"El saber de mis hijos hará mi grandeza"

UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Programa de Licenciatura en Matemáticas

Control adaptado de sistemas estocásticos bajo el enfoque de normas ponderadas

T E S I S

Que para obtener el título de:

Licenciado en Matemáticas

Presenta:

María Elena Martínez Manzanares

Director de tesis: Dr. Jesús Adolfo Minjárez Sosa.

Hermosillo, Sonora, México, 5 de octubre 2018

Este trabajo fue financiado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) dentro del proyecto "Aproximación, Estimación y Control de Sistemas Estocásticos y Juegos Dinámicos", con número de referencia CB2015-01/254306, bajo la dirección del Dr. Jesús Adolfo Minjárez Sosa.

SINODALES

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

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

Dra. María Teresa Robles Alcaraz
Universidad de Sonora, Hermosillo, México

Dr. Óscar Vega Amaya
Universidad de Sonora, Hermosillo, México

Agradecimientos

Agradezco a mi mamá María Betsabé, a mi hermano Jorge Isaac, a mi abuela María Luisa, a mi abuelo José y a toda mi familia; a mis amigos Luz Esmeralda, Félix Alejandro, Jesús Arturo, Irenisolina, Paola Alejandra, Joselyn y a todos los que me brindaron su paciencia, compañía y cariño.

A todos aquellos que me ayudaron a llegar a esta etapa de mi formación académica.

Índice general

Introducción

La Teoría de Control Óptimo es un área de las matemáticas aplicadas que estudia problemas de decisiones secuenciales con el objetivo de encontrar las mejores "decisiones" para su funcionamiento óptimo. Es decir, se tiene un sistema dinámico cuya evolución en el tiempo puede ser influenciada mediante decisiones que toma un controlador, teniendo como objetivo encontrar su comportamiento óptimo. La evolución del sistema puede ser a tiempo continuo o discreto, y si están inmersos factores aleatorios diremos que tenemos un sistema de control estocástico. En este trabajo se estudiarán sistemas de control a tiempo discreto que evolucionan de la siguiente manera: en cada etapa el controlador observa el estado actual del sistema para posteriormente tomar una decisión o acción sobre el mismo; a continuación se produce un *costo* (o *recompensa*) y el sistema pasa a un nuevo estado con cierta probabilidad de transición.

Una pregunta natural que surge a partir de lo descrito anteriormente es cómo el controlador identifica las decisiones más adecuadas a partir del estado actual del sistema. Para resolver esto, se definen los conceptos *política de control* e *índice de funcionamiento*. Una política de control es, en términos generales, una sucesión de funciones que relacionan el estado del sistema (ya sea considerando toda la información anterior o sólo el estado actual) con el conjunto de acciones factibles a tomar, y el índice de funcionamiento "mide" el comportamiento del sistema al utilizar distintas Políticas, es decir, ayudan a discernir entre todas las políticas cuál es la mejor. En esta tesis utilizaremos un Índice de Funcionamiento llamado *costo descontado*.

Al conjunto de componentes que describen un sistema de control estocástico se le llama *Modelo de Control Markoviano* (MCM) o simplemente Modelo de Control. Entonces, dado un MCM, una familia de políticas de control y un índice de funcionamiento, el *Problema de Control Óptimo* (PCO) es encontrar una política que minimice tal índice. Este es precisamente el problema que estudia la Teoría de Control Óptimo Estocástico.

La tesis se centrará en el estudio de una clase particular de sistemas de control estocástico cuya evolución se describe mediante una ecuación en diferencias estocásticas de la forma

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

donde x_t y a_t representan el estado y acción elegida al tiempo t, respectivamente, F es una función conocida, y $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con función de probabilidad común ρ. En este contexto, la probabilidad de transición que describe la evolución del sistema toma la forma

$$\begin{aligned} P_{x,y}(a) &= P[x_{t+1} = y | x_t = x, a_t = a], \\ &= \sum_{\{k:F(x,a,k)=y\}} \rho(k). \end{aligned}$$

Como podemos observar, la función de probabilidad ρ es determinante para estudiar la evolución del sistema. Sin embargo existen situaciones donde ρ es desconocida, lo cual nos lleva a tener que implemetar métodos para su aproximación, y de esta manera obtener cierta información sobre la evolución del sistema al momento de elegir una acción o control. Si esto es posible de realizar, decimos que tenemos un problema de Control Estocástico Adaptado y a la política resultante la llamaremos política adaptada.

El objetivo principal del trabajo es estudiar esta clase de sistemas cuando ρ es desconocida y resolver el problema de control adaptado asociado. Entonces, con el fin de poder implementar un método de estimación para ρ, supondremos que las variables aleatorias $\{\xi_t\}$ son observables. A partir de aquí, es posible obtener estimadores estadísticos ρ_t que aproximan a ρ conforme avanza el sistema, de tal manera que las decisiones en cada etapa dependerán de la estimación correspondiente. Por lo tanto, la política adaptada toma la forma $\hat{\pi} = \{f_t^{\rho_t}\}$ donde $a_t = f_t^{\rho_t}(x_t)$.

Un punto que se debe observar en esta clase de políticas es que las primeras decisiones se toman con muy poca información respecto a la función de probabilidad desconocida ρ, contrario al caso cuando ya han transcurrido varias etapas, es decir, cuando t es grande. Este hecho implica que una política adaptada, en general, puede no ser óptimo para el caso descontado, y por lo tanto su optimalidad la estudiaremos en un sentido asintótico.

En resumen, el problema que estudiamos en la presente tesis es el siguiente: mostrar la existencia de políticas adaptadas que sean asintóticamente óptimas respecto al índice de optimalidad de costo descontado, en sistemas de control estocástico de la forma (1)

con función de probabilidad desconocida. Además, asumiremos que el costo por etapa es posiblemente no acotado.

Esto último lo trataremos bajo el esquema de normas ponderadas, es decir, asumiremos que el costo por etapa esta dominado por una función W la cual satisface una condición de crecimiento. En el caso de costo acotado, la función W es cualquier constante mayor o igual que la cota.

El presente trabajo está estructurado en tres capítulos. En el Capítulo 1 estableceremos a detalle los elementos que definen el del Problema de Control Óptimo y enunciaremos un ejemplo. En el Capítulo 2 se enunciarán una serie de condiciones y propiedades bajo las cuales el PCO tiene solución, es decir, se garantiza la existencia de políticas óptimas. En el tercer y último capítulo estudiaremos el Problema de Control Estocástico Adaptado, introduciendo el Modelo de Control Adaptado y las Políticas de Control Adaptadas.

Capítulo 1

El Modelo de Control Markoviano

1.1. Introducción

En este capítulo introducimos el Modelo de Control Markoviano (MCM) el cual describe el comportamiento de un sistema de control estocástico. Además, definiremos los conceptos de políticas de control e índice de funcionamiento con el fin de plantear el problema de control óptimo. Finalmente, para ilustrar estos conceptos presentamos un ejemplo de un sistema de producción-inventario.

1.2. Descripción del Modelo

Definición 1.2.1. *Un modelo de control markoviano (MCM) en tiempo discreto, es un arreglo* $(\mathbb{X}, \mathbb{A}, \{A(x) : x \in \mathbb{X}\}, P, c)$ *que consta de los siguientes elementos:*

(a) $\mathbb{X}$ *representa el espacio de estados; y supondremos que es un conjunto numerable.*

(b) $\mathbb{A}$ *es el espacio de controles o acciones; y supondremos que es un conjunto numerable.*

(c) $A(x) \subset \mathbb{A}$ *es el conjunto de acciones admisibles para el estado* $x \in \mathbb{X}$*. Además definimos*

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

al cual llamaremos el conjunto de pares estado-acción admisibles.

(d) *P representa la ley de transición entre los estados. Es decir*

$$P_{x,y}(a) := Pr[x_{t+1} = y | x_t = x, a_t = a]$$

que satisface las siguientes propiedades:

(i) $P_{x,y}(a) \geq 0 \; \forall \, x, y \in \mathbb{X}, a \in A(x)$;

(ii) $\sum_{y \in \mathbb{X}} P_{x,y}(a) = 1, \forall x \in \mathbb{X}, a \in A(x)$.

(e) $c : \mathbb{K} \to \mathbb{R}$ *representa la función de costo por etapa.*

Este modelo representa un sistema estocástico que evoluciona de la siguiente manera. En el tiempo $t = 0$, el controlador observa el estado inicial $x_0 \in \mathbb{X}$. Después toma una decisión o acción $a_0 \in A(x_0)$, lo cuál tendrá como consecuencia un costo $c(x_0, a_0)$, y el sistema se mueve a un siguiente estado x_1 con probabilidad $P_{x_0,x_1}(a_0)$. Este proceso se repite para cada etapa t; si el número de etapas es finito, diremos que el sistema tiene *horizonte finito*, y en caso contrario, diremos que tiene *horizonte infinito*.

Un caso particular de este tipo de sistemas de control es cuando la dinámica la define una ecuación en diferencias estocásticas de la forma

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

donde

(i) x_t es el estado al tiempo t.

(ii) a_t es la acción al tiempo t.

(iii) $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ que toman valores en un conjunto numerable S con función de probabilidad ρ.

En efecto, en este caso la ley de transición P toma la forma:

$$\begin{aligned} P_{x,y}(a) &= Pr[x_{t+1} = y | x_t = x, a_t = a], \\ &= P[F(x_t, a_t, \xi_t) = y | x_t = x, a_t = a], \\ &= P[F(x, a, \xi_t) = y], \\ &= \sum_{\{k : F(x,a,k) = y\}} \rho(k). \end{aligned}$$

1.3. Políticas de Control

Recordemos que en cada etapa t del sistema, después de observar el estado actual x_t, el controlador debe de tomar una acción a_t del conjunto de acciones admisibles $A(x_t)$. En esta sección describimos cómo se determina el control en cada tiempo.

Definición 1.3.1. *Dado un MCM, para cada $t \in \mathbb{N}_0$, definimos el espacio de historias admisibles hasta la etapa t mediante*

$$\mathbb{H}_0 := \mathbb{X},$$
$$\mathbb{H}_t := \mathbb{K}^t \times \mathbb{X},\ t \in \mathbb{N}.$$

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

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

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

Definimos el conjunto

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

donde a cada elemento de $\mathbb{F}$ se le llama selector.

Definición 1.3.2.

(a) *Una política de control es una sucesión $\pi = \{g_t\}$ de funciones $g_t : \mathbb{H}_t \to \mathbb{A}$ tal que $g_t(h_t) \in A(x_t)$ para todo $h_t \in \mathbb{H}_t, t \in \mathbb{N}_0$. Es decir, $a_t = g_t(h_t) \in A(x_t)$.*

(b) *Una política de control markoviana es una sucesión $\pi = \{f_t\}$, donde $f_t \in \mathbb{F}$, $\forall\, t \in \mathbb{N}_0$. Esto es, $a_t = f_t(x_t)$.*

(c) *Una política markoviana es estacionaria si existe $f \in \mathbb{F}$ tal que $f_t = f\ \forall\, t \in \mathbb{N}_0$. Es decir, $a_t = f(x_t)$ para toda t. En este caso denotamos $\pi = f$.*

Denotamos por Π al conjunto de todas las políticas e identificamos al conjunto de políticas estacionarias con el conjunto $\mathbb{F}$. En el caso de un MCM con horizonte de planeación finito N, una política es de la forma $\pi = \{f_0, f_1, \ldots, f_{N-1}\}$.

