorema de punto fijo de Banach es el siguiente resultado.

Proposición 2.5 *Bajo la Hipótesis 2.1, existe una función $\tilde{u} \in B_W(\mathbb{X})$ tal que para todo $x \in \mathbb{X}$.*

$$\tilde{u}(x) = T\tilde{u}(x) \tag{2.7}$$

y

$$||T^n u - \tilde{u}||_W \leq (\alpha^* \beta)^n ||u - \tilde{u}||_W, \quad \forall u \in B_W(\mathbb{X}), \; n \in \mathbb{N}_0. \tag{2.8}$$

Definamos ahora la sucesión de funciones de iteración de valores $\{v_n\} \in B_W(\mathbb{X})$ como $v_0 = 0$,

$$v_n(x) = Tv_{n-1}(x) = T^n v_0, \;\; n \in \mathbb{N}, x \in \mathbb{X}. \tag{2.9}$$

Observe que de (2.8) y (2.9), tomando $u = v_0$ obtenemos

$$||v_n - \tilde{u}||_W \leq (\alpha^* \beta)^n ||\tilde{u}||_W, \;\; n \in \mathbb{N}_0. \tag{2.10}$$

Teorema 2.6 *Si la Hipótesis 2.1 se cumple, entonces:*

(a) La función de valor óptimo V^* es la única solución en $B_W(\mathbb{X})$ de la EO, es decir,

$$V^*(x) = TV^*(x), \;\; x \in \mathbb{X}. \tag{2.11}$$

(b) Para cada $n \in \mathbb{N}$,

$$||v_n - V^*||_W \leq \frac{L(\alpha^* \beta)^n}{1 - \alpha^* \beta}.$$

(c) Existe $f^* \in \mathbb{F}$ tal que

$$V^*(x) = T_{f^*}V^*(x) = c(x, f^*) + \alpha(x, f^*) \sum_{y \in \mathbb{X}} V^*(y) P_{x,y}(f^*), \;\; x \in \mathbb{X}.$$

y además, f^* es una política óptima, ésto es

$$V^*(x) = V(x, f^*), \;\; x \in \mathbb{X}.$$

Demostración. Nótese que si se demuestra que $\tilde{u} = V^*$, entonces quedará comprobada la validez de las partes (a) y (b) como consecuencia de las expresiones (2.7) y (2.8). Así que, para tal fin es suficiente demostrar que se satisfacen $\tilde{u} \geq V^*$ y $\tilde{u} \leq V^*$.

(i) $\tilde{u} \geq V^*$.

Sea $\tilde{f} \in \mathbb{F}$ tal que para toda $x \in \mathbb{X}$

$$\tilde{u}(x) = c(x, \tilde{f}) + \alpha(x, \tilde{f}) \sum_{y \in \mathbb{X}} \tilde{u}(y) P_{x,y}(\tilde{f}). \tag{2.12}$$

Iterando (2.12) se obtiene

$$\tilde{u}(x) = E_x^{\tilde{f}}\left[\sum_{n=0}^{m-1}\Gamma_n c(x_n,\tilde{f})\right] + E_x^{\tilde{f}}\Gamma_m\tilde{u}(x_m). \tag{2.13}$$

Por otra parte, para toda $\pi \in \Pi$, $x \in \mathbb{X}$ se tiene

$$E_x^{\pi}W(x_n) \leq \beta^n W(x), \ \ \forall n \in \mathbb{N}_0. \tag{2.14}$$

Este hecho se demuestra por inducción de la siguiente manera. Primero, nótese que para $n = 0$, es claro que

$$E_x^{\pi}W(x_0) \leq \beta^0 W(x) = W(x).$$

Ahora supongamos que (2.14) es válido para algún $n \geq 1$. Entonces, por la Hipótesis 2.1(c) se tiene

$$\begin{aligned} E_x^{\pi}\left[W(x_{n+1}) \mid h_n, a_n\right] &= \sum_{y\in\mathbb{X}} W(y)P_{x_n,y}(a_n) \\ &\leq \beta W(x_n), \end{aligned}$$

luego, tomando esperanza en ambos lados de esta desigualdad, y aplicando la hipótesis de inducción se obtiene

$$E_x^{\pi}\left[W(x_{n+1})\right] \leq \beta E_x^{\pi}[W(x_n)] \leq \beta^{n+1}W(x),$$

lo cual implica (2.14).

Ahora, por la Hipótesis 2.1(a)-(c) se tiene lo siguiente

$$\begin{aligned} 0 &\leq \left|E_x^{\tilde{f}}\Gamma_m\tilde{u}(x_m)\right| \\ &= \left|E_x^{\tilde{f}}\prod_{k=0}^{m-1}\alpha(x_k,\tilde{f})\tilde{u}(x_m)\right| \\ &\leq \left|E_x^{\tilde{f}}\prod_{k=0}^{m-1}\alpha^*\tilde{u}(x_m)\right| \\ &\leq E_x^{\tilde{f}}\left[(\alpha^*)^m|\tilde{u}(x_m)|\right] \\ &\leq (\alpha^*)^m||\tilde{u}||_W E_x^{\tilde{f}}\left[W(x_m)\right] \\ &\leq (\beta\alpha^*)^m||\tilde{u}||_W W(x). \end{aligned} \tag{2.15}$$

Por lo tanto, tomando límite cuando $m \to \infty$ en (2.13), se sigue

$$\begin{aligned}\tilde{u}(x) &= \lim_{m\to\infty}\left[E_x^{\tilde{f}}\left(\sum_{n=0}^{m-1}\Gamma_n c(x_n,\tilde{f})\right)+E_x^{\tilde{f}}\Gamma_m\tilde{u}(x_m)\right]\\ &= E_x^{\tilde{f}}\left[\sum_{n=0}^{\infty}\Gamma_n c(x_n,\tilde{f})\right]\\ &= V(x,\tilde{f}), \quad \forall\, x\in\mathbb{X}. \end{aligned} \tag{2.16}$$

Entonces, por definición de función de valor,

$$\tilde{u}(x) = V(x,\tilde{f}) \geq V^*(x) = \inf_{\pi\in\Pi} V(x,\pi),$$

ésto es,

$$\tilde{u}(x) \geq V^*(x). \tag{2.17}$$

(ii) $\tilde{u} \leq V^*$.

Puesto que $\tilde{u} = T\tilde{u}$, entonces para toda $(x,a)\in\mathbb{K}$ se tiene

$$\tilde{u}(x) \leq c(x,a) + \alpha(x,a)\sum_{y\in\mathbb{X}}\tilde{u}(y)P_{x,y}(a). \tag{2.18}$$

Por la Observación 1.4(b), para cada $\pi\in\Pi$, y cada $x\in\mathbb{X}$,

$$E_x^{\pi}\left[c(x_n,a_n)+\alpha(x_n,a_n)\tilde{u}(x_{n+1})-\tilde{u}(x_n)\mid h_n,a_n\right]\geq 0, \ \ \forall n\geq 0, \tag{2.19}$$

lo cual implica que, para cada $n\geq 1$

$$E_x^{\pi}\left[\prod_{i=0}^{n-1}\alpha(x_i,a_i)c(x_n,a_n)+\prod_{i=0}^{n}\alpha(x_i,a_i)\tilde{u}(x_{n+1})-\prod_{i=0}^{n-1}\alpha(x_i,a_i)\tilde{u}(x_n)\mid h_n,a_n\right]\geq 0. \tag{2.20}$$

Ahora, tomando esperanza E_x^{π} en (2.20), tenemos

$$E_x^{\pi}\left[\prod_{i=0}^{n-1}\alpha(x_i,a_i)c(x_n,a_n)+\prod_{i=0}^{n}\alpha(x_i,a_i)\tilde{u}(x_{n+1})-\prod_{i=0}^{n-1}\alpha(x_i,a_i)\tilde{u}(x_n)\right]\geq 0, \ \ \forall n\geq 1. \tag{2.21}$$

Luego, sumando desde $n = 0$ a $n = m-1$ en ambos lados de (2.21) y observando que $x_0 = x$, se sigue que

$$E_x^{\pi}\left[\sum_{n=0}^{m-1}\prod_{i=0}^{n-1}\alpha(x_i,a_i)c(x_n,a_n)+\prod_{i=0}^{m-1}\alpha(x_i,a_i)\tilde{u}(x_m)\right]\geq\tilde{u}(x),\ \ \forall m\geq 2, \quad (2.22)$$

y luego, tomando límite cuando $m\to\infty$ en (2.22) y usando (2.15) resulta

$$V(x,\pi)\geq\tilde{u}(x),$$

y, debido a que $\pi\in\Pi$ y $x\in\mathbb{X}$ son arbitrarios se obtiene

$$\tilde{u}(x)\leq V^*(x). \quad (2.23)$$

Por lo tanto, de (2.17) y (2.23) se concluye que

$$\tilde{u}(x)=V^*(x),\ \ \forall x\in\mathbb{X},$$

lo que prueba la validez de las partes (a) y (b).

