00384
4

UNIVERSIDAD NACIONAL AUTONOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS
FACULTAD DE CIENCIAS

PROCESOS DE CONTROL MARKOVIANOS: OPTIMALIDAD EN EL SENTIDO DE BLACKWELL

T E S I S

Que para obtener el grado académico de:

DOCTORA EN CIENCIAS (MATEMÁTICAS)

P r e s e n t a:

GUADALUPE CARRASCO LICEA

Director de tesis: DR. ONÉSIMO HERNÁNDEZ LERMA

México, D.F. Diciembre, 2002

A Martín

A Gabriel
y Alejandra

Por Rosalba

A todos los estudiantes universitarios que se negaron a entregar la Universidad a los dueños del dinero

A los cegeacheros que siguen defendiendo la educación gratuita en la UNAM

A mis amigos y compañeros: ***los luchos***

Agradezco

A Onésimo, por ser un gran maestro y por la honestidad reflejada en el apoyo al trabajo académico de alguien que piensa diferente a él.

A mis amigos *los controleros*: Adolfo, Daniel, Fernando, Juan, Oscar, Raquiel y Rigo, por las horas invertidas en discusiones sobre matemáticas y sobre todo lo demás y, en especial, por su apoyo en los días de encierro.

A Ivonne, por sus eternas dudas sobre qué sentido tiene estudiar posgrados, por su compañía en los congresos y, fundamentalmente, por ser una fuente inagotable de solidaridad.

A Ramón y Rebeca, por ser mis amigos, desde siempre y hasta siempre.

A mis papás y a mis hermanos, por todo. ¡Gracias familia!

Contenido

Capítulo 1

Introducción

En 1962, David Blackwell sugirió en su artículo [2] calificar a una política π^* como óptima si es $\alpha-$óptima bajo todo factor de descuento α cercano a 1, criterio que actualmente se conoce como optimalidad de Blackwell. Más precisamente, sea $\mathbb{P}$ el conjunto de todas las políticas y $V_\alpha(\pi, x) := E_x^\pi\left[\sum_{t=0}^{\infty}\alpha^t r(x_t, a_t)\right]$ la recompensa $\alpha-$descontada para una política $\pi \in \mathbb{P}$ dado que el estado inicial es x. Diremos que π^* es *Blackwell óptima* (BO) si existe $\alpha^*(x) \in (0,1)$ tal que

$$V_\alpha(\pi^*, x) - V_\alpha(\pi, x) \geq 0 \qquad \text{para toda } \pi \in \mathbb{P} \quad \text{y toda } \alpha \in (\alpha^*(x), 1) \tag{1.0.1}$$

En los siguientes capítulos de este trabajo, se exponen las definiciones precisas de los conceptos mencionados en el párrafo anterior así como otros conceptos que se encontrarán en el resto de la introducción. Aquí nos referimos a ellos sólo con la intención de dar una visión general del desarrollo del criterio de optimalidad de Blackwell.

Blackwell demostró la existencia de políticas que satisfacen (1.0.1) para modelos de control con espacios finitos de estados y de acciones. En 1969, Veinott [25] introdujo una escala infinita de criterios sensibles al factor de descuento, definidos de la siguiente manera. Para cada $n = -1, 0, 1...$, una política π^* es *$n-$descontada óptima* si

$$\liminf_{\alpha\uparrow 1} \mathbf{r}^{-n}\left(V_\alpha(\pi^*, x) - V_\alpha(\pi, x)\right) \geq 0 \qquad \text{para cada } \pi \in \mathbb{P},$$

donde $\mathbf{r} = (1-\alpha)/\alpha$ es la tasa de interés correspondiente al factor de descuento α.

El enfoque introducido por Veinott [25] y por Miller y Veinott [22] se basa en el uso de la expansión en serie de Laurent para la recompensa $\alpha-$descontada alrededor de $\alpha = 1$, que es de la forma

$$V_\alpha = (1+\mathbf{r})\left[\mathbf{r}^{-1}y_{-1} + \sum_{n=0}^{\infty}\mathbf{r}^n y_n\right].$$

Para modelos con espacios finitos de estados y de acciones, Veinott demostró que una política es BO en la clase de políticas estacionarias y deterministas si y sólo si es $(N-1)$−descontada óptima, donde N es el número de elementos del espacio de estados, y desarrolló un algoritmo para encontrar políticas n−descontadas óptimas. El enfoque basado en la serie de Laurent ha sido ampliamente utilizado en investigaciones acerca de estos criterios de optimalidad en modelos de control más generales.

En modelos con espacio numerable de estados, la definición de optimalidad de Blackwell requiere una modificación sugerida por Dekker y Hordijk [5] para obtener cierto tipo de resultados de existencia. La modificación consiste en que el intervalo donde la política BO mejora a otra política, depende no sólo del estado inicial sino también de la política con la que se está comparando. Es decir, en modelos cuyo espacio de estados no sea finito diremos que π^* es BO si para cada $\pi \in \mathbb{P}$ y cada estado inicial x, existe un intervalo de factores de descuento $(\alpha^*(x,\pi),1)$ en el que $V_\alpha(\pi^*,x) - V_\alpha(\pi,x) \geq 0$.

Para este tipo de modelos se tiene que una política es Blackwell óptima en el conjunto de políticas estacionarias y deterministas si y sólo si es n−descontada óptima para toda $n = -1, 0, 1, \ldots$, lo cual suele expresarse diciendo que se trata de una política ∞−descontada óptima.

Las contribuciones al desarrollo de este tipo de criterios en modelos con espacio de estados numerable incluyen los trabajos de Hordijk y Sladký [17], Dekker y Hordijk [5, 6], Lasserre [21] y Cavazos-Cadena y Lasserre [4] entre otros.