En un MCM con horizonte finito $N < \infty$, definimos el espacio muestral $\Omega_N := \mathbb{K}^N \times \mathbb{X}$ cuyos elementos son las *trayectorias* de la forma

$$\omega = (x_0, a_0, \ldots, x_{N-1}, a_{N-1}, x_N),$$

con $(x_k, a_k) \in \mathbb{K}$ si $k = 0, 1, \ldots, N-1$ y $x_N \in \mathbb{X}$. Si el MCM es de horizonte infinito, el espacio muestral es $\Omega = \mathbb{K}^\infty$ y las trayectorias son de la forma $\omega = (x_0, a_0, \ldots)$ donde $(x_k, a_k) \in \mathbb{K}$ para cada $k \in \mathbb{N}_0$.

Para cada estado inicial $x \in \mathbb{X}$ y cada política $\pi = \{g_0, g_1, \ldots\} \in \Pi$, existe una probabilidad P_x^π definida en una familia de subconjuntos de Ω tal que las variables x_t y a_t satisfacen

$$P_x^\pi(x_0 = x) = 1,$$

$$a_t = g_t(h_t),$$

$$P_x^\pi(x_{t+1} = y | h_t, a_t) = P_{x_t,y}(a_t).$$

En el caso de horizonte finito $N < \infty$, la probabilidad P_x^π se define por

$$P_x^\pi(x_0, a_0, \ldots, x_{N-1}, a_{N-1}, x_N) = \delta_x(x_0) P_{x_0,x_1}(a_0) \cdots P_{x_{N-1},x_N}(a_{N-1}),$$

donde $a_k = f_k(x_0, a_0, \ldots, x_k)$, $k = 0, 1, \ldots, N-1$, y $\delta_x(\cdot)$ es la probabilidad concentrada en x. Denotamos por E_x^π al operador esperanza respecto a P_x^π.

1.4. Indices de Funcionamiento

Un *índice de funcionamiento* es una función $w : \Pi \times \mathbb{X} \to \mathbb{R}$ que "mide" el rendimiento del sistema al utilizar distintas políticas de control dado el estado inicial. A continuación definimos los índices más importantes.

Definición 1.4.1. *Sean $x \in \mathbb{X}$ y $\pi \in \Pi$.*

(a) *Definimos el costo total esperado hasta la N-énsima etapa por*

$$J_N(\pi, x) := E_x^\pi[\sum_{t=0}^{N-1} c(x_t, a_t) + c_N(x_N)],$$

donde $c_N : \mathbb{X} \to \mathbb{R}$ es una función definida en $\mathbb{X}$ que representa un "costo terminal".

(b) *De manera similar, definimos el costo total con horizonte infinito como*

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

Es importante señalar que la función de costo total esperado para $N = \infty$ diverge en muchos casos. Este problema se evita (dependiendo si nos interesa analizar el sistema en sus primeras etapas o en el futuro) definiendo nuevos índices de funcionamiento como los que se presentan a continuación.

Definición 1.4.2. *Sean $x \in \mathbb{X}$ y $\pi \in \Pi$.*

(a) *Definimos el costo total esperado α-descontado con horizonte finito como*

$$V_\alpha^N(\pi, x) := E_x^\pi[\sum_{t=0}^{N-1} \alpha^t c(x_t, a_t) + \alpha^N c_N(x_N)],$$

donde $\alpha \in (0, 1)$ representa el factor de descuento.

(b) *Similarmente, definimos el costo total α-descontado con horizonte infinito como*

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

con $\alpha \in (0, 1)$ el factor de descuento.

(c) *Se define el costo promedio esperado como*

$$H(\pi, x) := \limsup_{N\to\infty} \frac{1}{N} E_x^\pi[\sum_{t=0}^{N-1} c(x_t, a_t)].$$

La razón del nombre "factor de descuento" para α en la definición de V_α es debido a que tiene una intrepretación monetaria: si el sistema se analiza durante períodos largos, el término α^t representa la depreciación del dinero en la etapa. Entonces, dado un costo L en el tiempo t, su valor en el tiempo presente es $\alpha^t L$.

1.5. Problema de Control Óptimo

Dado un MCM $(\mathbb{X}, \mathbb{A}, \{A(x) : x \in \mathbb{X}\}, P, c)$, una familia de políticas de control admisibles Π y uno de los índices de funcionamiento definidos en la Sección 1.4 al cual representamos por $\omega(\pi, x)$, el Problema de Control Óptimo (PCO) consiste en encontrar una política $\pi \in \Pi$ tal que

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

Llamaremos a π^* política óptima bajo el índice respectivo y a la función $\omega^*(\cdot)$ la función de valor óptimo.

1.6. Ejemplo: un sistema de producción-inventario

En el inicio de cada período, un controlador observa la cantidad de un artículo (nivel de inventario) que se encuentra disponible para su venta. Con base en esta información se ordena a la unidad de producción una cantidad adicional de artículos o conservar el nivel de inventario con el fin de satisfacer la demanda que se presentará durante el período. Definimos las siguientes variables:

- x_t: nivel de inventario al inicio del período t.
- a_t: cantidad de artículos ordenados al inicio del período t.
- ξ_t: demanda durante el período t. Supondremos que $\{\xi_t\}$ es una sucesión de variables aleatorias i.i.d. con función de probabilidad ρ.

Suponemos que se satisfacen las siguientes condiciones:

- El almacén tiene una capacidad infinita.
- La solicitud de artículos adicionales se hace al inicio de cada período y se surte inmediatamente.
- Los costos de producción del artículo no varían en diferentes períodos.
- Es posible conocer la demanda no satisfecha.

De estas condiciones obtenemos que $\mathbb{X} = \mathbb{A} = A(x) = \{0, 1, 2, \ldots\}$.

El sistema de producción-inventario evoluciona de acuerdo

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

con $t \in \mathbb{N}_0$ y $x_0 = x$. Verbalmente, la cantidad de artículos en el período $t+1$ será *lo que se tenía, más lo que se ordenó, menos lo que se vendió* en el período t. Podemos notar que la ley de transición del sistema viene dada por

$$\begin{aligned} P_{x,y}(a) &= Pr[x_{t+1} = y | x_t = x, a_t = a], \\ &= P[(x_t + a_t - \xi_t)^+ = y | x_t = x, a_t = a], \\ &= P[(x + a - \xi_t)^+ = y], \\ &= P[(x + a - \xi_t)^+ = y], \\ &= \sum_{k \in S_{(x,a,y)}} \rho(k), \end{aligned}$$

donde $S_{(x,a,y)} := \{s \in \mathbb{N}_0 | (x + a - s)^+ = y\}$.

Por último, la función de costo por etapa viene dada por

$$c(x, a) = \lambda a + h_1 E_\rho[(x + a - \xi_t)^+] + h_2 E_\rho[(\xi_t - x - a)^+],$$

donde

$$\begin{aligned} \lambda &: \text{precio (unitario) de producción,} \\ h_1 &: \text{costo (unitario) de almacenamiento,} \\ h_2 &: \text{costo (unitario) por demanda no satisfecha.} \end{aligned}$$

Capítulo 2

Criterio de costo descontado

2.1. Introducción

En este capítulo estudiaremos el PCO con el criterio de costo α-descontado. Con este propósito, introducimos condiciones que garantizan la existencia de políticas óptimas. Estas políticas serán caracterizadas por medio de una ecuación de optimalidad. Asimismo describiremos un algoritmo que aproxima a la solución de dicha ecuación. Estos resultados se desarrollarán asumiendo que el costo por etapa es posiblemente no acotado en el contexto de normas ponderadas. Para una fácil referencia, recordemos el índice de costo descontado y definamos PCO bajo este índice.

Sea $(\mathbb{X}, \mathbb{A}, \{A(x) : x \in \mathbb{X}\}, P, c)$ un MCM. El costo total esperado α-descontado es

$$V_\alpha(\pi, x) := E_x^\pi[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)], \quad \pi \in \Pi, x \in \mathbb{X},$$

donde $\alpha \in (0, 1)$ es el factor de descuento.

Para este índice, el problema de control óptimo consiste en encontrar $\pi^* \in \Pi$ tal que

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

2.2. Ecuación de optimalidad

Para cada $u : \mathbb{X} \to \mathbb{R}, \alpha \in (0,1)$, definimos el operador T_α como

$$T_\alpha u(x) := \inf_{a \in A(x)} [c(x,a) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a)], \qquad x \in \mathbb{X}. \tag{2.1}$$

Diremos que una función $u : \mathbb{X} \to \mathbb{R}$ es solución de la EO α-decontada si $T_\alpha u(x) = u(x)$, es decir

$$u(x) = \min_{a \in A(x)} [c(x,a) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a)], \quad x \in \mathbb{X}. \tag{2.2}$$

El análisis de la EO es clave para resolver el PCO, ya que es suficiente encontrar una solución y minimizar, como notaremos más adelante. Antes de entrar con los detalles, veremos la motivación de la EO. En términos generales, la forma de la EO es consecuencia del algoritmo de Programación Dinámica (PD) con horizonte finito N, el cual presentamos a continuación, y cuya demostración puede consultarse en [3].

Recordemos que el costo α-descontado total en N estapas cuando se utiliza la política π y el estado inicial x se define como

$$V_\alpha^N(\pi, x) := E_x^\pi [\sum_{t=0}^{N-1} \alpha^t c(x_t, a_t) + \alpha^N c_N(x_N)],$$

con función de valor óptimo

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

Teorema 2.2.1. ***(Algoritmo de Programación Dinámica)*** *Para $t = 0, 1, 2, \ldots, N$ se definen las funciones de programación dinámica v_t en $\mathbb{X}$ recursivamente por*

$$\begin{aligned} v_N(x) &= \alpha^N c_N(x), & (2.3)\\ v_t(x) &= \min_{a \in A(x)} \{\alpha^t c(x,a) + \sum_{y \in \mathbb{X}} v_{t+1}(y) P_{x,y}(a)\}, & (2.4)\\ &= \min_{a \in A(x)} \{\alpha^t c(x,a) + \sum_{k} v_{t+1}[F(x,a,k)]\rho(k)\}. \end{aligned}$$

donde $\alpha \in (0,1)$ *es el factor de descuento y* c_N *es la función de costo terminal. Si para cada* $t = N-1, N-2, \ldots, 0$, *existe* $f_t^* \in \mathbb{F}$ *tal que*

$$v_t(x) = \alpha^t c(x, f_t^*) + \sum_{y \in \mathbb{X}} v_{t+1}(y) P_{x,y}(f_t^*) \quad \forall x \in \mathbb{X}$$

entonces

(i) $v_0(x) = V_\alpha^N(x) = \text{ínf}_{\pi \in \Pi} V_\alpha^N(\pi, x)$.