Finalmente, para demostrar (c), tomamos f^* en lugar de $\tilde{f}$, en la demostración de las partes (a) y (b). Entonces, por (2.16) se obtiene

$$V(x,f^*)=V^*(x),$$

lo cual implica que f^* es óptima.

■

2.5 Algoritmo de iteración de políticas

En el Teorema 2.6 se establece un algoritmo de aproximación para la función de valor óptimo V^* por medio de las funciones de iteración de valor $\{v_n\}$. En este caso la sucesión $\{v_n\}$ resulta ser creciente, converge a V^* y se define en forma recursiva.

Ahora se presentará el algoritmo de iteración de políticas, el cual proporciona una aproximación decreciente a V^* en el conjunto de políticas de control.

Para definir el algoritmo, primero obsérvese que a partir de la propiedad de Markov y aplicando propiedades de esperanza condicional, para cada política estacionaria $f \in \mathbb{F}$ y cada $x \in \mathbb{X}$, el costo correspondiente $V(x,f)$ satisface lo siguiente,

$$\begin{aligned}
V(x,f) &= E_x^f\left[\sum_{n=0}^{\infty}\Gamma_n c(x_n,a_n)\right] \\
&= c(x,f)+\alpha(x,f)E_x^f\left[\sum_{n=1}^{\infty}\prod_{k=0}^{n-1}\alpha(x_k,f)c(x_n,f)\right] \\
&= c(x,f)+\alpha(x,f)\sum_{y\in\mathbb{X}}E^f[c(x_1,f)+\alpha(x_1,f)\sum_{n=2}^{\infty}\prod_{k=0}^{n-1}\alpha(x_k,f) \\
&\quad c(x_n,f)\mid x_1]P_{x,y}(f) \\
&= c(x,f)+\alpha(x,f)\sum_{y\in\mathbb{X}}V(y,f)P_{x,y}(f) \\
&= T_fV(x,f),\ \ x\in\mathbb{X} \qquad (2.24)
\end{aligned}$$

donde T_f es el operador definido en (2.6).

Sea $f_0 \in \mathbb{F}$ una política estacionaria con función de costo

$$v_0(\cdot) := V(\cdot,f_0) \in B_W(\mathbb{X}).$$

Entonces por (2.24)

$$v_0(x) = c(x,f_0)+\alpha(x,f_0)\sum_{y\in\mathbb{X}}v_0(y)P_{x,y}(f_0) = T_{f_0}v_0(x),\ \ x\in\mathbb{X} \qquad (2.25)$$

Ahora, sea $f_1 \in \mathbb{F}$ tal que

$$Tv_0(x) = T_{f_1}v_0(x),\ \ x\in\mathbb{X}, \qquad (2.26)$$

y definamos

$$v_1(\cdot) := V(\cdot,f_1).$$

En general, definamos a la sucesión $\{v_n\} \in B_W(\mathbb{X})$ como sigue. Dado $f_n \in \mathbb{F}$, calculamos

$$v_n(\cdot) := V(\cdot,f_n) \in B_W(\mathbb{X}).$$

Luego, sea $f_{n+1} \in \mathbb{F}$ tal que

$$T_{f_{n+1}} v_n(x) = T v_n(x), \quad x \in \mathbb{X} \tag{2.27}$$

esto es,

$$c(x, f_{n+1}) + \alpha(x, f_{n+1}) \sum_{y \in \mathbb{X}} v_n(y) P_{x,y}(f_{n+1}) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} v_n(y) P_{x,y}(a) \right\}$$

Entonces definimos $v_{n+1}(\cdot) = V(\cdot, f_{n+1})$.

Teorema 2.7 *Bajo la Hipótesis 2.1, $\{v_n\}$ es una sucesión decreciente en $B_W(\mathbb{X})$ tal que $v_n \searrow V^*$.*

Demostración. Nótese que, de (2.24)-(2.27),

$$\begin{aligned} v_0(x) := V(x, f_0) &= c(x, f_0) + \alpha(x, f_0) \sum_{y \in \mathbb{X}} v_0(y) P_{x,y}(f_0) \\ &\geq \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} v_0(y) P_{x,y}(a) \right\} \\ &= T v_0(x) = T_{f_1} v_0(x) \\ &= c(x, f_1) + \alpha(x, f_1) \sum_{y \in \mathbb{X}} v_0(y) P_{x,y}(f_1). \end{aligned}$$

Luego, iterando esta desigualdad obtenida, no es difícil verificar que

$$v_0(x) \geq T v_0(x) \geq V(x, f_1) = v_1(x), \quad x \in \mathbb{X}.$$

En general, obsérvese que para cada $n \in \mathbb{N}$,

$$\begin{aligned} v_n(x) := V(x, f_n) &= c(x, f_n) + \alpha(x, f_n) \sum_{y \in \mathbb{X}} v_n(y) P_{x,y}(f_n) \\ &\geq \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \sum_{y \in \mathbb{X}} v_n(y) P_{x,y}(a) \right\} \\ &= T v_n(x) = T_{f_{n+1}} v_n(x) \\ &= c(x, f_{n+1}) + \alpha(x, f_{n+1}) \sum_{y \in \mathbb{X}} v_0(y) P_{x,y}(f_{n+1}). \end{aligned}$$

Ahora, por iteración de la desigualdad previa se obtiene

$$v_n(x) \geq Tv_n(x) \geq V(x, f_{n+1}) \geq v_{n+1}(x), \quad x \in \mathbb{X}. \tag{2.28}$$

En consecuencia, $\{v_n\}$ es una sucesión decreciente, y por lo tanto, existe una función $v \in B_W(\mathbb{X})$ tal que $v_n \searrow v$.

Ahora, dado que $v_n(\cdot) := V(\cdot, f_n) \geq V^*(\cdot)$, $\forall\ n \in \mathbb{N}$, entonces se tiene

$$v(x) \geq V^*(x), \quad x \in \mathbb{X}. \tag{2.29}$$

Tomando límite cuando $n \to \infty$ en (2.28), y aplicando en Lema 3.3 en [17] concluimos que se cumple la igualdad

$$Tv(x) = v(x). \tag{2.30}$$

Por otro lado, por la Observación 1.4(b), para cada $\pi \in \Pi$ y cada $x \in \mathbb{X}$

$$\begin{aligned}
E_x^{\pi}\left[\Gamma_{n+1} v(x_{n+1}) \mid h_n, a_n\right] &= \Gamma_{n+1} \sum_{y \in \mathbb{X}} v(y) P_{x_n,y}(a_n) \\
&= \Gamma_n [c(x_n, a_n) + \alpha(x_n, a_n) \sum_{y \in \mathbb{X}} v(y) P_{x_n,y}(a_n) \\
&\quad - c(x_n, a_n)] \\
&\geq \Gamma_n \min_{a \in A(x_n)} \left[c(x_n, a) + \alpha(x_n, a) \sum_{y \in \mathbb{X}} v(y) P_{x_n,y}(a) \right] \\
&\quad - \Gamma_n c(x_n, a_n) \\
&= \Gamma_n T v(x_n) - \Gamma_n c(x_n, a_n) \geq \Gamma_n v(x_n) - \Gamma_n c(x_n, a_n).
\end{aligned}$$

Esto implica que,

$$\Gamma_n c(x_n, a_n) \geq E_x^{\pi}\left[\Gamma_n v(x_n) - \Gamma_{n+1} v(x_{n+1}) \mid h_n, a_n\right].$$

Por lo tanto, para toda $k \in \mathbb{N}$,

$$E_x^{\pi} \sum_{n=0}^{k-1} \Gamma_n c(x_n, a_n) \geq v(x) - E_x^{\pi}[\Gamma_k v(x_k)].$$

Haciendo $k \to \infty$, nótese que por (2.15) se obtiene,

$$V(x, \pi) \geq v(x).$$

Dado que $\pi \in \Pi$ es arbitraria se obtiene la desigualdad,

$$V^*(x) \geq \nu(x);$$

lo cual, combinando con (2.29) demuestra

$$V^*(x) = \nu(x).$$

Por consiguiente,se tiene que $\nu_n \searrow V^*$. Esto demuestra el Teorema.

■

Capítulo 3

Modelos de control con factor de descuento aleatorio

3.1 Introducción

En este capítulo estudiaremos modelos de control de Markov con factor de descuento no constante de la forma

$$\tilde{\alpha}(x_n, a_n, \xi_n), \tag{3.1}$$

donde x_n y a_n representan el estado y el control al tiempo n, respectivamente, y $\{\xi_n\}$ es una sucesión de variables aleatorias i.i.d. con distribución θ.