En su artículo de 1988, Dekker y Hordijk probaron la existencia de políticas estacionarias BO en modelos con recompensas no acotadas bajo la norma del supremo pero que tienen una norma ponderada acotada. Usaron conjuntos de acciones compactos así como probabilidades de transición y recompensas continuas respecto a las acciones. Para obtener la existencia de políticas BO en la clase de políticas estacionarias y deterministas, usaron la hipótesis de que todas las cadenas de Markov generadas por políticas estacionarias satisfacen una condición de ergodicidad geométrica uniforme en una norma ponderada. En el artículo de 1992 [6] sustituyeron esta hipótesis por otra condición de recurrencia uniforme que se puede verificar más fácilmente. Además demostraron que una política BO en la clase de las políticas estacionarias, es también BO en el conjunto de todas las políticas. Todas estas ideas han sido utilizadas en muchos de los trabajos posteriores y una parte de ellas se usará también en el trabajo que aquí presentamos.

Los trabajos de Yushkevich [28, 29, 30] iniciaron el estudio de los criterios de optimalidad sensible al factor de descuento para modelos cuyo espacio de estados es un espacio de Borel, es decir, un subconjunto de Borel de un espacio métrico, separable y completo. En su artículo de 1997, Yushkevich trabajó sobre conjuntos compactos de

acciones, con recompensas acotadas y continuas y supuso la existencia y continuidad de densidades de transición, que además satisfacen una condición tipo Doeblin. Esta condición le permitió garantizar la convergencia uniforme de la serie de Laurent para recompensas α–descontadas. Con este tipo de hipótesis demostró la existencia de políticas BO en el espacio de políticas estacionarias aleatorizadas con distribución inicial absolutamente continua respecto a una medida de referencia.

Uno de los trabajos más recientes sobre optimalidad de Blackwell, se debe a Hordijk y Yushkevich y fue publicado en dos partes debido a su extensión: [18] y [19]. En él, los autores dan condiciones para la existencia de políticas BO en el conjunto de todas las políticas para modelos con espacio de estados de Borel, conjuntos de acciones compactos y recompensas no acotadas. En la primera parte del trabajo usaron una topología débil-fuerte basada en funciones de Carathéodory para mostrar que el conjunto de políticas estacionarias aleatorizadas es compacto y presentaron condiciones que les permitieron garantizar la continuidad de los coeficientes de la serie de Laurent de la recompensa α–descontada. De esta manera, garantizaron la existencia de políticas BO en la subclase de políticas estacionarias aleatorizadas. En la segunda parte extienden este resultado a todas las políticas. Sus hipótesis incluyen la ergodicidad geométrica de las cadenas generadas por políticas estacionarias aleatorizadas, la continuidad absoluta de la distribución inicial respecto a una medida de referencia, la existencia y continuidad de densidades de transición y una condición de integrabilidad uniforme para la densidad de transición en n pasos.

Para ampliar la información acerca de los trabajos publicados sobre optimalidad de Blackwell y temas relacionados, se pueden consultar los artículos de Yushkevich [28, 30] así como la nota bibliográfica del capítulo 10 del libro de Puterman [24]. En [20] se encuentra una presentación bastante clara de la evolución del concepto detallando las hipótesis utilizadas por varios de los autores y esbozando las características de las demostraciones de los principales artículos publicados en los últimos 40 años sobre el tema.

El resultado central del presente trabajo consiste en utilizar otra vía para demostrar la existencia de políticas BO en el conjunto de todas las políticas, para modelos tan generales como los utilizados en los artículos de Hordijk y Yushkevich [18, 19]. El enfoque se estructura alrededor de la construcción de una sucesión infinita de modelos encajados en el sentido de que el conjunto de acciones admisibles en un estado x para el modelo n, contiene al conjunto de acciones admisibles en el mismo estado para el modelo $n+1$.

Para lograr este objetivo, caracterizaremos la optimalidad de Blackwell en términos de la optimalidad promedio, la cual se define de la siguiente manera: la recompensa promedio esperada obtenida al aplicar una política π cuando el estado inicial es x, está

dada por

$$J(\pi, x) := \liminf_{N\to\infty} \frac{1}{N} E_x^{\pi}\left[\sum_{t=0}^{N-1} r(x_t, a_t)\right],$$

y una política π^* es promedio óptima si $J(\pi^*, x) := \sup_{\pi\in\mathbb{P}} J(\pi, x)$. Este criterio se relaciona directamente con los criterios sensibles al factor de descuento. De hecho, la optimalidad $-1-$descontada es equivalente a la optimalidad promedio en modelos numerables y en este sentido, desde los trabajos de Veinott, se ha podido interpretar la optimalidad de Blackwell como un criterio límite de una serie infinita de criterios que inicia con la optimalidad promedio. En este trabajo veremos que también es posible interpretarla como resultado de aplicaciones sucesivas de optimalidad promedio a una colección infinita $\{\mathfrak{M}_0, \mathfrak{M}_1, \mathfrak{M}_2, \dots\}$ de modelos encajados muy generales, donde el concepto límite se aplica a los modelos y no al criterio utilizado. En otras palabras, se muestra que una política es BO si y sólo si sus acciones maximizan las ecuaciones de optimalidad para recompensa promedio correspondientes a los modelos de la sucesión.

El camino que aquí presentamos para garantizar la existencia de políticas BO, no sólo resulta más directo que el desarrollado en los artículos [18, 19], sino que además utiliza hipótesis menos restrictivas.