(ii) *La política* $\pi^* = (f_0^*, f_1^*, \ldots, f_{N-1}^*)$ *es óptima, es decir*

$$V_\alpha^N(\pi^*, x) = V_\alpha^N(x) = v_0(x).$$

Podemos expresar las ecuaciones del algoritmo de programación dinámica en términos de las funciones $w_t(x) := \alpha^{-t} v_t(x), t = 0, 1, 2, \ldots, N$, como

$$w_N(x) = c_N(x), \tag{2.5}$$

$$w_t(x) = \min_{a \in A(x)} [c(x,a) + \alpha \sum_{y \in \mathbb{X}} w_{t+1}(y) P_{x,y}(a)] \tag{2.6}$$

para toda $x \in \mathbb{X}, t = N-1, N-2, \ldots, 0$.
En efecto, ya que partiendo de (2.3),

$$c_N(x) = \alpha^{-N} \alpha^N c_N(x) = \alpha^{-N} v_N(x) = w_N(x),$$

y partiendo de (2.4),

$$\begin{aligned}
w_t(x) &:= \alpha^{-t} v_t(x), \\
&= \alpha^{-t} \min_{a \in A(x)} [\alpha^t c(x,a) + \sum_{y \in \mathbb{X}} v_{t+1}(y) P_{x,y}(a)], \\
&= \min_{a \in A(x)} [c(x,a) + \alpha^{-t} \sum_{y \in \mathbb{X}} (\alpha^{t+1} \alpha^{-(t+1)}) v_{t+1}(y) P_{x,y}(a)], \\
&= \min_{a \in A(x)} [c(x,a) + \alpha^{-t} \alpha^{t+1} \sum_{y \in \mathbb{X}} \alpha^{-(t+1)} v_{t+1}(y) P_{x,y}(a)], \\
&= \min_{a \in A(x)} [c(x,a) + \alpha \sum_{y \in \mathbb{X}} w_{t+1}(y) P_{x,y}(a)].
\end{aligned}$$

Como las ecuaciones (2.3) y (2.4) son equivalentes a (2.5) y (2.6) respectivamente, el Algoritmo de Programación Dinámica sigue siendo válido sustituyendo las funciones v_t

por w_t y notemos que

$$\begin{aligned} w_0(x) &= V_\alpha^N(x), \\ &= \inf_\pi E_x^\pi[\sum_{k=0}^{N-1} \alpha^k c(x_k, a_k) + \alpha^N c_N(x_N)]. \end{aligned}$$

Es decir, obtuvimos el costo óptimo α-descontado para un problema de N etapas, y la política $\pi^* = \{f_0^*, f_1^*, \ldots, f_{N-1}^*\}$, donde $f_t^* \in \mathbb{F}, t = 0, 1, \ldots, N-1$, es óptima y minimiza el lado derecho de (2.6). Observemos que el algoritmo de PD resuelve el PCO de forma recursiva de adelante hacia atrás, lo cual es posible porque el horizonte es finito. Para el caso $N = \infty$, este algoritmo no es aplicable en su forma actual, y por lo tanto haremos algunas modificaciones para formularlo de forma recursiva hacia adelante.

Para esto, partiremos de las relaciones (2.5)-(2.6) y definimos $\nu_t := w_{N-t}, t = 0, 1, \ldots, N$. Entonces

$$\nu_0(x) = c_N(x),$$

$$\nu_{t+1}(x) = \min_{a \in A(x)} [c(x, a) + \alpha \sum_{y \in \mathbb{X}} \nu_t(y) P_{x,y}(a)]. \tag{2.7}$$

para toda $x \in \mathbb{X}, t = 0, 1, 2, \ldots, N$. Además, si $f_t \in \mathbb{F}, t = N-1, N-2, \ldots, 0$, minimiza el lado derecho de (2.6), entonces $g_t = f_{N-t}, t = 1, 2, \ldots, N$, minimiza (2.7). Más aún, observe que $\nu_N(x)$ es el costo óptimo de un problema en N etapas ya que

$$\nu_N(x) = w_0(x) = V_\alpha^N(x) \quad \forall x \in \mathbb{X}.$$

Entonces

$$\nu_n(x) = \inf_{\pi \in \Pi} V_\alpha^n(\pi, x), \quad \forall n \in \mathbb{N}, x \in \mathbb{X}.$$

En el caso particular que el costo terminal sea cero, *i.e.*, $c_N \equiv 0$, el algoritmo nos queda

$$\nu_0 \equiv 0 \tag{2.8}$$

$$\begin{aligned} \nu_{t+1}(x) &= \min_{a \in A(x)} [c(x, a) + \alpha \sum_{y \in \mathbb{X}} \nu_t(y) P_{x,y}(a)] = T_\alpha \nu_t(x), \\ &= T_\alpha^{t+1} \nu_0(x), \quad x \in \mathbb{X}, t = 0, 1, \ldots, N, \end{aligned} \tag{2.9}$$

al cual se le conoce como *Algoritmo de Iteración de Valores*. Esta relación motiva la definición de la EO.

Nos interesa estudiar el comportamiento de las funciones ν_t cuando $t \to \infty$. En lo que resta del capítulo veremos bajo que condiciones efectivamente ν_t converge a V_α^*.

2.3. Condiciones de optimalidad

En esta sección introducimos condiciones que garantizan la existencia de una solución a la EO y políticas óptimas y discutimos algunos resultados que son consecuencia de dichas condiciones.

Definición 2.3.1. *Sea $W : \mathbb{X} \to [1, \infty)$ una función aribitraria. Denotamos por $\mathbb{B}_W$ al espacio lineal normado de todas las funciones $u : \mathbb{X} \to \mathbb{R}$ con norma*

$$\|u\|_W := \sup_X \frac{|u(x)|}{W(x)} < \infty.$$

Hipótesis 2.3.2.

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

(b) *Existe una función $W : \mathbb{X} \to [1, \infty)$ tal que*

$$\max_{A(x)} |c(x, a)| \leq \bar{c} W(x) \quad \forall x \in \mathbb{X},$$

con $\bar{c}$ una constante mayor que cero.

(c) *Existe una constante $\beta \in (0, 1/\alpha)$ tal que*

$$\max_{a \in A(x)} \sum_{y \in \mathbb{X}} W(y) P_{x,y}(a) \leq \beta W(x) \quad \forall x \in \mathbb{X}. \tag{2.10}$$

Observación 2.3.3. *La Hipótesis 2.3.2 (a) garantiza que para cada $u \in \mathbb{B}_W$ existe $f \in \mathbb{F}$ tal que*

$$T_\alpha u(x) = c(x, f) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(f) \quad \forall x \in \mathbb{X}.$$

Proposición 2.3.4. *Suponga que se cumple la Hipótesis 2.3.2. Sea $T : \mathbb{B}_W \to \mathbb{B}_W$ un mapeo monótono. Si existe un número positivo $\gamma < 1$ tal que*

$$T(u + rW)(x) \leq Tu(x) + \gamma r W(x) \quad \forall u \in \mathbb{B}_W, \quad x \in \mathbb{X}, \tag{2.11}$$

y para toda $r \in \mathbb{R}_+$, entonces T es un operador de contracción módulo γ.

Demostración. Sean u, v funciones en $\mathbb{B}_W$. Entonces,

$$\frac{u(x) - v(x)}{W(x)} \leq ||u - v||_W \quad \forall x \in \mathbb{X},$$

lo cual implica la desigualdad

$$u(x) \leq v(x) + W(x)||u - v||_W \quad \forall x \in \mathbb{X}.$$

Del hecho de que T es monótono y (2.11), tomando $r = ||u - v||_W$, tenemos que

$$Tu(x) \leq T(v + rW)(x) \leq Tv(x) + \gamma r W(x),$$

es decir,

$$Tu(x) - Tv(x) \leq \gamma W(x)||u - v||_W.$$

Intercambiando u y v obtenemos que $Tu(x) - Tv(x) \geq -\gamma W(x)||u - v||_W$, lo cual implica

$$|Tu(x) - Tv(x)| \leq \gamma W(x)||u - v||_W.$$

Por lo tanto, $||Tu - Tv||_W \leq \gamma ||u - v||_W$. ■

Proposición 2.3.5. *Si se satisface la Hipótesis 2.3.2, entonces el operador T_α es monótono.*

Demostración. Sean $u, u' \in \mathbb{B}_W$ tales que $u \leq u'$. Entonces

$$c(x,a) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \leq c(x,a) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a) \quad \forall (x,a) \in \mathbb{K}.$$

Tomando ínfimo en $a \in A(x)$ en ambos lados de la desigualdad anterior se obtiene $Tu(x) \leq Tu'(x)$ para toda $x \in \mathbb{X}$. ■

Proposición 2.3.6. *Suponga que se satisface las Hipótesis 2.3.2. Entonces T_α es un operador de contracción en $\mathbb{B}_W$ con módulo $\gamma := \alpha\beta < 1$; es decir, T_α mapea $\mathbb{B}_W$ en sí mismo y*

$$||T_\alpha u - T_\alpha u'||_W \leq \gamma ||u - u'||_W \quad u, u' \in \mathbb{B}_W. \tag{2.12}$$

Entonces, por el Teorema de Punto Fijo de Bananch, existe una única función $u^* \in \mathbb{B}_W$ tal que $u^* = T_\alpha u^*$.

Demostración. Probaremos primero que T_α mapea $\mathbb{B}_W$ en sí mismo. Sea $u \in \mathbb{B}_W$. Supongamos que $T_\alpha u(x)$ alcanza el mínimo en $a^* \in A(x)$. De (2.10) tenemos que

$$\begin{aligned}
|T_\alpha u(x)| &= |\inf_{A(x)}[c(x,a) + \alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a)]|, \\
&= |c(x,a^*) + \alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a^*)|, \\
&\leq |c(x,a^*)| + |\alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a^*)|, \\
&\leq \bar{c}W(x) + |\alpha||u||_W[\sum_{y\in\mathbb{X}} W(y)P_{x,y}(a^*)]|, \\
&= \bar{c}W(x) + \alpha||u||_W[\sum_{y\in\mathbb{X}} W(y)P_{x,y}(a^*)], \\
&\leq (\bar{c} + (\alpha\beta)||u||_W)W(x) \quad \forall x \in \mathbb{X}.
\end{aligned}$$

De aquí se sigue que $T_\alpha u \in \mathbb{B}_W$.

Como el operador T_α es monótono, por la Proposición 2.3.4, para probar (2.12) es suficiente demostrar que

$$T_\alpha(u + rW)(x) \leq T_\alpha u(x) + (\alpha\beta)rW(x) \quad \forall x \in \mathbb{X}, \quad u \in \mathbb{B}_W, \quad r > 0.$$

En efecto,

$$\begin{aligned}
T_\alpha(u + rW)(x) &= \min_{a\in A(x)}[c(x,a) + \alpha \sum_{y\in\mathbb{X}} (u + rW)(y)P_{x,y}(a)], \\
&= \min_{a\in A(x)}[c(x,a) + \alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a) + \alpha r \sum_{y\in\mathbb{X}} W(y)P_{x,y}(a)], \\
&\leq \min_{a\in A(x)}[c(x,a) + \alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a) + \alpha r\beta W(x)], \\
&= T_\alpha(u) + (\alpha\beta)rW(x) \qquad \forall x \in \mathbb{X}.
\end{aligned}$$

■

Proposición 2.3.7. *Sea $\pi \in \Pi$ y $x \in \mathbb{X}$. Entonces, bajo la Hipótesis 2.3.2, para cada $t \in \mathbb{N}_0$ se cumplen las siguientes propiedades:*

(a) $E_x^\pi W(x_t) \leq \beta^t W(x)$;

(b) $|E_x^\pi c(x_t, a_t)| \leq E_x^\pi |c(x_t, a_t)| \leq \bar{c}\beta^t W(x)$.