El factor de descuento (3.1) juega el siguiente papel en la evolución del sistema. En el estado inicial $x_0 \in \mathbb{X}$, el controlador elige una acción $a_0 \in A(x_0)$. Entonces se incurre en un costo $c(x_0, a_0)$, y el sistema avanza a un nuevo estado x_1 de acuerdo con una ley de transición, y se presenta una perturbación aleatoria ξ_1. Una vez que el sistema se encuentra en el estado $x_1 \in \mathbb{X}$ el controlador elige una acción $a_1 \in A(x_1)$ y se incurre en un costo descontado $\tilde{\alpha}(x_0, a_0, \xi_1)c(x_1, a_1)$. Luego, el sistema avanza al estado $x_2 \in \mathbb{X}$ y se repite el proceso.

De la descripción previa, dada la historia compuesta del estado-acción y perturbación aleatoria, en cada etapa $n \geq 1$, el controlador incurre en un costo descontado

$$\tilde{\alpha}(x_0, a_0, \xi_1)\tilde{\alpha}(x_1, a_1, \xi_2) \cdots \tilde{\alpha}(x_{n-1}, a_{n-1}, \xi_n)c(x_n, a_n). \tag{3.2}$$

Por lo tanto, los costos son descontados a una tasa de descuento multiplicativa, y suponiendo la posibilidad de costos $c(\cdot,\cdot)$ no acotados, el objetivo es estudiar la optimalidad bajo el índice de funcionamiento definido por la acumulación de estos costos a lo largo de la evolución del sistema.

3.2 Modelo de Control

Consideramos el modelo de control de Markov (MCM)

$$\tilde{\mathfrak{M}} = (\mathbb{X}, \mathbb{A}, \mathbb{S}, \{A(x) : x \in \mathbb{X}\}, P, \tilde{\alpha}, c) \tag{3.3}$$

donde $\mathbb{X}$, $\mathbb{A}$, $\{A(x) : x \in \mathbb{X}\}$, P y c son como en la Definición 1.1, mientras que, $\mathbb{S}$ representa el espacio donde toman valores las perturbaciones aleatorias y $\tilde{\alpha} : \mathbb{K} \times \mathbb{S} \to (0,1)$ es la función del factor de descuento.

Supondremos que $\{\xi_n\}$ es una sucesión de v.a. i.i.d. definidas sobre un espacio de probabilidad $(\Omega_0, \mathcal{F}, P_0)$ con valores en $\mathbb{S}$ y distribución $\theta \in \mathbb{P}(\mathbb{S})$, es decir,

$$\theta(s) = P_0[\xi_n = s], \quad s \in \mathbb{S},\ n \in \mathbb{N}_0.$$

Cabe aclarar que en adelante, c.s significará casi seguramente respecto a la probabilidad P_0.

Definición 3.1 *Dado el MCM $\tilde{\mathfrak{M}}$ y $n \in \mathbb{N}_0$ definimos el espacio de historias admisibles hasta la n-ésima etapa como*

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

Entonces, un elemento de $\mathbb{H}_n$ es un vector (o historia) de la forma

$$h_n = (x_0, a_0, s_1, \ldots, x_{n-1}, a_{n-1}, s_n, x_n),$$

con $(x_k, a_k, s_{k+1}) \in \mathbb{K} \times \mathbb{S}$ para $k = 0, 1, \ldots, n-1$ y $x_n \in \mathbb{X}$.

Como definimos en el Capítulo 1, en general una regla de decisión aleatorizada es una función $\pi_n : \mathbb{H}_n \to \mathbb{P}(\mathbb{A})$, y son markovianas si $\pi_n : \mathbb{X} \to \mathbb{P}(\mathbb{A})$. Similarmente se definen las reglas de decisión deterministas, así como las políticas de control correspondientes.

Preservaremos la misma notación Π para el conjunto de todas las políticas y $\mathbb{F}$ para las estacionarias, tomando en cuenta las consideraciones específicas de este caso.

Además, si consideramos el espacio muestral $\Omega = (\mathbb{K} \times \mathbb{S})^{\infty} \times \mathbb{X}$, para cada $\pi \in \Pi$ y estado inicial $x \in \mathbb{X}$ existe una probabilidad P_x^{π} en la familia de subconjuntos de Ω tal que

$$
\begin{aligned}
&P_x^{\pi}[x_0 = x] = 1, \\
&P_x^{\pi}[a_n = a \mid h_n] = \pi_n(a \mid h_n), \quad a \in \mathbb{A} \\
&P_x^{\pi}[\xi_{n+1} = s \mid h_n, a_n] = \theta(s), \quad s \in \mathbb{S} \\
&P_x^{\pi}[x_{n+1} = y \mid h_n, a_n, \xi_{n+1}] = P_{x_n,y}(a_n), \quad y \in \mathbb{X}.
\end{aligned}
$$

3.3 Índice de funcionamiento

Tomando en cuenta la descripción del modelo $\tilde{\mathcal{M}}$, podemos definir el índice de funcionamiento de la siguiente manera.

Definición 3.2 *Sea $\pi \in \Pi$ y $x_0 = x$. El costo total esperado con factor de descuento (x, a, ξ)-dependiente se define como:*

$$
V(x, \pi) := E_x^{\pi}\left[\sum_{n=0}^{\infty} \tilde{\Gamma}_n c(x_n, a_n)\right], \tag{3.4}
$$

donde

$$
\begin{aligned}
&\tilde{\Gamma}_n := \prod_{k=0}^{n-1} \tilde{\alpha}(x_k, a_k, \xi_{k+1}), \ \ si \ \ n \in \mathbb{N} \ \ y \\
&\tilde{\Gamma}_0 := 1.
\end{aligned}
$$

Por lo tanto, el problema de control óptimo, consiste en encontrar una política $\pi' \in \Pi$ tal que

$$
V(x, \pi') = \inf_{\pi \in \Pi} V(x, \pi) =: V'(x), \ \ x \in \mathbb{X}.
$$

Para garantizar la existencia de la política óptima, asumiremos que se cumple la siguiente hipótesis.

Hipótesis 3.3 *a) Se satisface la Hipótesis 2.1(a) y (d).*

b) La función $\tilde{\alpha}$ satisface

$$\alpha_0^* := \sup_{(x,a,s)\in\mathbb{K}\times\mathbb{S}} \tilde{\alpha}(x,a,s) < 1$$

y la Hipótesis 2.1(c) se cumple con α_0^* en lugar de α^*

ENTONCES, BAJO ESTA CONDICIÓN, ES CONOCIDO QUE LA POLÍTICA ÓPTIMA SE ENCUENTRA EN LA CLASE DE LAS POLÍTICAS MARKOVIANAS, LO CUAL SIGNIFICA QUE, LAS POLÍTICAS MARKOVIANAS SON SUFICIENTES PARA RESOLVER EL PCO. POR LO TANTO, SI DENOTAMOS POR Π_M AL CONJUNTO DE POLÍTICAS MARKOVIANAS, NUESTRO PROBLEMA ES ENCONTRAR UNA POLÍTICA $\varphi^* \in \Pi_M$ TAL QUE argumentar por que es conocido

$$V(x,\varphi^*) = \inf_{\varphi\in\Pi_M} V(x,\varphi) =: V(x), \;\; x \in \mathbb{X}. \tag{3.5}$$

Mas aún, en la clase Π_M podemos escribir el índice de funcionamiento (3.4) en términos de la distribución θ de las v.a.'s ξ_t tal como lo establece el siguiente resultado.