El tipo de construcción que presentamos aquí fue utilizada por Cavazos–Cadena y Lasserre en su artículo [4] para demostrar la existencia de políticas BO en modelos numerables. Sus hipótesis incluyen las condiciones usuales de continuidad y compacidad además de una condición de recurrencia fuerte para las cadenas generadas por políticas estacionarias y deterministas. Dicha condición consiste en que para cada política estacionaria y determinista $\pi = (f, f, f, \dots)$ existe un estado $z(f)$ tal que el tiempo esperado de arribo a él desde cualquier estado inicial, esté acotado. Además, requieren que la colección de medidas invariantes de las cadenas de Markov correspondientes a políticas estacionarias, sea tensa. Este artículo de Cavazos-Cadena y Lasserre motivó nuestra investigación y sirvió de guía para construir la demostración de un resultado similar en modelos más generales.

La idea subyacente en la construcción de modelos encajados no es nueva. En la formulación de las ecuaciones de optimalidad para los criterios $n-$descontados, que aparecen implícitamente en [25] y explícitamente en [5], se consideran conjuntos de acciones anidados en el sentido de que se buscan acciones que maximicen cada ecuación sólo entre aquellas acciones que maximizaron la ecuación anterior. En el trabajo [8], Federgruen y Schweitzer propusieron un método para resolver sistemas finitos de ecuaciones funcionales anidadas, del tipo de las ecuaciones de optimalidad para modelos numerables, pero no investigaron su implementación en el estudio de procesos de control Markovianos.

En nuestra presentación decidimos utilizar funciones de costo por etapa en lugar

de recompensas. Evidentemente, los resultados obtenidos son equivalentes en modelos con recompensas.

La estructura del trabajo es la siguiente: en el capítulo 2 incluimos los conceptos básicos de procesos de Markov controlados. El capítulo 3 incluye las hipótesis que serán utilizadas, la construcción de la sucesión de modelos y la demostración del resultado principal así como de todos los resultados preliminares que se requieren. La última sección de este capítulo está dedicada a hacer una comparación detallada entre las hipótesis requeridas y los procedimientos usados en los artículos [18] y [19], y en este trabajo Finalmente, el capítulo 4 está destinado a presentar un ejemplo que satisface las hipótesis que requerimos para la existencia de políticas BO en el que, además, encontramos la única política óptima en el sentido de Blackwell que existe en ese modelo.

Capítulo 2

Procesos de Markov controlados

2.1 Introducción

En este capítulo expondremos brevemente las ideas básicas acerca de los procesos de control Markovianos. El objetivo que perseguimos es especificar la clase de modelos de control, las clases de políticas de control y los criterios de optimalidad que serán necesarios para el desarrollo posterior del trabajo. Al mismo tiempo, se introduce la notación y terminología que usaremos en adelante.

2.2 Modelos de control Markoviano

Un modelo de control Markoviano (MCM) es la representación matemática de cierto tipo de sistemas dinámicos cuya evolución está afectada por elementos estocásticos. Aquí consideraremos modelos estacionarios a tiempo discreto compuestos por 5 elementos:

$$\mathfrak{M} := (X, A, \{A(x) : x \in X\}, Q, c). \tag{2.2.1}$$

(a) X es el espacio de estados del sistema dinámico.

(b) A es el espacio de acciones o controles de que se dispone.

Supondremos que X y A son espacios de Borel, es decir, subconjuntos de Borel (no vacíos) de espacios métricos, separables y completos. Denotaremos las σ–álgebras de Borel correspondientes a estos espacios por $\mathfrak{B}(X)$ y $\mathfrak{B}(A)$.

(c) $A(x)$ es el subconjunto de A que contiene las acciones admisibles cuando el sistema se encuentra en el estado x. El conjunto de parejas viables estado–acción se denota por

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

el cual supondremos que es un subconjunto de Borel de $X \times A$ y que contiene la gráfica de una función medible $f : X \to A$. Cuando hablemos de medibilidad de funciones y conjuntos, nos referiremos siempre a medibilidad con respecto a la σ–ágebra de Borel correspondiente.

(d) Q es una probabilidad de transición sobre X dado $\mathbb{K}$, es decir, Q satisface las condiciones:

(d.1) $Q(\cdot \mid x, a)$ es una medida de probabilidad en X para cada (x, a) en $\mathbb{K}$,

(d.2) $Q(B \mid \cdot\,, \cdot\,)$ es una función medible en $\mathbb{K}$ para cada $B \in \mathfrak{B}(X)$.

(e) c es una función medible definida en $\mathbb{K}$ que representa el costo por etapa.

Denotaremos por x_t al estado en que se encuentra el sistema al tiempo t y por a_t a la acción que se aplica en ese mismo tiempo. La evolución del sistema se da de la siguiente forma: supongamos que al tiempo t se observa que el sistema se encuentra en el estado x, es decir, $x_t = x$; si se elige la acción $a \in A(x)$, entonces se genera un costo $c(x, a)$ y el sistema pasa a ocupar un nuevo estado, que se observará en el tiempo $t + 1$, de acuerdo a una medida de probabilidad $Q(\cdot \mid x, a)$, es decir,

$$Q(B \mid x, a) = Pr[x_{t+1} \in B \mid x_t = x, a_t = a].$$

Supongamos que el nuevo estado al que arribó el sistema es $x_{t+1} = x'$. Al aplicar una nueva acción $a_{t+1} = a'$ se genera un nuevo costo y el sistema pasa a ocupar un nuevo estado que será observado al tiempo $t+2$. Este proceso se repite hasta un cierto tiempo finito N o indefinidamente. Los modelos que usaremos en el presente trabajo son de horizonte infinito, lo que significa que la repetición del proceso descrito será indefinida.