Demostración. (a) De la Hipótesis 2.3.2 (c), tenemos

$$\begin{aligned} E_x^\pi[W(x_t)|x_0, a_0, \ldots, x_{t-1}, a_{t-1}] &= \sum_{y\in\mathbb{X}} W(y)P_{x_{t-1},y}(a_{t-1}) \\ &\leq \beta W(x_{t-1}) \end{aligned}$$

Calculando esperanza a ambos lados obtenemos

$$E_x^\pi[W(x_t)] \leq \beta E_x^\pi[W(x_{t-1})].$$

Iterando esta desigualdad, se obtiene

$$E_x^\pi[W(x_t)] \leq \beta^t W(x).$$

(b) La primera desigualdad en (b) es trivial. Para ver la segunda desigualdad, note que de Hipótesis 2.3.2 (b) se sigue

$$|c(x_t, a_t)| \leq \bar{c}W(x_t).$$

Tomando esperanza en ambos lados y considerando el inciso anterior tenemos que

$$\begin{aligned} E_x^\pi|c(x_t, a_t)| &\leq E_x^\pi \bar{c}W(x_t) = \bar{c}E_x^\pi W(x_t), \\ &\leq \bar{c}\beta^t W(x). \end{aligned}$$

■

Lema 2.3.8. *Suponga que se cumple la Hipótesis 2.3.2. Si $u \in \mathbb{B}_W$ satisface la desigualdad $u \leq T_\alpha u$, entonces $u(\cdot) \leq V_\alpha(\pi, \cdot)$ para toda $\pi \in \Pi$. Por lo tanto, $u \leq V^*$.*

Demostración. Notemos que $u \leq T_\alpha u$ implica que

$$u(x) \leq c(x, a) + \alpha \sum_{y\in\mathbb{X}} u(y)P_{x,y}(a) \quad \forall x \in \mathbb{X}, a \in A(x). \tag{2.13}$$

Sea $x \in \mathbb{X}$ y $\pi \in \Pi$ una política arbitraria. Entonces

$$\begin{aligned} E_x^\pi[\alpha^{t+1}u(x_{t+1})|h_t, a_t] &= \alpha^{t+1}\sum_{y\in\mathbb{X}} u(y)P_{x_t,y}(a_t), \\ &= \alpha^{t+1}\sum_{y\in\mathbb{X}} u(y)P_{x_t,y}(a_t) + \alpha^t c(x_t,a_t) - \alpha^t c(x_t,a_t), \\ &= \alpha^t[c(x_t,a_t) + \alpha\sum_{y\in\mathbb{X}} u(y)P_{x_t,y}(a_t)] - \alpha^t c(x_t,a_t), \\ &\geq \alpha^t u(x_t) - \alpha^t c(x_t,a_t). \end{aligned}$$

La última desigualdad es debe a (2.13). Entonces, como $u(x_t) = E_x^\pi[u(x_t)|h_t,a_t]$ tenemos

$$E_x^\pi[\alpha^{t+1}u(x_{t+1})|h_t,a_t] \geq \alpha^t E_x^\pi[u(x_t)|h_t,a_t] - \alpha^t c(x_t,a_t),$$

lo cual implica

$$\alpha^t c(x_t,a_t) \geq E_x^\pi[\alpha^t u(x_t) - \alpha^{t+1}u(x_{t+1})|h_t,a_t].$$

Tomando esperanza en ambos lados obtenemos

$$E_x^\pi\left[\alpha^t c(x_t,a_t)\right] \geq E_x^\pi[\alpha^t u(x_t) - \alpha^{t+1}u(x_{t+1})],$$

y sumando desde $t=0$ hasta n, obtenemos

$$\begin{aligned} E_x^\pi\left[\sum_{t=0}^{n}\alpha^t c(x_t,a_t)\right] &\geq E_x^\pi[\alpha^0 u(x) - \alpha^1 u(x_1)] + E_x^\pi[\alpha^1 u(x_1) - \alpha^2 u(x_2)] + \ldots + \\ &\quad E_x^\pi[\alpha^n u(x_n) - \alpha^{n+1}u(x_{n+1})], \quad \text{(suma telescópica)} \\ &\geq u(x) - E_x^\pi[\alpha^{n+1}W(x_{n+1})||u||_W], \\ &\geq u(x) - ||u||_W\alpha^{n+1}\beta^{n+1}W(x), \quad \text{por Proposición 2.3.7 (a).} \end{aligned}$$

Haciendo n tender a infinito obtenemos $V_\alpha(\pi,x) \geq u(x)$. Puesto que $\pi \in \Pi$ y $x \in \mathbb{X}$ son arbitrarios, concluimos que $V_\alpha^*(x) \geq u(x)$ para todo $x \in \mathbb{X}$. ■

2.4. Existencia de Políticas Óptimas

Sean $\{\nu_n\}$ las funciones de iteración de valores definidas en (2.8) y (2.9), es decir,

$$\begin{aligned} \nu_0 &\equiv 0 \\ \nu_{t+1}(x) &= \min_{a\in A(x)} [c(x,a) + \alpha \sum_{y\in\mathbb{X}} \nu_t(y) P_{x,y}(a)] = T_\alpha \nu_t(x) \\ &= T_\alpha^{t+1} \nu_0(x), \quad x \in \mathbb{X}, t = 0, 1, \ldots. \end{aligned}$$

Observe que bajo la Hipótesis 2.3.2, como $T_\alpha : \mathbb{B}_W \to \mathbb{B}_W$, tenemos que $\{\nu_n\} \subset \mathbb{B}_W$.

A continuación establecemos el resultado principal de este capítulo.

Teorema 2.4.1. *Suponga que la Hipótesis 2.3.2 se cumple. Sea β la constante de (2.10) y $\gamma := \alpha\beta$. Entonces:*

(a) *La función α-descontada $V_\alpha^*(x)$ es la única solución de la EO en el espacio $\mathbb{B}_W$, y*

$$||\nu_n - V^*||_W \leq \bar{c}\frac{\gamma^n}{1-\gamma}, \quad n = 1, 2, \ldots, \tag{2.14}$$

donde $\bar{c}$ es la constante en la Hipótesis 2.3.2 (b).

(b) *Existe una política estacionaria $f^* \in \mathbb{F}$ tal que se alcanza el mínimo en el lado derecho de la EO para toda $x \in \mathbb{X}$, esto es,*

$$V_\alpha^*(x) = c(x, f^*) + \alpha \sum_{y\in\mathbb{X}} V_\alpha^*(y) P_{x,y}(f^*) \quad \forall x \in \mathbb{X}. \tag{2.15}$$

Además, la política f^* es óptima, es decir, $V_\alpha(f^*, \cdot) = V_\alpha^*(\cdot)$.

Demostración. (a) Por la Proposición 2.3.6 y el Teorema de Punto Fijo de Banach (Proposición A.0.8), T_α tiene un único punto fijo $u^* \in \mathbb{B}_W$, es decir,

$$T_\alpha u^* = u^*,$$

y

$$||T_\alpha^n u - u^*||_W \leq \gamma^n ||u - u^*||_W \quad \forall u \in \mathbb{B}_W, n = 0, 1, \ldots. \tag{2.16}$$

Por lo tanto, para demostrar (a) solamente tenemos que mostrar:

(a_1) $V_\alpha^* \in \mathbb{B}_W$, con norma $||V_\alpha^*||_W \leq \frac{\bar{c}}{1-\gamma}$, y

(a_2) $V_\alpha^* = u^*$

En este caso, (2.14) se sigue de (2.9) y (2.16) con $u \equiv 0$.

Para probar (a_1), sea $\pi \in \Pi$ una política arbitraria y sea $x \in \mathbb{X}$ un estado arbitrario. Note que (a_1) se sigue de la Proposición 2.3.7 (b) y las siguientes relaciones:

$$\begin{aligned}
|V_\alpha(\pi, x)| &= |E_x^\pi[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)]|, \\
&\leq E_x^\pi[\sum_{t=0}^{\infty} \alpha^t |c(x_t, a_t)|], \\
&= \sum_{t=0}^{\infty} \alpha^t E_x^\pi |c(x_t, a_t)|, \\
&\leq \sum_{t=0}^{\infty} \alpha^t \bar{c} \beta^t W(x), \\
&= \bar{c} \sum_{t=0}^{\infty} (\alpha^t \beta^t) W(x), \\
&= \frac{\bar{c} W(x)}{1-\gamma}, \quad x \in \mathbb{X}.
\end{aligned}$$

Como $\pi \in \Pi$ y $x \in \mathbb{X}$ son arbitrarios, concluimos que

$$|V_\alpha^*(x)| \leq \frac{\bar{c} W(x)}{1-\gamma} \quad \forall x \in \mathbb{X}, \tag{2.17}$$

lo cual prueba que $V_\alpha^* \in \mathbb{B}_W$. Para probar (a_2), notemos primero que

$$\lim_{t \to \infty} \alpha^t E_x^\pi u(x_t) = 0 \quad \forall \pi \in \Pi, x \in \mathbb{X}, u \in \mathbb{B}_W. \tag{2.18}$$

En efecto, de la definición de la norma $||\cdot||_W$ tenemos que

$$\frac{|u(x)|}{W(x)} \leq \|u\|_W.$$

De esto último y de la Proposición 2.3.7 (a), vemos que

$$\alpha^t E_x^\pi |u(x_t)| \leq ||u||_W \alpha^t E_x^\pi W(x_t) \leq ||u||_W (\alpha\beta)^t W(x).$$

Tomando límite, obtenemos (2.18).

Por la Propisición 2.3.6 existe una única función $u^* \in \mathbb{B}_W$ tal que

$$\begin{aligned} u^*(x) &= T_\alpha u(x), \\ &= \inf_{A(x)} [c(x,a) + \alpha \sum_{y \in \mathbb{X}} u(y) P_{x,y}(a)], \end{aligned}$$

para todo $x \in \mathbb{X}$. De la Observación 2.3.3, existe $f^* \in \mathbb{F}$ tal que

$$u^*(x) = c(x, f^*) + \alpha \sum_{y \in \mathbb{X}} u^*(y) P_{x,y}(f^*) \quad \forall x \in \mathbb{X}. \tag{2.19}$$

Iterando obtenemos

$$u^*(x) = E_x^{f^*} \sum_{t=0}^{n-1} \alpha^t c(x_t, f^*) + \alpha^n E_x^{f^*} u^*(x_n) \quad \forall x \in \mathbb{X}, \quad n = 1, 2, \ldots. \tag{2.20}$$

De (2.18) y haciendo n tender a infinito se obtiene

$$u^*(x) = E_x^{f^*} [\sum_{t=0}^{\infty} \alpha^t c(x_t, f^*)] = V_\alpha(f^*, x). \tag{2.21}$$

Por consiguiente, de la definición de V_α^*, obtenemos que

$$u^*(x) \geq V_\alpha^*(x).$$

La otra desigualdad, se sigue del Lema 2.3.8 sustituyendo u por u^*, es decir, se cumple $u^*(x) \leq V_\alpha^*(x)$. Por lo tanto, $u^*(x) = V_\alpha^*(x), x \in \mathbb{X}$.