Lema 3.4 *Para cada $x \in \mathbb{X}$ y cada $\varphi \in \Pi_M$,*

$$V(x,\varphi) := E_x^{\varphi}\left[\sum_{n=0}^{\infty} \Gamma_n^{\theta} c(x_n,a_n)\right], \tag{3.6}$$

donde $\Gamma_n^{\theta} = \prod_{n=0}^{n-1} \alpha_{\theta}(x_n,a_n)$, $\Gamma_0^{\theta} = 1$, y

$$\alpha_{\theta}(x,a) = E(\tilde{\alpha}(x,a,\xi)) = \sum_{s\in\mathbb{S}} \tilde{\alpha}(x,a,s)\theta(s), \quad (x,a) \in \mathbb{K}. \tag{3.7}$$

Demostración. La relación (3.6) es una consecuencia de las propiedades de la probabilidad P_x^π. En efecto, para cada $x \in \mathbb{X}$ y $\varphi \in \Pi_M$

$$\begin{aligned}
E_x^\varphi \tilde{\alpha}(x_0,a_0,\xi_1)c(x_1,a_1) &= \sum_{a_0}\sum_{s_1}\sum_{x_1}\sum_{a_1}\tilde{\alpha}(x_0,a_0,s_1)c(x_1,a_1)\varphi_1(a_1 \mid x_1) \\
&\quad P_{x_0,x_1}(a_0)\theta(s_1)\varphi_0(a_0 \mid x_0) \\
&\quad = \sum_{a_0}\sum_{s_1}\tilde{\alpha}(x_0,a_0,s_1)\theta(s_1)\sum_{x_1}\sum_{a_1}c(x_1,a_1)\varphi_1(a_1 \mid x_1) \\
&\quad P_{x_0,x_1}(a_0)\varphi_0(a_0 \mid x_0) \\
&\quad = \sum_{a_0}\alpha_\theta(x_0,a_0)\sum_{x_1}\sum_{a_1}c(x_1,a_1)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0) \\
&\quad \varphi_0(a_0 \mid x_0) \\
&\quad = \sum_{a_0}\sum_{x_1}\sum_{a_1}\alpha_\theta(x_0,a_0)\varphi_0(a_0 \mid x_0)c(x_1,a_1)P_{x_0,x_1}(a_0) \\
&\quad \varphi_1(a_1 \mid x_1) \\
&\quad = E_x^\varphi \alpha_\theta(x_0,a_0)c(x_1,a_1).
\end{aligned}$$

Ahora obsérvese que

$$\begin{aligned}
E_x^\varphi \tilde{\Gamma}_2 c(x_2,a_2) &= \sum_{a_0}\sum_{s_1}\sum_{x_1}\sum_{a_1}\sum_{s_2}\sum_{x_2}\sum_{a_2}\tilde{\alpha}(x_0,a_0,s_1)\tilde{\alpha}(x_1,a_1,s_2)c(x_2,a_2) \\
&\quad \varphi_2(a_2 \mid x_2)P_{x_1,x_2}(a_1)\theta(s_2)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0)\varphi_0(a_0 \mid x_0)\theta(s_1) \\
&\quad = \sum_{a_0}\sum_{s_1}\tilde{\alpha}(x_0,a_0,s_1)\theta(s_1)\sum_{x_1}\sum_{a_1}\sum_{s_2}\sum_{x_2}\sum_{a_2}\tilde{\alpha}(x_1,a_1,s_2) \\
&\quad c(x_2,a_2)\varphi_2(a_2 \mid x_2)P_{x_1,x_2}(a_1)\theta(s_2)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0)\varphi_0(a_0 \mid x_0) \\
&\quad = \sum_{a_0}\alpha_\theta(x_0,a_0)\sum_{x_1}\sum_{a_1}\sum_{s_2}\tilde{\alpha}(x_1,a_1,s_2)\theta(s_2)\sum_{x_2}\sum_{a_2}c(x_2,a_2) \\
&\quad \varphi_2(a_2 \mid x_2)P_{x_1,x_2}(a_1)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0)\varphi_0(a_0 \mid x_0) \\
&\quad = \sum_{a_0}\alpha_\theta(x_0,a_0)\varphi_0(a_0 \mid x_0)\sum_{x_1}\sum_{a_1}\alpha_\theta(x_1,a_1)\sum_{x_2}\sum_{a_2}c(x_2,a_2) \\
&\quad \varphi_2(a_2 \mid x_2)P_{x_1,x_2}(a_1)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0) \\
&\quad = \sum_{a_0}\sum_{x_1}\sum_{a_1}\sum_{x_2}\sum_{a_2}\alpha_\theta(x_0,a_0)\varphi_0(a_0 \mid x_0)\alpha_\theta(x_1,a_1)c(x_2,a_2) \\
&\quad \varphi_2(a_2 \mid x_2)P_{x_1,x_2}(a_1)\varphi_1(a_1 \mid x_1)P_{x_0,x_1}(a_0) \\
&\quad = E_x^\varphi \Gamma_2 c(x_2,a_2).
\end{aligned}$$

Aplicando argumentos similares, se obtiene que para $n \geq 3$, se tiene

$$E_x^{\varphi}\tilde{\Gamma}_n c(x_n,a_n) = E_x^{\varphi}\Gamma_n c(x_n,a_n).$$

En consecuencia, de (3.4) se obtiene (3.6).

■

3.4 Ecuación de optimalidad

Obsérvese que para todo $(x,a) \in \mathbb{K}$,

$$\alpha_\theta(x,a) = \sum_{s\in\mathbb{S}} \tilde{\alpha}(x,a,s)\theta(s) \leq \alpha_0^* \sum_{s\in\mathbb{S}} \theta(s) = \alpha_0^* < 1.$$

Entonces

$$\sup_{(x,a)\in\mathbb{K}} \alpha_\theta(x,a) < 1.$$

Para facilitar la exposición de los resultados se hará uso de la notación

$$\alpha_0^* := \sup_{(x,a)\in\mathbb{K}} \alpha_\theta(x,a) < 1. \tag{3.8}$$

Por lo tanto, sobre la clase de las políticas markovianas Π_M, podemos aplicar los resultados del Capítulo 2 para resolver el PCO (3.5). En efecto, comparando los índices (1.3) y (3.6) podemos tomar $\Gamma_n^\theta = \Gamma_n$ y $\alpha_\theta(x,a) = \alpha(x,a)$ y definir el operador

$$T_\theta u(x) := \min_{a\in A(x)} \left\{ c(x,a) + \alpha_\theta(x,a) \sum_{y\in\mathbb{X}} u(y) P_{x,y}(a) \right\} \quad u \in B_W(\mathbb{X}), \;\; x \in \mathbb{X}. \tag{3.9}$$

Como consecuencia se tiene la validez del siguiente resultado.

Teorema 3.5 *Bajo las Hipótesis 3.3,*

(a) La función de valor óptimo V es el único punto fijo del operador T_θ:

$$V(x) = T_\theta V(x), \;\; x \in \mathbb{X}$$

y

$$\|V\|_W \leq \frac{L}{1-\alpha_0^*\beta} \tag{3.10}$$

(b) Existe $f^* \in \mathbb{F}$ tal que

$$V(x) = c(x,f^*) + \alpha_\theta(x,f^*)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(f^*), \;\; x \in \mathbb{X},$$

y la política estacionaria f^* es óptima.

3.5 Estimación empírica y control

En esta sección estudiaremos el PCO asociado al modelo $\tilde{\mathfrak{M}}$ asumiendo que las v.a. $\{\xi_n\}$ son observables e i.i.d. con distribución θ desconocida.

Bajo la hipótesis de observabilidad del proceso $\{\xi_n\}$, podemos implementar esquemas de estimación estadística de θ y combinarlo con procedimientos de minimización a fin de obtener políticas casi óptimas. En este trabajo usaremos la distribución empírica como estimador de θ.

La evolución de este proceso de estimación y control es de la siguiente manera. En la n-ésima etapa, antes de elegir el control $a_n \in A(x_n)$, el controlador estima θ mediante la distribución empírica

$$\theta_n(k) := \frac{1}{n}\sum_{i=1}^{n} \delta_k(s), \quad \forall n \in \mathbb{N}, \;\; k \in \mathbb{S}$$

donde

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

Luego, el controlador combina este proceso con la historia del sistema para seleccionar un control (o acción) $a = a_n(\theta_n) \in A(x_n)$. Entonces, se genera un costo descontado a una tasa de descuentos multiplicativa de la forma

$$\tilde{\alpha}(x_0,a_0,\xi_1)\tilde{\alpha}(x_1,a_1,\xi_2)\cdots\tilde{\alpha}(x_{n-1},a_{n-1},\xi_n)c(x_n,a_n),$$

y el sistema avanza a un nuevo estado $x_{n+1} = y \in \mathbb{X}$ de acuerdo a la ley de transición

$$P(x_{n+1} = y \mid x_n = x, a_n = a).$$

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

Es importante señalar que, de acuerdo a la Definición 3.2, el costo total esperado con factor de descuento que depende del estado-acción-perturbación aleatoria, a su vez depende fuertemente de las acciones seleccionadas durante las primeras etapas, que es precisamente cuando la información respecto a la distribución desconocida θ resulta deficiente. Razón por la cual, bajo estas circunstancias no es posible, en general, obtener una política óptima, de manera que aquí estudiaremos la optimalidad de las políticas en un sentido asintótico.

A la política resultante de un proceso de estimación y control se le llama política adaptada.