La elección del estado inicial x_0 puede hacerse aplicando una distribución de probabilidad inicial ν, es decir, $Pr[x_0 \in B] = \nu(B)$, con $B \in \mathfrak{B}(X)$. En particular, ν puede ser la medida de Dirac concentrada en un punto $x \in X$.

2.3 Políticas de control

De manera general, podemos decir que una política de control es una regla para elegir acciones en cada tiempo t. Esta elección de acciones representa la forma en que un observador puede influir en la evolución del sistema dinámico. El espacio de historias admisibles hasta el tiempo t se define de la siguiente manera:

$$H_0 := X, \quad \text{y} \quad H_t := \mathbb{K}^t \times X \quad \text{para} \quad t \in \mathbb{N},$$

donde $\mathbb{N} = \{1, 2, \dots\}$. Un elemento de H_t es de la forma

$$h_t := (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t) \quad \text{donde} \quad a_k \in A(x_k) \ \text{ para } \ k = 0, 1, ..., t-1,$$

y lo llamaremos una $t-$historia.

Denotaremos por $\mathbb{F}$ el conjunto de funciones medibles $f : X \to A$ que satisfacen que $f(x)$ está en $A(x)$ para toda $x \in X$. Nos referiremos a los elementos de $\mathbb{F}$ como *funciones de decisión.* La hipótesis de que $\mathbb{K}$ contiene a la gráfica de una función medible de X en A nos asegura que $\mathbb{F}$ es un conjunto no vacío.

La clase más general de políticas está formada por las políticas en las que la elección de la acción se hace en forma aleatoria y considerando la $t-$historia ocurrida para cada t. Cada una de estas políticas es una sucesión $\pi = \{\pi_t\}$ de distribuciones de probabilidad en A dado H_t, en la cual cada ditribución condicional π_t satisface la restricción:

$$\pi_t(A(x_t) \mid h_t) = 1, \qquad h_t \in H_t \ , \ t \in \mathbb{N}_0,$$

donde $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. El conjunto de tales políticas será denotado por $\mathbb{P}$. Si la elección de la acción depende exclusivamente del estado en que se encuentra el sistema, es decir, si cada distribución de probabilidad en la definición anterior satisface que

$$\pi_t(\ \cdot \mid h_t) = \pi_t(\ \cdot \mid x_t), \qquad h_t \in H_t, \quad t \in \mathbb{N}_0,$$

entonces π se conoce como una *política Markoviana* y el conjunto de tales políticas se designará por Φ. Cuando la elección de la acción se hace en forma determinista, es decir, cuando cada distribución π_t se reduce a una función f_t en $\mathbb{F}$, la política será llamada determinista o, más precisamente, *determinista y Markoviana.* Un subconjunto especial de las políticas deterministas es el de aquellas en las que se usa la misma función f en todo t, es decir, $\pi = \{f, f, f, ...\}$. A este último tipo de políticas se les conoce como políticas *deterministas y estacionarias.* Identificaremos el conjunto de estas políticas con el conjunto de funciones de decisión en X en virtud de que cada elemento de $\mathbb{F}$ genera una política estacionaria y determinista y, recíprocamente, cada política de este tipo está generada por un elemento de $\mathbb{F}$.

Consideremos el espacio medible $(\Omega, \mathcal{A})$, donde $\Omega := (X \times A)^\infty$ y $\mathcal{A}$ es la $\sigma-$álgebra producto correspondiente. Sea $\pi \in \mathbb{P}$ una política de control y supongamos que el estado inicial es x. El Teorema de Ionescu-Tulcea ([1] Teorema 2.7.2, p.109) garantiza la existencia de una medida de probabilidad P_x^π definida sobre $(\Omega, \mathcal{A})$, que satisface las siguientes propiedades para cada $t \in \mathbb{N}_0$

$$\begin{aligned} P_x^\pi[x_0 = x] &= 1; \\ P_x^\pi[a_t \in C \mid h_t] &= \pi_t(C \mid h_t), \quad \forall\, C \in \mathfrak{B}(A); \\ P_x^\pi[x_{t+1} \in B \mid h_t, a_t] &= Q(B \mid x_t, a_t) \quad \forall\, B \in \mathfrak{B}(X). \end{aligned}$$

Al proceso estocástico $\{\Omega, \mathcal{A}, P_x^\pi, \{x_t\}\}$ lo llamaremos *proceso de control Markoviano* (PCM). En particular, si π es una política determinista y estacionaria, el proceso de estados generado $\{x_t\}$ es de Markov respecto a la medida P_x^π (ver [11]).

El operador esperanza respecto a P_x^π será denotado por E_x^π.

2.4 Criterios de optimalidad

Ahora nos proponemos fijar criterios para evaluar el desarrollo del sistema. Para ello definimos a continuación los índices de funcionamiento que serán de interés en este trabajo.