(b) Esta parte se sigue directamente de (2.21) puesto que $u^*(\cdot) = V_\alpha^*(f^*, \cdot)$. Así

$$u^*(\cdot) = V_\alpha^*(\cdot) = V(f^*, \cdot).$$

■

Capítulo 3

Estimación y Control en Sistemas Estocásticos

3.1. Introducción

Generalmente en los problemas de aplicación algunas de las componentes del modelo de control no son completamente conocidas por el controlador. Esto lleva a implementar esquemas que permitan ir 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 diseñar políticas de control que minimicen el índice de funcionamiento en consideración.

En el capítulo uno se introdujeron los MCM definidos por medio de ecuaciones en diferencias de la forma

$$x_{t+1} = F(x_t, a_t, \xi_t), t = 0, 1, \ldots \tag{3.1}$$

donde

(i) x_t es el estado al tiempo t,

(ii) a_t es la acción al tiempo t,

(iii) $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ que toman valores en un conjunto numerable S con función de probabilidad ρ.

Dado que la evolución del sistema es aleatoria y el comportamiento probabilístico lo determina ρ, resulta fundamental conocer esta función de probabilidad para estudiar la dinámica del sistema. No obstante, en muchos casos, suponer que ρ es conocida es poco realista. Ejemplos de estas situaciones aparecen cuando ξ_t representa la tasa de interés o la demanda de cierto artículo. Entonces, el problema que se presenta cuando ρ es desconocida por el controlador se puede plantear como un problema de control adaptado.

El material que se presenta en este capítulo está basado en los resultados de [2].

3.2. Modelo de Control

Consideremos un sistema de control estocástico que evoluciona en el tiempo mediante la ecuación en diferencia estocástica (3.1), donde la sucesión $\{\xi_t\}$ está formada por v.a. i.i.d. y observables tomando valores en un conjunto numerable $S \subset \mathbb{R}$, con función de probabilidad desconocida $\rho(s)$. Entonces

$$\rho(s) = P[\xi_t = s], \quad \forall t \in \mathbb{N}_0, \quad s \in S,$$

y

$$P[\xi_t \in B] = \sum_{s \in B} \rho(s) \quad \forall t \in \mathbb{N}_0.$$

Entonces la ley de transición toma la forma

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

donde

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

Sea

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

el modelo de control adaptado asociado a (3.1).

Observemos que para una función $u : \mathbb{X} \to \mathbb{R}$ arbitraria se cumple la igualdad

$$\sum_{y \in X} u(y) P_{x,y}(a) = \sum_{s \in S} u[F(x, a, s)] \rho(s) \quad \forall (x, a) \in \mathbb{K},$$

siempre y cuando alguna de las sumas este bien definida. Supondremos que todas las hipótesis consideradas en el capítulo anterior se cumplen en el presente contexto, por lo tanto el Teorema 2.4.1 es válido. Para una fácil referencia las escribiremos de nuevo. En particular observemos

$$T_\alpha u(x) = \min_{a\in A(x)} \{c(x,a) + \alpha \sum_{s\in S} u[F(x,a,s)]\rho(s)\} \quad x \in \mathbb{X},$$

y diremos que una función $u \in \mathbb{B}_W$ es solución de la ecuación de optimalidad α-descontada si $u(x) = T_\alpha u(x), x \in \mathbb{X}$, es decir,

$$u(x) = \min_{a\in A(x)} \{c(x,a) + \alpha \sum_{s\in S} u[F(x,a,s)]\rho(s)\} \quad x \in \mathbb{X}. \tag{3.2}$$

Hipótesis 3.2.1.

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

(b) *Existe una función $W : \mathbb{X} \to [1, \infty)$ tal que*

$$\max_{A(x)} |c(x,a)| \leq \bar{c} W(x) \qquad \forall x \in \mathbb{X},$$

(c) *Existe una constante $\beta \in (0, 1/\alpha)$ tal que*

$$\max_{A(x)} \sum_{s\in S} W[F(x,a,s)]\rho(s) \leq \beta W(x) \quad \forall x \in \mathbb{X}.$$

Teorema 3.2.2. *Suponga que se cumple la Hipótesis 3.2.1. Sea β la constante de Hipótesis 3.2.1 (c) y $\gamma := \alpha\beta$. Entonces:*

(a) *La función α-descontada V_α^* es la única solución de la EO en el espacio $\mathbb{B}_W$, y*

$$||\nu_n - V_\alpha^*||_W \leq \bar{c}\frac{\gamma^n}{1-\gamma}, \quad n = 1, 2, \ldots, \tag{3.3}$$

donde $\bar{c}$ es la constante en la Hipótesis 3.2.1 (b) y $\{\nu_n\}$ es la sucesión de funciones de iteración de valores.

(b) *Existe una política estacionaria $f^* \in \mathbb{F}$ tal que se alcanza el mínimo en el lado derecho para toda $x \in \mathbb{X}$, esto es,*

$$V_\alpha^*(x) = c(x, f^*) + \alpha \sum_{s\in S} V_\alpha^*[F(x, f^*, s)]\rho(s) \quad \forall x \in \mathbb{X}, \tag{3.4}$$

y f^* es óptima, es decir, $V_\alpha(f^*, \cdot) = V_\alpha(\cdot)$.

Como la función de probabilidad ρ es desconocida, observemos que la teoría desarrollada en el capítulo anterior no proporciona la política óptima, específicamente porque la ecuación de optimalidad depende de ρ. A partir de este hecho, el objetivo es introducir un procedimiento que combine métodos de estimación estadística de ρ y procesos de control para aproximar a la función de valor V_α^* y a la política óptima.

3.3. Estimación de la función de probabilidad

Sea $q \geq 1$ un número real fijo. Denotamos por l_q al espacio métrico de las sucesiones $\{x_j\}$ tales que $\sum_{j=1}^{\infty} |x_j|^q < \infty$, es decir

$$l_q := \{x = (x_1, x_2, \ldots) : \sum_{j=1}^{\infty} |x_j|^q < \infty\}.$$

Abusando de la notación, diremos que una función $\sigma : S \to \mathbb{R}$ pertenece a l_q si

$$\sum_{s \in S} |\sigma(s)|^q < \infty.$$

El objetivo de esta sección es mostrar la existencia de un estimador de $\rho \in l_q$ que herede sus mismas propiedades. Para esto necesitamos suponer lo siguiente:

Hipótesis 3.3.1.

(a) *Existe $q \in (1, 2)$ y una función $\tilde{\rho} \in l_q$ tal que $\rho \in l_q$ y $\rho(\cdot) \leq \tilde{\rho}(\cdot)$.*

(b) *Para cada $s \in S$,*

$$\psi(s) := \sup_{(x,a) \in \mathbb{K}} \frac{1}{W(x)} W[F(x, a, s)] < \infty,$$

y además

$$\sum_{s \in S} \psi^2(s) \tilde{\rho}^{(2-q)}(s) < \infty.$$

(c) *Supondremos que* $\inf_{x \in \mathbb{X}} W(x) = 1$.

Observación 3.3.2. *De la Hipótesis 3.3.1 (b), tenemos que para todo $(x, a) \in \mathbb{K}$ y $s \in S$,*

$$W[F(x, a, s)] \leq W(x)\psi(s).$$

Además, de la Hipótesis 3.3.1 (c) tenemos que $\psi(s) \geq 1 \quad \forall s \in S$.

Teorema 3.3.3. *Bajo la Hipótesis 3.3.1, existe un estimador*

$$\rho_t(s) = \rho_t(s; \xi_0, \xi_1, \ldots, \xi_{t-1}) \in l_q, s \in S, t \in \mathbb{N},$$

de ρ *tal que:*

(a) ρ_t *es una función de probabilidad.*

(b) $\rho_t(\cdot) \leq \tilde{\rho}(\cdot)$.

(c) $\sum_{s\in S} W[F(x,a,s)]\rho_t(s) \leq \beta W(x) \quad \forall t \in \mathbb{N}_0, (x,a) \in \mathbb{K}$.

(d) $E||\rho_t - \rho|| \to 0$ *cuando* $t \to \infty$, *donde para una función* $\sigma : S \to \mathbb{R}$,

$$||\sigma|| := \sup_{(x,a)\in\mathbb{K}} \frac{1}{W(x)} \sum_{s\in S} W[F(x,a,s)]|\sigma(s)| \tag{3.5}$$

El estimador ρ_t será usado en la siguiente sección para construir políticas adaptadas. En el resto de la sección nos ocuparemos en demostrar el Teorema 3.3.3.

Definamos los siguientes conjuntos de funciones de probabilidad

$$\begin{aligned} D_1 &:= \{\sigma \in l_q : \sigma \quad \text{es una función de probabilidad y} \quad \sigma(\cdot) \leq \tilde{\rho}(\cdot)\}, \\ D_2 &:= \{\sigma \in l_q : \sigma \quad \text{es una función de probabilidad y} \\ &\quad \sum_{s\in S} W[F(x,a,s)]\sigma(s) \leq \beta W(x), (x,a) \in \mathbb{K}\}, \\ D &:= D_1 \cap D_2. \end{aligned}$$

Lema 3.3.4. *Bajo la Hipótesis 3.3.1, el conjunto* D *es cerrado y convexo en* l_q.

Demostración. Sea $\{\sigma_n\}$ una sucesión en D tal que $\sigma_n \xrightarrow{l_q} \sigma$, es decir,

$$\left(\sum_{s\in S} |\sigma_n(s) - \sigma(s)|^q\right)^{1/q} \to 0, \tag{3.6}$$

cuando $n \to \infty$. Probaremos que $\sigma \in D$. Para hacer esto notemos primero que

$$\sigma(s) \leq \tilde{\rho}(s) \quad \forall s \in S. \tag{3.7}$$

Ahora mostraremos que

$$\sum_{s\in S} W[F(x,a,s)]\sigma(s) \leq \beta W(x) \quad \forall (x,a) \in \mathbb{K}, \tag{3.8}$$

para lo cual es suficiente ver que, cuando $n \to \infty$,

$$\sum_{s\in S} W[F(x,a,s)]\sigma_n(s) \to \sum_{s\in S} W[F(x,a,s)]\sigma(s) \quad \forall (x,a) \in \mathbb{K}. \tag{3.9}$$

De la Observación 3.3.2, para toda $(x,a) \in \mathbb{K}$,

$$\begin{aligned} I_n &:= |\sum_{s\in S} W[F(x,a,s)][\sigma_n(s) - \sigma(s)]|, \\ &\leq W(x)|\sum_{s\in S} \psi(s)[\sigma_n(s) - \sigma(s)]|, \\ &\leq W(x)\sum_{s\in S} \psi(s)|\sigma_n(s) - \sigma(s)|^{\frac{2-q}{2}}|\sigma_n(s) - \sigma(s)|^{q/2}. \end{aligned}$$

Por la Desigualdad de Cauchy-Schwarz y (3.7),

$$\begin{aligned} I_n &\leq W(x)\left[\sum_{s\in S} \psi^2(s)|\sigma_n(s) - \sigma(s)|^{2-q}\right]^{1/2}\left[\sum_{s\in S} |\sigma_n(s) - \sigma(s)|^q\right]^{1/2}, \\ &\leq W(x)\left[\sum_{s\in S} \psi^2(s)|2\tilde{\rho}(s)|^{2-q}\right]^{1/2}\left[\sum_{s\in S} |\sigma_n(s) - \sigma(s)|^q\right]^{1/2}, \\ &\leq MW(x)\left[\sum_{s\in S} |\sigma_n(s) - \sigma(s)|^q\right]^{1/2}, \end{aligned}$$

para alguna constante $M < \infty$ (ver Hipótesis 3.3.1 (b)). Tomando límite cuando $n \to \infty$, por (3.6) obtenemos que $I_n \to 0$, lo cual a su vez implica (3.9) y (3.8).