Definición 3.6 *Diremos que una política $\pi \in \Pi$ es asintóticamente óptima para el modelo de control $\tilde{\mathfrak{M}}$ si para cada $x \in \mathbb{X}$,*

$$|V^n(x,\pi) - E_x^\pi[V(x_n)]| \to 0 \quad \textit{cuando} \;\; n \to \infty,$$

donde

$$V^n(x,\pi) := E_x^\pi \left[\sum_{t=n}^{\infty} \Gamma_{n,t} c(x_t, a_t) \right] \tag{3.11}$$

es el costo total esperado descontado con factor de descuento no constante de la etapa n en adelante y

$$\Gamma_{n,t} := \prod_{k=n}^{t-1} \alpha_\theta(x_k, a_k) \;\; \textit{para } t > n, \;\; y \;\; \Gamma_{n,n} := 1. \tag{3.12}$$

3.5.1 Políticas adaptadas asintóticamente óptimas

Primero definamos la sucesión de funciones $\{V_n\} \in B_W(\mathbb{X})$ como $V_0 \equiv 0$, y para $n \in \mathbb{N}$ mediante la siguiente ecuación recursiva

$$V_n(x) = T_{\theta_n} V_{n-1}(x) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha_{\theta_n}(x,a) \sum_{y \in \mathbb{X}} V_{n-1}(y) P_{x,y}(a) \right\}. \tag{3.13}$$

Además, como en (3.10) se obtiene

$$||V_n||_W \leq \frac{L}{1-\beta\,\alpha_0^*} \quad \forall n \in \mathbb{N}. \tag{3.14}$$

Por otro lado, definamos la familia $\mathcal{A}$ de funciones $\tilde{\alpha} : \mathbb{S} \to (0,1)$ como

$$\mathcal{A} := \{\tilde{\alpha}(x,a,\cdot) : (x,a) \in \mathbb{K}\}. \tag{3.15}$$

Nótese que por la Hipótesis 3.3, se tiene que la familia $\mathcal{A}$ es uniformemente acotada. Además, dado que $\mathbb{S}$ es numerable, por la Proposición A.5 se obtiene

$$\sup_{(x,a)\in\mathbb{K}} \left| \sum_{s\in\mathbb{S}} \tilde{\alpha}(x,a,s)\theta_n(s) - \sum_{s\in\mathbb{S}} \tilde{\alpha}(x,a,s)\theta(s) \right| \to 0 \quad c.s.,$$

es decir, $\eta_n \to 0$ c.s. cuando $n \to \infty$ donde

$$\eta_n := \sup_{(x,a)\in\mathbb{K}} |\alpha_{\theta_n}(x,a) - \alpha_\theta(x,a)|. \tag{3.16}$$

Ahora, presentamos dos propiedades fundamentales de la sucesión $\{V_n\}$.

Proposición 3.7 *Si se satisface la Hipótesis 3.3, entonces*

a) $||V - V_n||_W \to 0$ *c.s. cuando* $n \to \infty$.

b) Además, para cada $n \in \mathbb{N}$ *existe* $f_n = f_n^{\theta_n} \in \mathbb{F}$ *tal que* $f_n(x) \in A(x)$ *minimiza el lado derecho de (3.13), esto es,*

$$V_n(x) = c(x,f_n) + \alpha_{\theta_n}(x,f_n) \sum_{y\in\mathbb{X}} V_{n-1}(y) P_{x,y}(f_n), \quad x \in \mathbb{X}.$$

Demostración. a) Del Teorema 3.5(a) y (3.13) se tiene que

$$V(x) = TV(x) \quad \text{y} \quad V_n(x) = T_\theta V_{n-1}(x).$$

Entonces, para cada $x \in \mathbb{X}$ y $n \in \mathbb{N}$

$$\begin{aligned} |V(x) - V_n(x)| = \Bigg| & \inf_{a\in A(x)} \left\{ c(x,a) + \alpha_\theta(x,a) \sum_{y\in\mathbb{X}} V(y) P_{x,y}(a) \right\} \\ & - \inf_{a\in A(x)} \left\{ c(x,a) + \alpha_{\theta_n}(x,a) \sum_{y\in\mathbb{X}} V_{n-1}(y) P_{x,y}(a) \right\} \Bigg| \\ \leq & \sup_{a\in A(x)} \left| \alpha_\theta(x,a) \sum_{y\in\mathbb{X}} V(y) P_{x,y}(a) - \alpha_{\theta_n}(x,a) \sum_{y\in\mathbb{X}} V_{n-1}(y) P_{x,y}(a) \right|. \end{aligned}$$

Ahora sumando y restando el término $\alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a)$ se obtiene lo que muestra el desarrollo a continuación

$$\begin{aligned}
|V(x)-V_n(x)| \leq \sup_{a\in A(x)} \Bigg\{ \Bigg| & \alpha_\theta(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a) - \alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a) \\
& + \alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} [V(y)-V_{n-1}(y)]P_{x,y}(a) \Bigg| \Bigg\} \\
& \leq \sup_{a\in A(x)} \left\{ [\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)] \sum_{y\in\mathbb{X}} |V(y)|P_{x,y}(a) \right\} \\
& + \sup_{a\in A(x)} \left\{ \alpha_{\theta_n}(x,a) \sum_{y\in\mathbb{X}} |V(y)-V_{n-1}(y)|P_{x,y}(a) \right\} \\
& \leq \sup_{a\in A(x)} \left\{ [\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)] \sum_{y\in\mathbb{X}} ||V||_W W(y)P_{x,y}(a) \right\} \\
& + \sup_{a\in A(x)} \left\{ \alpha_{\theta_n}(x,a) \sum_{y\in\mathbb{X}} ||V-V_{n-1}||_W W(y)P_{x,y}(a) \right\} \\
& \leq \sup_{a\in A(x)} \{|\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)|\beta||V||_W W(x)\} \\
& + \sup_{a\in A(x)} \{\alpha_0^*\beta||V-V_{n-1}||_W W(x)\} \\
& \leq \beta||V||_W W(x) \sup_{a\in A(x)} \{|\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)|\} + \alpha_0^*\beta||V-V_{n-1}||_W W(x).
\end{aligned}$$

Por lo cual, de (3.10) y (3.16),

$$\begin{aligned}
||V-V_n||_W &\leq \beta||V||_W \sup_{(x,a)\in\mathbb{K}} \{|\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)|\} + \alpha_0^*\beta||V-V_{n-1}||_W \\
&\leq \frac{L\beta}{1-\beta\alpha_0^*}\eta_n + \alpha_0^*\beta||V-V_{n-1}||_W. \qquad (3.17)
\end{aligned}$$

Ahora, sea $\gamma := \limsup ||V-V_n||_W < \infty$. Tomando límite superior en (3.17) obsérvese que de (3.7) se obtiene

$$\gamma \leq \beta\alpha_0^*\gamma \quad c.s.$$

Por lo anterior, necesariamente $\gamma \equiv 0$ ya que $\beta\alpha_0^* < 1$. Esto implica

$$\lim_{n\to\infty} ||V-V_n||_W = 0 \quad c.s.$$

b) Nótese que esta parte es consecuencia directa del Teorema 3.5(b), la cual garantiza la existencia de tales minimizadores.

■

Ahora definamos la política $\hat{\pi} = \{\hat{\pi}_n\}$ como

$$\hat{\pi}_n(h_n) = \hat{\pi}_n(h_n;\theta_n) := f_n(x_n), \quad n \in \mathbb{N} \tag{3.18}$$

y $\hat{\pi}_0$ alguna función fija. Nuestro objetivo es demostrar que $\hat{\pi}$ es una política asintóticamente óptima. Para este fin, es necesario introducir la siguiente hipótesis.

Hipótesis 3.8 *Existen constantes no-negativas $d < \infty$, $\lambda < 1$ y $p > 1$ tal que*

$$\sum_{y\in\mathbb{X}} W^p(y)P_{x,y}(a) \leq \lambda W^p(x) + d, \quad \forall (x,a) \in \mathbb{K}. \tag{3.19}$$

Observación 3.9 *Aplicando la desigualdad de Jensen a (3.19) se tiene que, para toda $(x,a) \in \mathbb{K}$,*

$$\sum_{y\in\mathbb{X}} W(y)P_{x,y}(a) \leq \lambda' W(x) + d', \tag{3.20}$$

donde $\lambda' = \lambda^{\frac{1}{p}}$ y $d' = d^{\frac{1}{p}}$.