Definición 2.4.1 *Sea $x \in X$ un estado arbitrario y $\pi \in \mathbb{P}$ cualquier política.*

1. *El costo total esperado al aplicar la política π dado que el estado inicial es x, se define como*

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

cuando se calcula para N etapas y se tiene un costo terminal igual a cero, y como

$$V_1(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} c(x_t, a_t) \right]$$

cuando el cálculo se hace a lo largo de todas las etapas de un modelo con horizonte infinito. Denotaremos el costo total esperado óptimo para un estado inicial x por

$$V_1^*(x) = \inf_{\pi \in \mathbb{P}} V_1(x, \pi).$$

2. *El costo esperado $\alpha-$descontado cuando se aplica la política π, dado el estado inicial x, está dado por*

$$V_\alpha(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right], \tag{2.4.1}$$

donde α es un número en el intervalo $(0,1)$ llamado factor de descuento. Denotamos el costo $\alpha-$descontado óptimo para un estado inicial x por:

$$V_\alpha^*(x) := \inf_{\pi \in \mathbb{P}} V_\alpha(x, \pi)$$

3. *El costo promedio esperado bajo la política π dado el estado inicial x es*

$$J(\pi, x) := \limsup_{N \to \infty} \frac{1}{N} V_{N,1}(\pi, x) = \limsup_{N \to \infty} \frac{1}{N} E_x^\pi \left[\sum_{t=0}^{N-1} c(x_t, a_t) \right]. \tag{2.4.2}$$

Diremos que el costo promedio óptimo es

$$J^*(x) := \inf_{\pi \in \mathbb{P}} J(x, \pi).$$

A la pareja formada por un modelo de control $\mathfrak{M}$ y un índice de funcionamiento, se le llama problema de control estocástico Markoviano. El objetivo que se persigue es encontrar políticas π^* que conduzcan al valor óptimo de alguno de los índices, es decir, políticas óptimas de acuerdo a un criterio preestablecido. Así pues, dado un estado inicial $x \in X$, diremos que

(a) $\pi^* \in \mathbb{P}$ es una política costo total óptima si

$$V_1(\pi^*, x) = V_1^*(x).$$

(b) $\pi^* \in \mathbb{P}$ es una política $\alpha-$descontada óptima si

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

(c) $\pi^* \in \mathbb{P}$ es una política promedio óptima si

$$J(\pi^*, x) = J^*(x). \tag{2.4.3}$$

De las definiciones de los índices puede verse que en el criterio de costo total esperado todos los costos por etapa tienen el mismo peso, pero es fácil encontrarse con una serie divergente. En el artículo [12] se presenta un compendio amplio de resultados sobre este criterio de optimalidad.

En el criterio de optimalidad $\alpha-$descontada se da un peso mayor a los costos generados en las primeras etapas debido a que α^t tiende a cero geométricamente haciendo que los costos por etapa pierdan peso progresivamente conforme t crece.

En contraposición, el criterio promedio esperado depende sólo del comportamiento asintótico de los costos por etapa, sin importar lo que haya ocurrido en las primeras etapas. De hecho, al estar este último criterio enteramente determinado por el comportamiento límite del sistema cuando t tiende a infinito, dos políticas que generen, por ejemplo, las sucesiones de costos $(1000, 0, 0, \ldots)$ y $(0, 0, 0, \ldots)$ tienen el mismo costo promedio esperado. En la literatura sobre control estocástico a tiempo discreto, se encuentran diversos caminos para suavizar esta naturaleza poco selectiva de la optimalidad promedio. Por lo general, se trabaja en subclases del conjunto de políticas promedio óptimas; por ejemplo, en el artículo [15] se consideran políticas promedio óptimas que además minimizan la varianza de los costos.

2.5 Normas ponderadas

En virtud de que en este trabajo vamos a usar funciones de costo no acotadas bajo la norma del supremo, requerimos otro tipo de normas. Sea $\omega : X \to [1, \infty)$ una función medible a la que nos referiremos como función de peso. Si u es una función en X de valores reales, definimos la norma–ω de u como

$$\|u\|_\omega := \left\|\frac{u}{\omega}\right\| = \sup_{x \in X} \frac{|u(x)|}{\omega(x)}. \tag{2.5.1}$$

Si $\omega(\cdot) \equiv 1$, la norma–ω coincide con la norma del sup. En general, como $\omega(\cdot) \geq 1$, se tiene que

$$\|u\|_\omega \leq \|u\|,$$

así que las funciones acotadas bajo la norma del sup son también acotadas bajo la norma–ω. Lo contrario no necesariamente sucede. Si $\|u\|_\omega < k$ entonces $|u(x)| < k\omega(x)$ para toda $x \in X$, pero ω puede ser una funcion no acotada bajo la norma del sup.

Al espacio de Banach formado por las funciones medibles en X con norma–ω finita lo denotaremos por $\mathbb{B}_\omega(X)$.

Cuando se trata con medidas finitas (con signo), lo usual es tomar la norma de variación total dada por

$$\|\mu\|_{VT} := \sup_{\|v\| \leq 1} \left| \int_X v \, d\mu \right| = |\mu|(X),$$

donde $|\mu| = \mu^+ + \mu^-$ denota la variación total de μ. Por analogía, la norma–ω de una medida μ finita y con signo en $\mathfrak{B}(X)$, será:

$$\|\mu\|_\omega := \sup_{\|v\|_\omega \leq 1} \left| \int_X v \, d\mu \right| = \int_X \omega \, d|\mu|$$

que se reduce a la norma anterior si $\omega(\cdot) \equiv 1$. En general, como $\omega(\cdot) \geq 1$ se tiene

$$\|\mu\|_\omega \geq \|\mu\|_{VT}.$$

Capítulo 3

Optimalidad en el sentido de Blackwell

3.1 Introducción

El criterio de optimalidad de Blackwell tiene la característica de que contempla tanto los costos generados en las primeras etapas como los asintóticos, sin las desventajas que presenta el costo total esperado.

Definición 3.1.1 *Sea $x \in X$ un estado inicial. Diremos que π^*es una política Blackwell óptima (BO), si para cada $\pi \in \mathbb{P}$ existe un número $\alpha^*(x,\pi) \in (0,1)$ tal que*

$$V_\alpha(\pi, x) - V_\alpha(\pi^*, x) \geq 0 \quad \textit{para toda} \quad \alpha \in (\alpha^*(x,\pi), 1). \tag{3.1.1}$$

Mientras más cerca esté α de 1 más peso se le da a los costos a largo plazo. Al mismo tiempo, en la medida que se utiliza el costo α–descontado para comparar el efecto de las políticas, no se dejan de contemplar los primeros costos generados.