A continuación mostraremos que σ es una función de probabilidad. Para esto, observemos que por (3.6)

$$\begin{aligned}
|1-\sum_{s\in S}\sigma(s)| &= |\sum_{s\in S}\sigma_n(s)-\sum_{s\in S}\sigma(s)|,\\
&\leq \sum_{s\in S}|\sigma_n(s)-\sigma(s)|,\\
&= \sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{\frac{2-q}{2}}|\sigma_n(s)-\sigma(s)|^{q/2},\\
&\leq \left[\sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{2-q}\right]^{1/2}\left[\sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{q}\right]^{1/2},\\
&\leq \left[\sum_{s\in S}|2\tilde{\rho}(s)|^{2-q}\right]^{1/2}\left[\sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{q}\right]^{1/2}, \quad \text{pero por la Observación 3.3.2}\\
&\leq \left[\sum_{s\in S}|2\psi^2(s)\tilde{\rho}(s)|^{2-q}\right]^{1/2}\left[\sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{q}\right]^{1/2},\\
&\leq M^*\left[\sum_{s\in S}|\sigma_n(s)-\sigma(s)|^{q}\right]^{1/2}\to 0,
\end{aligned}$$

para alguna constante $M^*<\infty$ y haciendo $n\to\infty$; por lo tanto,

$$\sum_{s\in S}\sigma(s)=1.$$

Como $\sigma\geq 0$ tenemos que σ es una función de probabilidad.

Finalmente, mostraremos ahora la convexidad del conjunto D. Sean σ_1, σ_2 elementos de D. Demostraremos que

$$(1-t)\sigma_1(s)+t\sigma_2(s)\in D \quad \forall t\in[0,1], \quad s\in S.$$

Notemos que

$$\begin{aligned}
\sum_{s\in S}[(1-t)\sigma_1(s)+t\sigma_2(s)] &= (1-t)\sum_{s\in S}[\sigma_1(s)]+t\sum_{s\in S}[\sigma_2(s)],\\
&= (1-t)+t,\\
&= 1,
\end{aligned}$$

ya que $\sigma_1(s), \sigma_2(s)$ son funciones de probabilidad. Ahora bien, como $\sigma_1(s), \sigma_2(s) \in D$, entonces

$$\begin{aligned} \sigma_1(s) \leq \tilde{\rho}(s) \quad \text{y} \quad \sigma_2(s) \leq \tilde{\rho}(s), \quad \text{entonces} \\ (1-t)\sigma_1(s) \leq (1-t)\tilde{\rho}(s) \quad \text{y} \quad t\sigma_2(s) \leq t\tilde{\rho}(s). \end{aligned}$$

Por lo tanto,

$$\begin{aligned} (1-t)\sigma_1(s) + t\sigma_2(s) &\leq (1-t)\tilde{\rho}(s) + t\tilde{\rho}(s), \\ &= \tilde{\rho}(s)[(1-t)+t], \\ &= \tilde{\rho}(s). \end{aligned}$$

Siguiendo un esquema similar al anterior, como $\sigma_1(s), \sigma_2(s) \in D$, tenemos que

$$\sum_{s\in S} W[F(x,a,s)](1-t)\sigma_1(s) \leq (1-t)\beta W(x) \quad \text{y} \quad \sum_{s\in S} W[F(x,a,s)](t)\sigma_2(s) \leq (t)\beta W(x).$$

Entonces

$$\begin{aligned} \sum_{s\in S} W[F(x,a,s)](1-t)\sigma_1(s) + \sum_{s\in S} W[F(x,a,s)](t)\sigma_2(s) &\leq \\ (1-t)\beta W(x) + (t)\beta W(x) &= \\ \beta W(x). \end{aligned}$$

■

Sea $\hat{\rho}_t(s) := \hat{\rho}_t(s; \xi_0, \xi_1, \ldots, \xi_{t-1}), s \in S, t \in \mathbb{N}$ un estimador en l_q arbitrario de ρ tal que

$$\begin{aligned} E||\hat{\rho}_t - \rho||_{l_q}^{q/2} &= E||\hat{\rho}_t - \rho||_{l_q}^{1/2}, \\ &= E\left(\sum_{s\in S} |\hat{\rho}_t(s) - \rho(s)|^q\right)^{1/2} \to 0, \end{aligned} \tag{3.10}$$

cuando $t \to \infty$.

Como D es cerrado y convexo, de [1, Ejercicio 15.4, p. 169] (ver también [4]), para cada $t \in \mathbb{N}$, existe $\rho_t \in D$ tal que

$$||\rho_t - \hat{\rho}_t||_{l_q} = \inf_{\sigma\in D} ||\sigma - \hat{\rho}_t||_{l_q}.$$

Observemos que

$$\begin{aligned} ||\rho_t - \rho||_{l_q} &\leq ||\rho_t - \hat{\rho}_t||_{l_q} + ||\hat{\rho}_t - \rho||_{l_q}, \\ &\leq 2||\hat{\rho}_t - \rho||_{l_q}, \end{aligned}$$

luego elevando a la potencia $q/2$ a ambos lados obtenemos

$$||\rho_t - \rho||_{l_q}^{q/2} \leq 2^{q/2}||\hat{\rho}_t - \rho||_{l_q}^{q/2}.$$

Entonces, por (3.10) tenemos

$$E||\rho_t - \rho||_{l_q}^{q/2} \leq 2^{q/2}E||\hat{\rho}_t - \rho||_{l_q}^{q/2} \to 0. \tag{3.11}$$

Demostración del Teorema 3.3.3. Por todo lo anterior, observemos que es suficiente demostrar

$$E||\rho_t - \rho|| \to 0 \tag{3.12}$$

cuando $t \to \infty$. Por la Desigualdad de Cauchy-Schwarz y la Hipótesis 3.3.1

$$\begin{aligned} ||\rho_t - \rho|| &= \max_{(x,a)\in\mathbb{K}} \frac{1}{W(x)} \sum_{s\in S} W[F(x,a,s)]|\rho_t(s) - \rho(s)|, \\ &\leq \sum_{s\in S} \psi(s)|\rho_t(s) - \rho(s)|, \\ &\leq \left[\sum_{s\in S} \psi^2(s)|\rho_t(s) - \rho(s)|^{2-q}\right]^{1/2} \left[\sum_{s\in S} |\rho_t(s) - \rho(s)|^q\right]^{1/2}, \\ &\leq \left[\sum_{s\in S} \psi^2(s)|2\tilde{\rho}(s)|^{2-q}\right]^{1/2} \left[\sum_{s\in S} |\rho_t(s) - \rho(s)|^q\right]^{1/2}, \\ &\leq M' \left[\sum_{s\in S} |\rho_t(s) - \rho(s)|^q\right]^{1/2}, \\ &= M'||\rho_t(s) - \rho(s)||_{l_q}^{q/2}, \end{aligned}$$

para alguna constante $M < \infty$. Por lo tanto, de (3.11)

$$E||\rho_t - \rho|| \leq M'E||\rho_t - \rho||_{l_q}^{q/2} \to 0,$$

cuando $t \to \infty$. ∎

3.4. Optimalidad de políticas adaptadas

Como se comentó anteriomente, una política adaptada es aquella que combina procesos de estimación y control para elegir una acción. En nuestro caso, la política adaptada dependerá del estimador ρ_t en el sentido que el control al tiempo t tomará la forma

$$a_t = f_t(x_t, \rho_t) = f_t^{\rho_t}(x_t).$$

Por otro lado, de acuerdo al Teorema 3.3.3 es claro 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 de más importancia a las decisiones tomadas en las primeras etapas, precisamente donde el método de estimación proporciona una información pobre respecto a la función de probabilidad desconocida ρ, 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, como se define a continuación.

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

$$\phi(x, a) := c(x, a) + \alpha \sum_{s \in S} V_\alpha^*[F(x, a, s)]\rho(s) - V^*(x), \quad (x, a) \in \mathbb{K}.$$

Definición 3.4.1.

(a) *Una política de control markoviana $\pi = \{f_t\}$ es puntualmente asintóticamente óptima (PAO) si para cada $x \in \mathbb{X}$ se cumple que*

$$\phi(x, f_t(x)) \to 0$$

cuando $t \to \infty$.

La política $\pi = \{f_t\}$ es W-uniformemente asintóticamente óptima (W-UAO) si

$$\sup_{x \in \mathbb{X}} \frac{\phi(x, f_t(x))}{W(x)} \to 0$$

cuando $t \to \infty$.

(b) *Similarmente, una política de control adaptada $\pi = \{f_t^{\rho_t}\}$ es PAO si*

$$E[\phi(x, f_t^{\rho_t}(x))] \to 0$$

cuando $t \to \infty$ *y (W-UAO) si*

$$E \sup_{x \in \mathbb{X}} \frac{\phi(x, f_t^{\rho_t}(x))}{W(x)} \to 0$$

cuando $t \to \infty$.