Lema 3.10 *Bajo la Hipótesis 3.8, para cada $\pi \in \Pi$ y cada $x \in \mathbb{X}$ se satisfacen las siguientes desigualdades*

$$(a) \quad \sup_{n\in\mathbb{N}_0} E_x^{\pi}[W^p(x_n)] < \infty. \tag{3.21}$$

$$(b) \quad \sup_{n\in\mathbb{N}_0} E_x^{\pi}[W(x_n)] < \infty. \tag{3.22}$$

Demostración. (a) Para cada $\pi \in \Pi$ y $x \in \mathbb{X}$, nótese que de la Observación 1.4 y la Hipótesis 3.8 se tiene

$$\begin{aligned} E_x^{\pi}\left[W^p(x_n) \mid h_{n-1}, a_{n-1}, \xi_n\right] &= \sum_{y\in\mathbb{X}} W^p(y)P_{x_{n-1},y}(a_n) \\ &\leq \lambda W^p(x_{n-1}) + d. \end{aligned}$$

Ahora, tomando esperanza en ambos lados de la desigualdad previa obtenemos

$$E_x^\pi [W^p(x_n)] \leq \lambda E_x^\pi [W^p(x_{n-1})] + d.$$

Luego, iterando la expresión obtenida obsérvese lo siguiente,

$$\begin{aligned} E_x^\pi [W^p(x_n)] &\leq \lambda E_x^\pi [W^p(x_{n-1})] + d \\ &\leq \lambda \left(\lambda E_x^\pi [W^p(x_{n-1})] + d\right) + d \\ &= \lambda^2 E_x^\pi [W^p(x_{n-2})] + \lambda d + d \\ &\vdots \\ &\leq \lambda^n E_x^\pi [W^p(x_0)] + d\left(1 + \lambda + \lambda^2 + \ldots + \lambda^{n-1}\right), \end{aligned}$$

y, puesto que $\lambda < 1$, lo anterior hace posible afirmar que

$$E_x^\pi [W^p(x_n)] \leq W^p(x) + \frac{d}{1-\lambda} < \infty \ \ \forall n \in \mathbb{N}.$$

En consecuencia,

$$\sup_{n \in \mathbb{N}_0} E_x^\pi [W^p(x_n)] < \infty.$$

(b) La demostración de esta parte sigue un procedimiento similar excepto que aquí se aplica la desigualdad (3.20).

■

Lema 3.11 *Bajo la Hipótesis 3.3, una política $\pi \in \Pi$ es asintoticamente óptima para el modelo de control $\tilde{\mathfrak{M}}$ si*

$$\lim_{n \to \infty} E_x^\pi [\Phi(x_n, a_n)] = 0, \ \ x \in \mathbb{X},$$

donde

$$\Phi(x,a) := c(x,a) + \alpha_\theta(x,a) \sum_{y \in \mathbb{X}} V(y) P_{x,y}(a) - V(x), \ (x,a) \in \mathbb{K}.$$

Demostración. Para cada $x \in \mathbb{X}$, $\pi \in \Pi$ y $n \in \mathbb{N}$, por la Hipótesis 3.3(b), así como las expresiones (3.8), (2.14) y (3.10)

$$\begin{aligned} E_x^\pi [\Gamma_{n,m} V(x_m)] &= E_x^\pi \left[\prod_{k=n}^{m-1} \alpha_\theta(x_k, a_k) V(x_m) \right] \\ &\leq E_x^\pi \left[\prod_{k=n}^{m-1} \alpha_0^* |V(x_m)| \right] \\ &\leq E_x^\pi \left[(\alpha_0^*)^{m-n} ||V||_W W(x_m) \right] \\ &= (\alpha_0^*)^{m-n} ||V||_W E_x^\pi [W(x_m)] \\ &\leq (\alpha_0^*)^{m-n} \beta^m W(x) ||V||_W \\ &\leq \frac{(\alpha_0^*)^{m-n} \beta^m L}{1 - \beta \alpha_0^*} W(x). \end{aligned}$$

Ahora, como $\beta \alpha_0^* < 1$,

$$\lim_{m \to \infty} E_x^\pi [\Gamma_{n,m} V(x_m)] = 0. \tag{3.23}$$

Además dado que Φ es no-negativa, se tiene que cada $x \in \mathbb{X}$ y cada $\pi \in \Pi$

$$\lim_{t \to \infty} E_x^\pi [\Phi(x_t, a_t)] = 0 \;\; \Rightarrow \;\; \lim_{n \to \infty} \sum_{t=n}^{\infty} E_x^\pi [\Gamma_{n,t} \Phi(x_t, a_t)] = 0. \tag{3.24}$$

Ahora, para cada $x \in \mathbb{X}$, $\pi \in \Pi$ y $t \in \mathbb{N}$

$$\Phi(x_t, a_t) = E_x^\pi [c(x_t, a_t) + \alpha_\theta(x_t, a_t) V(x_{t+1}) - V(x_t) \mid h_t, a_t]. \tag{3.25}$$

Entonces de (3.25), (3.11) y (3.12), para cada $t \geq n$

$$\begin{aligned} \sum_{t=n}^{\infty} E_x^\pi [\Gamma_{n,t} \Phi(x_t, a_t)] &= \sum_{t=n}^{\infty} E_x^\pi [\Gamma_{n,t} (E_x^\pi [c(x_t, a_t) + \alpha_\theta(x_t, a_t) V(x_{t+1}) \\ &\quad - V(x_t) \mid h_t, a_t])] \\ &= \sum_{t=n}^{\infty} E_x^\pi E_x^\pi [\Gamma_{n,t} c(x_t, a_t) \mid h_t, a_t] + \sum_{t=n}^{\infty} E_x^\pi E_x^\pi [\Gamma_{n,t} (\alpha_\theta(x_t, a_t) \\ &\quad V(x_{t+1}) - V(x_t) \mid h_t, a_t])] \\ &= E_x^\pi \sum_{t=n}^{\infty} \Gamma_{n,t} c(x_t, a_t) + \sum_{t=n}^{\infty} [E_x^\pi \Gamma_{n,t+1} V(x_{t+1}) \\ &\quad - E_x^\pi \Gamma_{n,t} V(x_t)] \\ &= V^{(n)}(x, \pi) - E_x^\pi \Gamma_{n,n} V(x_n) + \lim_{m \to \infty} E_x^\pi \Gamma_{n,m} V(x_m) \\ &= V^{(n)}(x, \pi) - E_x^\pi V(x_n), \end{aligned} \tag{3.26}$$

donde la última igualdad de sigue de (3.23). Finalmente, si en (3.26) tomamos límite cuando $n \to \infty$, entonces por (3.24) y la Definición 3.6 se obtiene el resultado deseado, es decir

$$\lim_{n\to\infty} \left| V^{(n)}(x,\pi) - E_x^{\pi} V(x_n) \right| = 0.$$

■

Ahora, ya tenemos las condiciones para establecer el resultado principal.

Teorema 3.12 *Bajo la Hipótesis 3.3 la política $\hat{\pi}$ es asintóticamente óptima.*

Demostración. De acuerdo al lema anterior, para demostrar el teorema es suficiente mostrar que

$$\lim_{n\to\infty} E_x^{\hat{\pi}} [\Phi(x_n, a_n)] = 0.$$

Para tal fin definimos, para cada $n \in \mathbb{N}$ la función $\Phi_n : \mathbb{K} \to \mathbb{R}$ como sigue

$$\Phi_n(x,a) := c(x,a) + \alpha_{\theta_n}(x,a) \sum_{y\in\mathbb{X}} V_{n-1}(y) P_{x,y}(a) - V_n(x).$$

Por lo tanto, de la Proposición 3.7(b) se deduce que

$$\Phi_n(x, f_n) = 0 \ \ \forall n \in \mathbb{N}, \ \ x \in \mathbb{X}. \tag{3.27}$$

Ahora, si $\{(x_n, a_n)\}$ es una sucesión de pares estado-acción correspondiente a la política $\hat{\pi}$, obsérvese que de (3.27)

$$\begin{aligned} \Phi(x_n, a_n) &\leq |\Phi(x_n, a_n) - \Phi_n(x_n, a_n)| \\ &\leq \sup_{a\in A(x_n)} |\Phi(x_n, a) - \Phi_n(x_n, a)| \\ &\leq W(x_n)\mathcal{L}_n, \end{aligned}$$

donde

$$\mathcal{L}_n := \sup_{(x,a)\in\mathbb{K}} \frac{|\Phi(x,a) - \Phi_n(x,a)|}{W(x)}.$$

Por lo tanto, ahora demostraremos

$$\lim_{n\to\infty} E_x^{\hat{\pi}}(W(x_n)\mathcal{L}_n) = 0. \tag{3.28}$$

Para cada $n \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}_n &= \sup_{(x,a)\in\mathbb{K}} \frac{|\Phi(x,a)-\Phi_n(x,a)|}{W(x)} \\ &= \sup_{(x,a)\in\mathbb{K}} \Bigg| \alpha_\theta(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a) - V(x) + V_n(x) \\ &\quad - \alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V_{n-1}(y)P_{x,y}(a)\Bigg| \frac{1}{W(x)}.\end{aligned}$$