De aquí en adelante, incluiremos explícitamente la función costo en la notación de los índices y los valores óptimos para evitar confusiones al utilizar distintas funciones de costo por etapa. Así, el costo esperado α–descontado asociado a una función de costo c, bajo una política π y dado el estado inicial x, será denotado por $V_\alpha(\pi, c, x)$, y el correspondiente costo esperado α–descontado óptimo será denotado $V_\alpha^*(c, x)$. Análogamente, el costo promedio esperado dado el estado inicial x, bajo la política π y usando la función de costo c, será denotado por $J(\pi, c, x)$, y el costo promedio óptimo correspondiente, por $J^*(c, x)$.

3.2 Hipótesis

Como mencionamos en la introducción, nos proponemos caracterizar la optimalidad de Blackwell en términos de optimalidad promedio. Requerimos entonces condiciones que nos garanticen la existencia de políticas promedio óptimas en modelos cuyos espacios de estados y acciones sean de Borel, con funciones de costo no necesariamente acotadas. La formulación menos restrictiva que conocemos de este tipo de condiciones es la dada en [16] que presentaremos en esta sección. Se trata de dos conjuntos de hipótesis sobre el modelo $\mathfrak{M}$ introducido en (2.2.1). El primero de ellos contiene las hipótesis usuales sobre continuidad, compacidad y cotas para modelos generales como el que estamos analizando.

Hipótesis 3.2.1 *Para cada estado $x \in X$:*

(a) *El conjunto $A(x)$ es compacto no vacío.*

(b) *La función costo por etapa $c(x,a)$ es continua en $A(x)$.*

(c) *La función $a \longmapsto \int_X v(y)\, Q(dy \mid x,a)$ es continua en $A(x)$ para cada función $v : X \to \mathbb{R}$ continua y acotada (con la norma del sup).*

(d) *Además, existe una función de peso $\omega \geq 1$ en X, y una constante $\overline{c} \geq 0$ tal que, para cada $x \in X$*

(d_1) $\sup_{a \in A(x)} |c(x,a)| \leq \overline{c}\,\omega(x)$;

(d_2) *la función $a \longmapsto \int_X \omega(y)\, Q(dy \mid x,a)$ es continua en $A(x)$.*

Observación 3.2.2 *La hipótesis 3.2.1(d_1) se puede escribir como*

$$\|\widehat{c}\|_\omega \leq \overline{c}$$

donde $\widehat{c} = \sup_{a \in A(x)} |c(x,a)|$. Obsérvese que $\widehat{c}$ es una función en X de valores reales a la que se le puede aplicar la norma–ω.

Una hipótesis usual para modelos del tipo que aquí consideramos, es pedir que todas las cadenas de Markov generadas por políticas estacionarias satisfagan una condición de ergodicidad geométrica uniforme (ver la desigualdad (3.2.2) en la siguiente página) y alguna condición fuerte de recurrencia. En lugar de ello, aquí pediremos hipótesis más débiles que garantizan que dichas condiciones se cumplen, basándonos en un resultado presentado por O. Vega-Amaya en su artículo [27] y en otros resultados que aparecen en [9]. Estas hipótesis son:

Hipótesis 3.2.3 *Existe una medida no trivial ν en X, una función medible no-negativa l en $\mathbb{K}$ y una constante positiva $K < 1$, tales que*

(a) $Q(B \mid x,a) \geq l(x,a)\,\nu(B)$ *para toda* $B \in \mathfrak{B}(X)$ *y toda pareja* $(x,a) \in \mathbb{K}$.

(b) $\nu(\omega) := \int_X \omega\, d\nu < \infty$, *donde* ω *es la función de peso dada en la Hipótesis 3.2.1(d)*

(c) Para toda pareja $(x,a) \in \mathbb{K}$

$$\int_X \omega(y)\, Q(dy \mid x,a) \leq K\omega(x) + l(x,a)\,\nu(\omega). \tag{3.2.1}$$

(d) Para cada $f \in \mathbb{F}$, $\nu(l_f) := \int_X l_f\, d\nu > 0$, *donde* $l_f := l(x, f(x))$.

También vamos a requerir la siguiente forma de la Hipótesis 3.2.3(d), más fuerte que la anterior:

Hipótesis 3.2.4 *Existe una constante $\gamma > 0$ tal que $\nu(l_f) > \gamma$ para toda $f \in \mathbb{F}$, con ν y l como en la Hipótesis 3.2.3(d).*

Las Hipótesis 3.2.3 y 3.2.4 garantizan los siguientes hechos:

Proposición 3.2.5 *Si se cumple la Hipótesis 3.2.3, entonces para cada $f \in \mathbb{F}$:*

(i) La cadena de Markov definida por $Q_f(\cdot \mid \cdot)$ es $\nu-$irreducible y Harris recurrente positiva con una única medida de probabilidad invariante μ_f.

(ii) $\mu_f(\omega) < \infty$.

(iii) Si además se cumple la Hipótesis 3.2.4, entonces la cadena de Markov del inciso (i) es geométricamente ergódica con norma$-\omega$; es decir, existen constantes R y β, con $R \geq 0$ y $0 < \beta < 1$, tales que

$$\sup_{f \in \mathbb{F}} \left| \int_X u(y)\, Q_f^t(dy \mid x) - \mu_f(u) \right| \leq R\beta^t \|u\|_\omega\, \omega(x) \tag{3.2.2}$$

para toda $x \in X$, $u \in \mathbb{B}_\omega(X)$ y $t \in \mathbb{N}_0$.