3.5. Existencia de políticas adaptadas

Para una función $u : \mathbb{X} \to \mathbb{R}$, definimos el operador

$$T_t u(x) = \min_{a \in A(x)} \{c(x,a) + \alpha \sum_{s \in S} u[F(x,a,s)]\rho_t(s)\}, x \in \mathbb{X}, t \in \mathbb{N}_0.$$

Observe que por las propiedades de $\rho_t \in D$ y aplicando los argumentos del capítulo anterior, tenemos que $T_t : \mathbb{B}_W \to \mathbb{B}_W$ y es de contracción con módulo $\alpha\beta$, es decir, para $u, v \in \mathbb{B}_W$,

$$||T_t u - T_t v||_W \leq \alpha\beta ||u - v||_W. \tag{3.13}$$

Sea $\{V_t\} \subset \mathbb{B}_W$ una sucesión de funciones definidas como

$$\begin{aligned} V_0 &:= 0, \\ V_t(x) &:= T_t V_{t-1}(x), \\ &= \min_{a \in A(x)} \{c(x,a) + \alpha \sum_{s \in S} V_{t-1}[F(x,a,s)]\rho_{t-1}(s)\}, x \in \mathbb{X}, t \in \mathbb{N}_0. \end{aligned} \tag{3.14}$$

Mostraremos por inducción que

$$||V_t||_W \leq \frac{\bar{c}}{1 - \alpha\beta} \quad \forall t \in \mathbb{N}. \tag{3.15}$$

Para $t = 0$ tenemos que $V_0 = 0$. Supongamos que (3.15) se cumple para $t \in \mathbb{N}$, es decir,

$$||V_t||_W \leq \frac{\bar{c}}{1 - \alpha\beta}.$$

Entonces,

$$|V_t(x)| \leq \frac{\bar{c}}{1 - \alpha\beta} W(x).$$

Demostraremos que $|V_{t+1}(x)| \leq \frac{\bar{c}}{1-\alpha\beta}W(x)$. Supongamos que $V_{t+1}(x)$ alcanza el mínimo en $a^* \in A(x)$. De (3.14) tenemos

$$\begin{aligned}
|V_{t+1}(x)| &= |\min_{a\in A(x)}\{c(x,a)+\alpha\sum_{s\in S}V_t[F(x,a,s)]\rho_t(s)\}|,\\
&= |c(x,a^*)+\alpha\sum_{s\in S}V_t[F(x,a^*,s)]\rho_t(s)|,\\
&\leq |c(x,a^*)|+|\alpha\sum_{s\in S}V_t[F(x,a^*,s)]\rho_t(s)|,\\
&\leq \bar{c}W(x)+|\frac{\bar{c}}{1-\alpha\beta}\alpha\sum_{s\in S}W[F(x,a^*,s)]\rho_t(s)| \quad \text{(por hipótesis de inducción)},\\
&\leq \bar{c}W(x)\left(1+\frac{\alpha\beta}{1-\alpha\beta}\right),\\
&\leq \bar{c}W(x)\frac{1}{1-\alpha\beta}. \qquad (3.16)
\end{aligned}$$

Por lo tanto, (3.15) es válida para toda $t \in \mathbb{N}_0$. Por otro lado, como $A(x)$ es finito, tenemos que para cada $t \in \mathbb{N}$ existe $f_t = f_t^{\rho_t} \in \mathbb{F}$ tal que

$$V_t(x) = c(x, f_t(x)) + \alpha\sum_{s\in S}V_{t-1}[F(x,f_t,s)]\rho(s), x \in \mathbb{X}. \qquad (3.17)$$

Definición 3.5.1. *Sea $f_t \in \mathbb{F}$ el minimizador en (3.17). Definimos la política adaptada markoviana $\bar{\pi} = \{f_t\}$.*

El objetivo es mostrar que V_t converge a V^* en la norma $||\cdot||_W$ y que $\bar{\pi}$ es asintóticamente óptima en el sentido de la Definición 3.4.1 (b), lo cual lo establece el siguiente teorema.

Teorema 3.5.2. *Suponga que las Hipótesis 3.2.1 y 3.3.1 se cumplen. Entonces*

(a) $E||V_t - V^*||_W \to 0$ *cuando* $t \to \infty$.

(b) *Para cada $x \in \mathbb{X}$, $E[\phi(x, f_t(x))] \to 0$ cuando $t \to \infty$, y más aún,*

$$E\sup_{x\in\mathbb{X}}\frac{\phi(x,f_t(x))}{W(x)} \to 0 \quad \text{cuando} \quad t\to\infty$$

Demostración. **(a)** De (3.13) y (3.14), para cada $t \in \mathbb{N}$, tenemos

$$\begin{aligned}
||V^* - V_t||_W &\leq ||TV^* - T_tV^*||_W + ||T_tV^* - T_tV_{t-1}||_W,\\
&\leq ||TV^* - T_tV^*||_W + \alpha\beta||V^* - V_{t-1}||_W. \qquad (3.18)
\end{aligned}$$

Por otro lado, para cada $x \in \mathbb{X}$ y $t \in \mathbb{N}$,

$$\begin{aligned} |TV^*(x) - T_tV^*(x)| &\leq \sup_{a\in A(x)} |\alpha \sum_{s\in S} V^*[F(x,a,s)]\rho(s) - \alpha \sum_{s\in S} V^*[F(x,a,s)]\rho_t(s)|, \\ &\leq \sup_{a\in A(x)} \alpha \sum_{s\in S} V^*[F(x,a,s)]|\rho(s) - \rho_t(s)|. \end{aligned}$$

De (2.17) tenemos que $|V^*(x)| \leq \bar{c}W(x)/(1-\alpha\beta)$. Entonces

$$|TV^*(x) - T_tV^*(x)| \leq \frac{\alpha\bar{c}}{1-\alpha\beta} \sup_{a\in A(x)} \sum_{s\in S} W[F(x,a,s)]|\rho(s) - \rho_t(s)|.$$

Por (3.5),

$$||TV^* - T_tV^*||_W \leq \frac{\alpha\bar{c}}{1-\alpha\beta}||\rho - \rho_t||,$$

lo cual implica, por el Teorema 3.3.3 (d),

$$E||TV^* - T_tV^*||_W \leq \frac{\alpha\hat{c}}{1-\alpha\beta}E||\rho - \rho_t|| \to 0. \tag{3.19}$$

Ahora, sea $L_0 := \limsup_{t\to\infty} E||V^* - V_t||_W$. Observe que $L_0 < \infty$ ya que $V^*, V_t \in \mathbb{B}_W$ y $||V^* - V_t||_W \leq ||V^*||_W + ||V_t||_W < M$ para alguna constante $M > 0$. Entonces, tomando esperanza y límite superior cuando $t \to \infty$ en (3.18) obtenemos, por (3.19),

$$L_0 \leq \alpha\beta L_0.$$

Como $\alpha\beta < 1$, concluimos que $L_0 = 0$, lo cual demuestra (a).

(b) Para cada $t \in \mathbb{N}$, definimos la función $\phi_t : \mathbb{K} \to \mathbb{R}$ como

$$\phi_t(x,a) := c(x,a) + \alpha \sum_{s\in S} V_{t-1}[F(x,a,s)]\rho(s) - V_t(x), \quad (x,a) \in \mathbb{K}.$$

Observe que por (3.17), $\phi_t(x, f_t(x)) = 0 \quad \forall t \in \mathbb{N}, x \in \mathbb{X}$. Entonces

$$\begin{aligned} \phi(x, f_t(x)) &= |\phi(x, f_t(x)) - \phi_t(x, f_t(x))|, \\ &\leq \sup_{a\in A(x)} |\phi(x,a) - \phi_t(x,a)|. \end{aligned} \tag{3.20}$$

Por otro lado, sumando y restando el término $\alpha \sum_{s\in S} V_{t-1}[F(x,a,s)]\rho(s)$ tenemos

$$\begin{aligned}
|\phi(x,a)-\phi_t(x,a)| &\leq |V^*(x)-V_t(x)| \\
&\quad +|\alpha\sum_{s\in S} V^*[F(x,a,s)]\rho(s) - \alpha\sum_{s\in S} V_{t-1}[F(x,a,s)]\rho(s)|, \\
&\leq ||V^*-V_t||_W W(x) + \alpha\sum_{s\in S}|V^*[F(x,a,s)]-V_{t-1}[F(x,a,s)]|\rho(s) \\
&\quad +\alpha\sum_{s\in S} V_{t-1}[F(x,a,s)]|\rho_t(s)-\rho(s)|, \\
&\leq ||V^*-V_t||_W W(x) + \alpha||V^*-V_{t-1}||_W \sum_{s\in S} W[F(x,a,s)]\rho(s) \\
&\quad +\alpha||V_{t-1}||_W \sum_{s\in S} W[F(x,a,s)]|\rho_t(s)-\rho(s)|, \\
&\leq ||V^*-V_t||_W W(x) + \alpha\beta||V^*-V_{t-1}||_W W(x) \\
&\quad +\frac{\alpha\bar{c}}{1-\alpha\beta}\sum_{s\in S} W[F(x,a,s)]|\rho_t(s)-\rho(s)|. \qquad (3.21)
\end{aligned}$$

De aquí, para cada $(x,a)\in\mathbb{K}$,

$$\sup_{a\in A(x)}|\phi(x,a)-\phi_t(x,a)| \leq ||V^*-V_t||_W W(x)+\alpha\beta||V^*-V_{t-1}||_W W(x)+\frac{\alpha\bar{c}}{1-\alpha\beta}||\rho_t-\rho||W(x).$$

Entonces, por (3.20), el Teorema 3.3.3 (d) y el inciso anterior,

$$E\phi(x,f_t(x))\to 0 \quad \text{cuando} \quad t\to\infty.$$

Más aún, dividiendo entre $W(x)$ en ambos lados de la desigualdad (3.21), por (3.5), obtenemos

$$\sup_{(x,a)\in\mathbb{K}}\frac{|\phi(x,a)-\phi_t(x,a)|}{W(x)} \leq ||V^*-V_t||_W+\alpha\beta||V^*-V_{t-1}||_W+\frac{\alpha\bar{c}}{1-\alpha\beta}||\rho_t-\rho||. \qquad (3.22)$$

Tomando esperanza y combinando el Teorema 3.3.3 (d) con el inciso anterior, concluimos que

$$E\sup_{x\in\mathbb{X}}\frac{\phi(x,f_t(x))}{W(x)}\to 0,$$

lo cual demuestra la parte (b). ■

Apéndice A

Teorema de Punto Fijo

Definición A.0.1. *Sea $\mathbb{X}$ un espacio vectorial (real o complejo). Una norma en $\mathbb{X}$ es una función $||\cdot|| : \mathbb{X} \to \mathbb{R}$, cuyo valor en $x \in \mathbb{X}$ se denota por $||x||$, y satisface las siguientes propiedades:*

(i) $||x|| \geq 0$,

(ii) $||x|| = 0$ *sí y sólo si* $x = 0$,

(iii) $||\alpha x|| = |\alpha|||x||$,

(iv) $||x + y|| \leq ||x|| + ||y||$.

para todo $x, y, z \in \mathbb{X}$ y todo α escalar del campo. Al par $(\mathbb{X}, ||\cdot||)$ se le llama espacio normado.

Definición A.0.2. *Un espacio métrico es una pareja (S, d), donde S es un conjunto no vacío, y $d : S \times S \to \mathbb{R}$ es una función tal que para $x, y, z \in S$ arbitrarios satisface las siguientes propiedades:*

(i) $d(x, x) = 0$,

(ii) $d(x, y) > 0$ *si* $x \neq y$,

(iii) $d(x, y) = d(y, x)$,

(iv) $d(x, y) \leq d(x,\ z) + d(z, y)$ *(desigualdad del triángulo).*

A la función d se le conoce como métrica.

Teorema A.0.3. *Una norma en $\mathbb{X}$ define una métrica d en $\mathbb{X}$ dada por*

$$d(x, y) = ||x - y|| \quad \forall x, y \in X$$

y es llamada la métrica inducida por la norma.

Demostración. Verificaremos que la función d satisface las 4 propiedades de una métrica. Sean $x, y, z \in \mathbb{X}$ arbitrarios.