Sumando y restando el término

$$\alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a)$$

se obtiene

$$\begin{aligned}\mathcal{L}_n &\le \sup_{(x,a)\in\mathbb{K}} \Bigg| \alpha_\theta(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a) - V(x) + V_n(y) - \alpha_{\theta_n}(x,a) \\ &\quad \sum_{y\in\mathbb{X}} V_{n-1}(y)P_{x,y}(a) + \alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a) - \alpha_{\theta_n}(x,a)\sum_{y\in\mathbb{X}} V(y)P_{x,y}(a)\Bigg| \\ &\le \sup_{(x,a)\in\mathbb{K}} |\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)| \frac{\sum_{y\in\mathbb{X}} ||V||_W W(y)P_{x,y}(a)}{W(x)} + ||V-V_n||_W \\ &\quad + \sup_{(x,a)\in\mathbb{K}} \frac{\alpha_0^* \sum_{y\in\mathbb{X}} ||V-V_{n-1}||_W W(y)P_{x,y}(a)}{W(x)} \\ &\le \frac{\beta L}{1-\beta\alpha_0^*} \sup_{(x,a)\in\mathbb{K}} |\alpha_\theta(x,a)-\alpha_{\theta_n}(x,a)| + ||V-V_n||_W + \beta\alpha_0^*||V-V_{n-1}||_W.\end{aligned}$$

Ahora, por la Proposición 3.7(a) y (3.16) se deduce que

$$\mathcal{L}_n \to 0 \;\; c.s. \;\; \text{cuando} \;\; n\to\infty. \tag{3.29}$$

Por otra parte, obsérvese que

$$\sup_{n\in\mathbb{N}} \mathcal{L}_n < M < \infty$$

para alguna constante M, y por (3.29) se tiene la convergencia en probabilidad

$$\mathcal{L}_n \xrightarrow{P_x^{\hat{\pi}}} 0 \text{ cuando } n \to \infty. \tag{3.30}$$

Además, por (3.21) se tiene

$$\sup_{n\in\mathbb{N}} E_x^{\hat{\pi}} (W(x_n)\mathcal{L}_n)^p \leq M^p \sup_{n\in\mathbb{N}} E_x^{\hat{\pi}} W^p(x_n) < \infty.$$

Esto implica que $\{W(x_n)\mathcal{L}_n\}$ es $P_x^{\hat{\pi}}$ uniformemente integrable (ver Teorema A.7).

Por otra parte, obsérvese que para dos números arbitrarios positivos m_1 y m_2,

$$[W(x_n)\mathcal{L}_n > m_1] = [W(x_n)\mathcal{L}_n > m_1, W(x_n) \leq m_2] \cup [W(x_n)\mathcal{L}_n > m_1, W(x_n) > m_2]$$

$$\subset \left[\mathcal{L}_n > \frac{m_1}{W(x_n)}, W(x_n) \leq m_2\right] \cup [W(x_n) > m_2],$$

de donde

$$[W(x_n)\mathcal{L}_n > m_1] \subset \left[\mathcal{L}_n > \frac{m_1}{m_2}\right] \cup [W(x_n) > m_2].$$

Este hecho implica

$$P_x^{\hat{\pi}}(W(x_n)\mathcal{L}_n > m_1) \leq P_x^{\hat{\pi}}\left(\mathcal{L}_n > \frac{m_1}{m_1}\right) + P_x^{\hat{\pi}}(W(x_n) > m_2).$$

Luego, aplicando la Desigualdad de Markov en el segundo término de la parte derecha de esta desigualdad se obtiene

$$P_x^{\hat{\pi}}(W(x_n)\mathcal{L}_n > m_1) \leq P_x^{\hat{\pi}}\left(\mathcal{L}_n > \frac{m_1}{m_1}\right) + \frac{E_x^{\hat{\pi}} W(x_n)}{m_2}.$$

asu vez, esta relación junto con (3.30) implica

$$W(x_n)\mathcal{L}_n \xrightarrow{P_x^{\hat{\pi}}} 0 \text{ cuando } n \to \infty. \tag{3.31}$$

Ahora, como $\{W(x_n)\mathcal{L}_n\}$ es $P_x^{\hat{\pi}}$ uniformemente integrable y además de (3.31) se garantiza su convergencia en probabilidad, entonces el Teorema A.8 hace posible concluir que

$$W(x_n)\mathcal{L}_n \xrightarrow{L^1} 0.$$

Ésto es, se obtiene (3.28), lo cual implica que $\hat{\pi}$ es asintóticamente óptima.

■

3.6 Ejemplo

En algunas aplicaciones la evolución del sistema está determinada por una ecuación en diferencias de la forma

$$x_{n+1} = F(x_n, a_n, \chi_n)$$

donde $\{\chi_n\}$ es una sucesión de v.a. i.i.d. con valores en algún conjunto numerable $\mathbb{D}$ siendo $F : \mathbb{X} \times \mathbb{A} \times \mathbb{D} \to \mathbb{X}$ una función conocida. En este caso, si g es la función de probabilidad común de las v.a. χ_n, es decir,

$$g(k) = P[\chi_n = k] \quad \forall\, k \in \mathbb{D},\ n \in \mathbb{N}_0,$$

entonces, para cada $(x,a) \in \mathbb{K}$ la ley de transición entre los estados toma la forma

$$\begin{aligned} P_{x,y}(a) &= P[x_{n+1} = y \mid x_n = x, a_n = a] \\ &= \sum_{k \in \mathbb{D}_y} g(k). \end{aligned}$$

donde $\mathbb{D}_y := \{k \in \mathbb{D} : F(x,a,k) = y\}$. Además, para una función u definida en $\mathbb{X}$

$$\sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) = \sum_{k \in \mathbb{D}} u[F(x,a,k)] g(k).$$

3.6.1 Un modelo de control autoregresivo

Consideremos un proceso de control de la forma

$$x_{n+1} = [\![\gamma G(a_n) x_n + \chi_n]\!], \quad n \in \mathbb{N}_0,$$

con x_0 dado, donde $[\![\cdot]\!]$ representa la parte entera, y tal que satisface las siguientes condiciones:

- Espacio de estados $\mathbb{X} = \mathbb{N}_0$.
- Conjunto de acciones admisibles finito $A(x) = \mathbb{A} \subset \mathbb{N}$.
- $G : \mathbb{A} \to B$ es una función del control donde $B = \{1, 2, \ldots, \bar{b}\}$, para algún $\bar{b} \in \mathbb{N}$.

- γ es una constante positiva tal que $\gamma\bar{b} < 1$.
- $\{\chi_n\}$ es una sucesión de v.a. i.i.d. discretas no-negativas con función de probabilidad g y esperanza $E[\chi_0] = \bar{\chi} < \infty$.
- La función de costo por etapa c satisface

$$|c(x,a)| \leq (x+l)^{\frac{1}{p}}, \quad (x,a) \in \mathbb{K}$$

  para algunas constantes $l > 1$ y $p > 1$.
- El proceso de perturbaciones aleatorias del factor de descuento $\{\varepsilon_n\}$ es una sucesión de v.a. i.i.d. que toman valores en un conjunto finito $\mathbb{S} = \{1,2,\ldots,\bar{s}\}$, $\bar{s} \in \mathbb{N}$.
- La función del factor de descuento es de la forma

$$\tilde{\alpha}(x,a,s) = \frac{s}{\kappa(x^2+a^2+1)}, \quad (x,a) \in \mathbb{K},\ s \in \mathbb{S},$$

  para alguna constante $\kappa > \bar{s}$.

Bajo estas condiciones obsérvese que

$$\tilde{\alpha}(x,a,s) \leq \frac{\bar{s}}{\kappa} \quad \forall (x,a,s) \in \mathbb{K} \times \mathbb{S}.$$

Por lo tanto,

$$\alpha^* := \sup_{(x,a,s)\in\mathbb{K}\times\mathbb{S}} \tilde{\alpha}(x,a,s) < 1.$$

Además, la ley de transición toma la forma

$$P_{x,y}(a) = \sum_{k\in\mathbb{D}_y} g(k)$$

donde $\mathbb{D}_y = \{k \in \mathbb{D} : [\![\gamma G(a)x + k]\!] = y\}$.

Por último, para verificar la Hipótesis 3.8, definamos

$$W(x) = (x+l)^{\frac{1}{p}}.$$

Entonces

$$\begin{aligned}
\sum_{y\in\mathbb{X}} W^p(y)P_{x,y}(a) &= \sum_{k\in\mathbb{D}} W^p\left([\![\gamma G(a)x+k]\!]\right) g(k) \\
&\leq \sum_{k\in\mathbb{D}} (\gamma G(a)x+k+l)g(k) \\
&= (\gamma G(a)x+l)\sum_{k\in\mathbb{D}} g(k) + \sum_{k\in\mathbb{D}} kg(k) \\
&\leq \gamma\bar{b}x+l+\bar{\chi} \\
&\leq \gamma\bar{b}(x+l)+l+\bar{\chi} \\
&\leq \lambda W^p(x)+d,
\end{aligned}$$

donde $\lambda = \gamma\bar{b} < 1$ y $d = l+\bar{\chi} < \infty$.