Demostración. Los incisos *(i)* y *(ii)* se siguen del Teorema 3.3 del artículo [27], mientras que la parte *(iii)* es una consecuencia de los incisos anteriores y de los Lemas 3.3 y 3.4 en [9]. ■

Las hipótesis que hemos enlistado y la Proposición 3.2.5 permiten obtener una serie de resultados importantes acerca de la optimalidad promedio. En la siguiente proposición incluimos varios de esos resultados para facilitar la referencia a ellos. La demostración se puede encontrar en el capítulo 10 del libro [14].

Proposición 3.2.6 *Supóngase que $\mathfrak{M}$ es un modelo definido como en (2.2.1) que satisface las hipótesis 3.2.1, 3.2.3. y 3.2.4. Entonces existe una constante $\rho_0 \in \mathbb{R}$ y una función $h_1 \in \mathbb{B}_\omega(X)$ tal que*

(i) $\rho_0 = J^*(c,x) := \inf_{\pi\in\mathbb{P}} J(\pi,c,x)$

(ii) $\|h_1\|_\omega \leq D\,\|\widehat{c}\|_\omega$ *donde* $D = R/(1-\beta)$ *y* R *y* β *son las constantes que aparecen en la desigualdad (3.2.2)*

(iii) ρ_0 *y* $-h_1$ *satisfacen la Ecuación de Optimalidad para Costo Promedio (EOCP), es decir,*

$$\rho_0 - h_1(x) = \inf_{A(x)}\left[c(x,a) - \int_X h_1(y)\,Q(dy\,|\,x,a)\right] \tag{3.2.3}$$

(iv) Para cada $x \in X$ el lado derecho de la EOCP dada en (3.2.3), tiene un minimizador f y la política estacionaria generada por f, es promedio óptima para la función de costo c.

3.3 Modelos encajados y el resultado principal

En esta sección introducimos una sucesión $\{\mathfrak{M}_n\}$ de procesos de control Markovianos definidos recursivamente, de tal manera que el conjunto de acciones admisibles en el modelo $\mathfrak{M}_{n+1}$ está contenido en el conjunto de acciones admisibles en el modelo $\mathfrak{M}_n$ para todo estado $x \in X$, y toda $n \in \mathbb{N}_0 = \{0,1,2,\dots\}$. En este sentido, hablaremos de una sucesión de modelos "encajados". El próposito de esto es caracterizar a las políticas BO en términos de políticas que satisfacen la EOCP en estos modelos.

En el resto de este capítulo, supondremos que se cumplen las Hipótesis 3.2.1, 3.2.3 y 3.2.4.

Observación 3.3.1 *a) Para lo que sigue, es importante hacer notar que las hipótesis 3.2.1(c) y (d_2) implican que la función $a \mapsto \int_X v(y)Q(dy\,|\,x,a)$ es continua en $A(x)$ para cada función $v \in \mathbb{B}_\omega(X)$. (Ver el Lema 8.3.7(a) en [14]).*
b) De la demostración del lema mencionado en (a) es claro que si $c(\cdot,\cdot)$ es no-negativa, entonces la hipótesis 3.2.1(d_2) no se requiere.

La construcción de la sucesión $\{\mathfrak{M}_n\}$, se obtiene aplicando los siguientes cuatro pasos:

1. Sea $\mathfrak{M}_0 := \mathfrak{M}$ el modelo original introducido en el Capítulo 2, tomando $h_0 = c$, $A_0(x) = A(x)\ \ \forall\ x$, y $\mathbb{K}_0 = \mathbb{K}$.

2. Sean $\rho_0 \in \mathbb{R}$ y $-h_1 \in \mathbb{B}_w(X)$ la constante y la función que satisfacen la EOCP asociada a $\mathfrak{M}_0$, es decir (como en (3.2.3)):

$$\rho_0 - h_1(x) = \inf_{a \in A_0(x)} \left[c(x,a) - \int_X h_1(y)\, Q(dy \,|\, x,a) \right], \qquad x \in X,$$

donde podemos suponer que $\|h_1\|_w \leq D\|c\|_w$ siendo D la constante que aparece en la Proposición 3.2.6(ii).

3. Definimos la función discrepancia como en [23], es decir:

$$\Phi_0(x,a) := c(x,a) - \int_X h_1(y) Q(dy \,|\, x,a) - \rho_0 + h_1(x), \qquad (x,a) \in \mathbb{K}_0. \quad (3.3.4)$$

La ecuación de optimalidad en el paso 2 implica que $\Phi_0 \geq 0$. Como $h_1 \in \mathbb{B}_w(X)$ y $c(x, \cdot\,)$ es continua en $A_0(x)$, entonces $\Phi_0(x, \cdot\,)$ es continua en $A_0(x)$ para cada $x \in X$ de acuerdo a la observación 3.3.1. Además, la parte (iii) de la Proposición 3.2.6 nos permite afirmar que $\min_{a \in A_0(x)} \Phi_0(x,a) = 0$ para cada $x \in X$. Entonces, el conjunto

$$A_1(x) := \{a \in A_0(x) \,|\, \Phi_0(x,a) = 0\} \qquad (3.3.5)$$

es un subconjunto cerrado y no vacío de $A_0(x)$. Como $A_0(x)$ es compacto para toda x, también lo es $A_1(x)$. Nótese que $h_1(\,\cdot\,)$ satisface las partes (b) y (d_1) de la Hipótesis 3.2.1, así que puede ser usada como una función costo. Definimos el modelo