(i) De la Definiciónn A.0.1 (ii) tenemos

$$d(x, x) = ||x - x|| = ||0|| = 0.$$

(ii) Si $x \neq y$ entonces $x - y \neq 0$, y por la Definición A.0.1 (i) y (ii) se sigue

$$d(x, y) = ||x - y|| > 0.$$

(iii) De la propiedad (iii) en la Definición A.0.1 tenemos

$$\begin{aligned} d(x, y) &= ||x - y|| = ||(-1)(y - x)||, \\ &= = |-1|||y - x|| = ||y - x||, \\ &= d(y, x). \end{aligned}$$

(iv) De la Definición A.0.1 propiedad (iv) se sigue que

$$\begin{aligned} d(x, y) &= ||x - y|| = ||(x - z) + (z - y)||, \\ &\leq ||x - z|| + ||z - y||, \\ &= d(x, z) + d(z, y). \end{aligned}$$

■

Definición A.0.4. *Sea (S, d) un espacio métrico. Se dice que (S, d) es un espacio métrico completo si cualquier sucesión de Cauchy en S converge en S.*

Definición A.0.5. *Un espacio de Banach es un espacio normado completo con la métrica inducida por la norma.*

Sea X un espacio métrico. Denotemos como $\mathbb{B}$ al espacio de funciones acotadas, es decir, una función $u : \mathbb{X} \to \mathbb{R}$ pertenece a $\mathbb{B}$ si

$$||u|| := \sup_{x \in \mathbb{X}} |u(x)| < \infty.$$

l espacio de funciones acotadas $\mathbb{B}$ es un espacio de Banach. Para ver una demostración de este resultado consulte [7, Teorema 7.15, p. 151].

Sea $W : \mathbb{X} \to [1, \infty)$ una función aribitraria y denotemos por $\mathbb{B}_W$ al espacio de funciones W-acotadas, es decir, $u : \mathbb{X} \to \mathbb{R}$ pertenece a $\mathbb{B}_W$ si

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

Observemos que

$$||u||_W = ||u/W||. \tag{A.1}$$

A la función Wse le conoce como función de peso y a $|| \cdot ||$ norma ponderada.

Teorema A.0.6. *El espacio de las funciones acotadas $\mathbb{B}_W$ es un espacio de Banach.*

Demostración. Sea $\{u_n\}$ una sucesión de Cauchy en $\mathbb{B}_W$. Note que por (A.1), $\{u_n/W\}$ es una sucesión de Cauchy en $\mathbb{B}$. En efecto, ya que $\{u_n\}$ es de Cauchy, para toda $\epsilon > 0$, existe $N \in \mathbb{N}$ tal que para todo $n, m > N$

$$||u_n - u_m||_W < \epsilon,$$

pero

$$||u_n - u_m||_W = \sup_{x \in \mathbb{X}} \frac{|u_n(x) - u_m(x)|}{W(x)} = \sup_{x \in \mathbb{X}} |\frac{u_n(x)}{W(x)} - \frac{u_m(x)}{W(x)}| = ||\frac{u_n}{W} - \frac{u_m}{W}||,$$

por lo tanto

$$||\frac{u_n}{W} - \frac{u_m}{W}|| < \epsilon.$$

Como $\{u_n/W\}$ es de Cauchy en un espacio completo entonces converge a una función $\tilde{u} \in \mathbb{B}$, esto quiere decir que

$$||\tilde{u}|| < \infty,$$

pero esto implica que

$$\begin{aligned} ||\frac{u_n}{W} - \tilde{u}|| &\rightarrow 0, \\ ||\frac{u_n}{W} - \frac{\tilde{u}W}{W}|| &\rightarrow 0. \end{aligned} \tag{A.2}$$

Definamos $u(x) := \tilde{u}(x)W(x) \quad \forall x \in \mathbb{X}$. Notemos que $u \in \mathbb{B}_W$. En efecto,

$$|u(x)| = |\tilde{u}(x)W(x)| \leq MW(x).$$

De esto último y (A.2) obtenemos

$$||u_n - u||_W \rightarrow 0.$$

■

Definición A.0.7. *Sea $(S,\ d)$ un espacio métrico. Se dice que un operador*

$$T : S \rightarrow S$$

es de contracción módulo $\alpha \in (0, 1)$, si

$$d(Tx, Ty) \leq \alpha d(x, y) \quad \forall x, y \in S.$$

Teorema A.0.8. ***(Teorema del Punto Fijo de Banach)*** *Si (S, d) es un espacio métrico completo y $T : S \rightarrow S$ es un operador de contracción en S con módulo α, entonces:*

(a) *Existe un único $x \in S$ tal que*

$$Tx = x.$$

(b) *Para cada $x_0 \in S$,*

$$\lim_{n\rightarrow\infty} T^n x_0 = x.$$

Demostración. **(a)** Sea $x_0 \in S$ arbitrario, definimos la "sucesión iterativa" $\{x_n\}$ por

$$\begin{aligned} x_0 &\quad , \\ x_1 &= Tx_0, \\ x_2 &= Tx_1 = T^2x_0, \\ &\vdots \\ x_n &= T^nx_0, \\ &\vdots \end{aligned} \tag{A.3}$$

Probraremos que $\{x_n\}$ es de Cauchy. Como T es un operador de contracción, existe $\alpha \in (0,1)$ tal que para todas $a, b \in S$

$$d(Ta, Tb) \leq \alpha d(a, b). \tag{A.4}$$

De (A.3) y (A.4) se sigue

$$\begin{aligned} d(x_{m+1}, x_m) &= d(Tx_m, Tx_{m-1}), \\ &\leq \alpha d(x_m, x_{m-1}), \\ &= d(Tx_{m-1}, Tx_{m-2}), \\ &\leq \alpha^2 d(x_{m-1}, x_{m-2}), \\ &\vdots \\ &\leq \alpha^m d(x_1, x_0). \end{aligned} \tag{A.5}$$

Ahora bien, de la desigualdad del triángulo y la fórmula para la suma de una sucesión geométrica, para $n > m$ obtenemos

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \ldots + d(x_{n-1}, x_n), \\ &\leq (\alpha^m + \alpha^{m+1} + \ldots + \alpha^{n-1}) d(x_0, x_1), \\ &= \alpha^m \frac{1 - \alpha^{n-m}}{1 - \alpha} d(x_0, x_1). \end{aligned}$$

Debido a que $\alpha \in (0,1)$, tenemos que el numerador $1 - \alpha^{n-m} < 1$. Entonces,

$$d(x_m, x_n) \leq \frac{\alpha^m}{1 - \alpha} d(x_0, x_1). \tag{A.6}$$

Haciendo el lado derecho de (A.6) tan pequeño como se quiera tomando m suficientemente grande, obtenemos que $\{x_m\}$ es Cauchy. Como S es completo, $\{x_m\}$ converge a un límite x, es decir, $x_m \to x$.

Probaremos que x es un punto fijo de T. Por (A.4) y la desigualdad del triángulo tenemos que

$$\begin{aligned} d(x, Tx) &\leq d(x, x_m) + d(x_m, Tx), \\ &\leq d(x, x_m) + \alpha d(x_{m-1}, x), . \end{aligned} \tag{A.7}$$

Dado que $x_m \to x$, la suma (A.7) se puede hacer más pequeña que cualquier $\epsilon > 0$. Entonces $d(x,\ Tx) = 0$, y $Tx = x$. Por lo tanto x es un punto fijo de T.

Además, x es el único punto fijo de T ya que si hubieran dos puntos fijos, $Tx = x$ y $T\bar{x} = \bar{x}$, por (A.4) obtenemos

$$d(x, \bar{x}) = d(Tx, T\bar{x}) \leq \alpha d(x, \bar{x}),$$

lo cual implica que $d(x, \bar{x}) = 0$ pues $\alpha \in (0, 1)$, por lo tanto $x = \bar{x}$.

(b) En la demostración del inciso anterior probamos que para $x_0 \in S$ arbitrario, la sucesión $\{x_n \ := T^n x_0\}$ converge al único punto fijo x. Por lo tanto, $\lim_{n\to\infty} T^n y = x$ para cada $y \in S$. ■

Observación A.0.9. *Una consecuencia del teorema anterior es la siguiente:*

$$d(T^n x_0, x) \leq \alpha^n d(x_0, x), \quad \forall n \in \mathbb{N}. \tag{A.8}$$

Demostración. En efecto, primero observemos que

$$d(Tx_0, x) = d(Tx_0, Tx) \leq \alpha d(x_0, x).$$

Ahora, supongamos que (A.8) se cumple para $n = k$, es decir,

$$d(T^k x_0, x) \leq \alpha^k d(x_0, x). \tag{A.9}$$

Entonces

$$\begin{aligned} d(T^{k+1}x_0, x) &= d(T(T^k x_0), Tx), \\ &\leq \alpha d(T^k x_0, Tx), \\ &\leq \alpha^{k+1} d(x_0, x), \end{aligned}$$

donde la última desigualdad es por (A.9). Esto demuestra (A.8). ■

Apéndice B

Abreviaturas y símbolos

Abreviaturas

- EO: Ecuación de Optimalidad.
- *v.a.i.i.d.* : Variables Aleatorias Independientes Idénticamente Distribuidas.
- MCM : Modelo de Control Markoviano.
- PAO : Puntualmente Asintóticamente Óptima.
- PCO : Problema de Control Óptimo.
- W-PAO : W-uniforme Puntualmente Asintóticamente Óptima.

Símbolos

- $\mathbb{X}$: Espacio de estados.
- $\mathbb{A}$: Espacio de controles o acciones.
- $\mathbb{N}$: Conjunto de los números naturales.
- $\mathbb{N}_0$: $\mathbb{N} \cup \{0\}$.
- $\mathbb{B}_W$: Espacio lineal normado de todas las funciones $u : \mathbb{X} \to \mathbb{R}$ con norma $||u||_W$, definida como

$$||u||_W := \sup_{x \in \mathbb{X}} \frac{|u(x)|}{W(x)}$$

- $\mathbb{K}$: Conjunto de pares estado-acción admisibles definido como

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

- $\mathbb{H}_t$: Espacio de historias admisibles hasta la etapa $t \in \mathbb{N}_0$ definido como

$$\begin{aligned} \mathbb{H}_0 &:= \mathbb{X} \\ \mathbb{H}_t &:= \mathbb{K}^t \times \mathbb{X}, t \in \mathbb{N}. \end{aligned}$$

- $\mathbb{F}$: Es el conjunto de selectores definido como $\mathbb{F} := \{f : \mathbb{X} \to \mathbb{A} | f(x) \in A(x), x \in \mathbb{X}\}$.

Bibliografía

[1] Devroye, L., and Lugosi, G. *Combinatorial Methods in Density Estimation*. Springer, New York, 2012. 30

[2] Gordienko, E. I., and Minjárez-Sosa, J. A. Adaptive control for discrete-time markov processes with unbounded costs: discounted criterion. *Kybernetika 34*, 2 (1998), 217–234. 24

[3] Hernández-Lerma, O., and Lasserre, J. B. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York, 2012. 12

[4] Köthe, G. Topological vector spaces. In *Topological Vector Spaces I*. Springer, New York, 1983, pp. 123–201. 30

[5] Luque-Vásquez, F., Minjárez-Sosa, J. A., and Vega-Amaya, O. *Introducción a la Teoría de Control Estocástico*. Universidad de Sonora, 1996.

[6] Minjárez-Sosa, J. A. Estimación empírica en sistemas de control de markov. Por aparecer.

[7] Rudin, W. *Principles of Mathematical Analysis*. McGraw-Hill Publishing Co., 1976. 39