Apéndice A

Convergencia de variables aleatorias e integrabilidad uniforme

En este trabajo de tesis se usará la siguiente notación.

Notación:

- $\mathbb{N}$: conjunto de los números naturales.
- $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$: conjunto de los números naturales y el cero.
- $\mathbb{P}(S)$: conjunto de todas las medidas de probabilidad definidas en un conjunto S.

CONVERGENCIA DE VARIABLES ALEATORIAS

Sean $\xi, \xi_0, \xi_1, \ldots$ v.a.'s definidas en un espacio de probabilidad común $(\Omega, \mathcal{F}, P)$.

Definición A.1 *Diremos que* $\{\xi_n\}_{n\in\mathbb{N}_0}$ *converge casi seguramente a* ξ*, denotado por*

$$\xi_n \overset{a.s.}{\to} \xi,$$

si

$$P\{\omega \in \Omega : \xi_n(\omega) \to \xi(\omega) \ \ cuando \ \ n \to \infty\} = 1.$$

Definición A.2 *Diremos que* $\{\xi_n\}_{n\in\mathbb{N}_0}$ *converge en probabilidad a* ξ*, denotado por*

$$\xi_n \overset{P}{\to} \xi,$$

si para cada $\varepsilon > 0$,

$$P\left[\{\omega \in \Omega : |\xi_n(\omega) - \xi(\omega)| \geq \varepsilon\}\right] \to 0 \ \text{ cuando } \ n \to \infty.$$

Definición A.3 *Diremos que la sucesión* $\{\xi_n\}_{n\in\mathbb{N}}$ *de variables aleatorias integrables converge en media o en* L^1 *a la variable aleatoria integrable* ξ*, denotado por*

$$\xi_n \xrightarrow{L^1} \xi,$$

si

$$\lim_{n\to\infty} E|\xi_n - \xi| = 0.$$

CLASES GLIVENKO-CANTELLI

Definición A.4 *Sea* $\mathcal{H}$ *una familia de funciones* $h : \mathbb{S} \to \mathbb{R}$*. Diremos que* $\mathcal{H}$ *es una clase Glivenko-Cantelli si*

$$\sup_{h\in\mathcal{H}} \left| \sum_{k\in\mathbb{S}} h(k)\theta_t(k) - \sum_{k\in\mathcal{H}} h(k)\theta(k) \right| \to 0 \ \text{ cuando } \ t \to \infty.$$

Proposición A.5 *Si* $\mathcal{H}$ *es una familia de funciones uniformemente acotada y* $\mathbb{S}$ *es un conjunto numerable, entonces* $\mathcal{H}$ *es una clase Glivenko-Cantelli.*

INTEGRABILIDAD UNIFORME

Definición A.6 *Una sucesión de v.a.'s* $\{\xi_n\}_{n\in\mathbb{N}}$ *es uniformemente integrable si*

$$E\left(|\xi_n(\omega)|1_{\{|\xi_n|>a\}}\right) \to 0 \ \text{ cuando } \ a \to \infty$$

uniformemente en n.

Teorema A.7 *Sea* $\{\xi_n\}$ *una sucesión de v.a.'s, si* $E|\xi_n|^p \leq M \ \forall \ n \in \mathbb{N}$ *y* $p > 1$ *entonces* $\{\xi_n\}$ *es uniformemente integrable.*

Teorema A.8 $\xi_n \xrightarrow{p} \xi$ *y* $\{\xi_n\}$ *es uniformemente integrable si y solo si* $E|\xi_n| < \infty$ $\forall n \in \mathbb{N}$ *y además* $\xi_n \xrightarrow{L^1} \xi$.

Bibliografía

[1] Ash RB. (1972) *Real Analysis and Probability.* Academic Press, New York.

[2] Bertsekas DP., Shreve SE. (1978) *Stochastic Optimal Control: The Discrete-Time Case.* Academic Press, New York.

[3] Bertsekas DP. (1987) *Dynamic Programing: Deterministic and Stochastic Models.* Prentice-Hall, Englewood Cliffs, N.J.

[4] Dynkin EA., Yushkevich AA. (1979) *Controlled Markov Processes*. Springer-Verlag, New York. Syst., Estimation and Control, 4:99-140.

[5] Gil-Alana LA. (2004) *Modelling the U.S. interest rate in terms of I(d) statistical model*. The Quarterly Review of Economics and Finance, 44:475-486.

[6] Gonzáles-Hernández J., López-Martínez RR., Pérez-Hernández R. (2007) *Markov control processes with randomized discount cost in Borel space*. Math. Meth. Oper. Res., 65:27-44.

[7] Gonzáles-Hernández J., López-Martínez RR., Minjárez-Sosa JA. (2008) *Adaptive police for sthochastic systems under a randomized discounted criterion*. Bol. Soc. Mat. Mex., 14:149-163.

[8] Gonzáles-Hernández J., López-Martínez RR., Minjárez-Sosa JA. (2009) *Approximation, estimation and control of stochastic systems under a randomized discount cost criterion*. Kybernetika, 45:737-754.

[9] Gonzáles-Hernández J., López-Martínez RR., Minjárez-Sosa JA., Gabriel-Argüelles JR., (2013) *Constrained Markov control processes with randomized discounted cost criteria: occupation measures and extremal points*. Risk and Decision Analysis, 4:163-176.

[10] Gonzáles-Hernández J., López-Martínez RR., Minjárez-Sosa JA., Gabriel-Argüelles JR., (2014) *Constrained Markov control processes with randomized discounted rate: infinite linear programming approach*. Optim. Control Appl. Meth., 35:575-591.

[11] González-Trejo TJ., Hernández-Lerma O., Hoyos-Reyes LF. (2003) *Minimax control of discrte time stochastic systems*. SIAM J. Control Optim., 34:217-234.

[12] Gordienko EI., Minjárez-Sosa JA. (1998) *Adaptive control for discrete time Markov processes with unbounded costs: discounted criterion*. Kybernika, 34:217-234.

[13] Hernández-Lerma O. (1989) *Adaptive Markov Control Processes*. Springer, New York.

[14] Hernández-Lerma O., Lassarre JB. (1996) *Discrete Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York.

[15] Hernández-Lerma O., Lassarre JB. (1999) *Further Topics on Discrete Time Markov Control Processes*. Springer, New York.

[16] Hernández-Lerma O., González-Hernández J. (2000) *Constrained Markov control processes in Borel spaces: the discount case*. Math. Meth. Oper. Res., 52:271-285.

[17] Hernández-Lerma O., Runggaldier W. (1994) *Monotone approximations for convex stochastic control problems* . J. Math. Syst., Estimation and Control, 4:99-140.

[18] Hilgert N., Minjárez-Sosa JA. (2001) *Adaptive policies for time varying stochastic systems under discounted criterion*. Math. Meth. Oper. Res., 54:491-505.

[19] Hinderer K. (1979) *foundations of non Stationary Dynamic Programming with Discrete Time Parameter*. Lecture Notes Oper. Res., 33, Springer, New York.

[20] López-Martínez RR., Hernández-Lerma O. (2003) *The Lagrange approach to constrained Markov processes: a survey and extension of results*. Morfismos, 7:1-26.

[21] Luque-Vásquez F., Minjárez-Sosa JA. (2014) *Iteration algorithms in Markov decision processes with state-action-dependent discount factors and unbounded costs*. Reporte interno, Depto. de Matemáticas UNISON.

[22] Minjárez-Sosa JA. (2015) *Markov control models with unknown random state-action-dependent discount factors*. TOP. DOI:10.1007/s11750-015-0360-5.

[23] Piunovskiy AB. (1997) *Optimal Control of Random Sequences in Problems with Constraints* Kluwer, Dordrecht.

[24] Puterman ML. (1994) *Markov Decision Processes. Discrete Stochastic Dynamic Programming*. Wiley, New York.

[25] Schäl M. (1975) *Conditions for optimality and for the limit on n-stage optimal policies to be optimal*. Z. Wahrs. Verw. Gebiete., 32:179-196.

[26] Stockey NL., Lucas Jr. RE. (1989) *Recursive Methods in Economic Dynamics*. Harvard University Press, Cambridge, MA.

[27] Wei Q., Guo X. (2011) *Markov decision processes with state-depent discount factor and unbounded rewards/cost.* Oper. Rest. Lett., 39:368-274.