$$\mathfrak{M}_1 := (X, A, \{A_1(x) \,|\, x \in X\}, Q, h_1).$$

4. Dado el modelo $\mathfrak{M}_n := (X, A, \{A_n(x) | x \in X\}, Q, h_n)$ para una $n \geq 1$, donde $h_n \in \mathbb{B}_w(X)$, y cada $A_n(x)$ es un subconjunto no vacío y compacto de A, construimos el modelo $\mathfrak{M}_{n+1}$ de la siguiente forma. Sean $\rho_n \in \mathbb{R}$ y $-h_{n+1} \in \mathbb{B}_w(X)$ la constante y la función que satisfacen la EOCP asociada al modelo $\mathfrak{M}_n$, es decir,

$$\rho_n - h_{n+1}(x) = \inf_{a \in A_n(x)} \left[h_n(x) - \int_X h_{n+1}(y) Q(dy \,|\, x,a) \right], \quad x \in X,$$

donde h_{n+1} satisface que $\|h_{n+1}\|_w \leq D\|h_n\|_w$ debido a la parte (ii) de la Proposición 3.2.6. Sea $\mathbb{K}_n := \{(x,a) | x \in X,\ a \in A_n(x)\}$ y definamos la función discrepancia Φ_n asociada a $\mathfrak{M}_n$ por

$$\Phi_n(x,a) := h_n(x) - \int_X h_{n+1}(y) Q(dy \,|\, x,a) - \rho_n + h_{n+1}(x), \quad (x,a) \in \mathbb{K}_n. \quad (3.3.6)$$

Como antes, tenemos que $\Phi_n(x, \cdot\,)$ es una función no negativa y continua en $A_n(x)$, y para alguna $a \in A_n(x)$, $\Phi_n(x,a) = 0$. Combinando este hecho con la compacidad de $A_n(x)$, se sigue que los conjuntos

$$A_{n+1}(x) := \{a \in A_n(x) \,|\, \Phi_n(x,a) = 0\}, \quad x \in X, \qquad (3.3.7)$$

son no vacíos y compactos. Entonces, el modelo $\mathfrak{M}_{n+1}$ queda definido como

$$\mathfrak{M}_{n+1} := (X, A, \{A_{n+1}(x) \mid x \in X\}, Q, h_{n+1}).$$

Observación 3.3.2 *En la construcción anterior se tiene que, como* $\|h_{n+1}\|_w \leq D\|h_n\|_w$ *con* $D = R/(1-\beta)$ *(ver la Proposición 3.2.6(ii)), se cumple que*

$$\|h_{n+1}\|_w \leq D^n\|h_1\|_w \leq D^{n+1}\|\widehat{c}\|_\omega. \tag{3.3.8}$$

y como $\widehat{c} \in \mathbb{B}_\omega(X)$, *se tiene que todas las funciones* h_n, $n \in \mathbb{N} = \{1, 2, \dots\}$, *satisfacen la condición 3.2.1*(d_1)

Para continuar, definimos los siguientes conceptos que usaremos en las secciones posteriores.

Definición 3.3.3 *Para cada* $x \in X$ *y* $n \in \mathbb{N}$, *sea*

$$\mathbb{P}_n(x) := \{\pi \in \mathbb{P} \mid P_x^\pi[a_t \in A_n(x_t)] = 1 \ \text{ para toda } t \in \mathbb{N}_0,$$

donde a_t *es la acción indicada por la política* π *al tiempo* t, *cuando el sistema se encuentra en el estado* x_t.

Esto es, $\mathbb{P}_n(x)$ es el conjunto de políticas que satisfacen la EOCP para el modelo $\mathfrak{M}_{n-1}$ P_x^π−casi seguramente cuando $x_0 = x$, es decir, $J(\pi, c, x) = \rho_{n-1}$.

Para cada $x \in X$, definimos

$$\mathbb{P}_\infty(x) := \bigcap_{n=0}^{\infty} \mathbb{P}_n(x),$$

Observación 3.3.4 *(a) Las funciones* Φ_n, $n \in \mathbb{N}_0$, *definidas en (3.3.6) satisfacen las hipótesis 3.2.1(b) y* (d_1), *y por tanto, pueden ser usadas como funciones de costo por etapa.*
(b) Dados $x \in X$ *y* $\pi \in \mathbb{P}_{n+1}(x)$, *para cada* $n \in \mathbb{N}_0$ *se cumple que:*

$$1 = P_x^\pi[a_t \in A_{n+1}(x_t) \ \forall\ t \in \mathbb{N}_0] = P_x^\pi[\Phi_n(x_t, a_t) = 0 \ \forall\ t \in \mathbb{N}_0].$$

Usando el hecho de que $\Phi_n \geq 0$, *la igualdad anterior es equivalente a* $E_x^\pi[\Phi_n(x_t, a_t)] = 0$ *para todo* $t \in \mathbb{N}_0$, *y de acuerdo a nuestra definición del costo* $\alpha-$*descontado esperado en 2.4.1, tenemos que se cumple la siguiente afirmación:*

$$\text{Si } \pi \in \mathbb{P}_{n+1}(x), \text{ entonces } V_\alpha(\pi, \Phi_n, x) = 0 \quad \text{para cualquier} \quad \alpha \in (0,1). \tag{3.3.9}$$

Consecuentemente, si $\pi^* \in \mathbb{P}_\infty(x)$, *entonces* $V_\alpha(\pi^*, \Phi_n, x) = 0$ *para toda* $n \in \mathbb{N}_0$ *y toda* $\alpha \in (0,1)$.

El siguiente teorema es el resultado principal de este trabajo.

Teorema 3.3.5 *Bajo las Hipótesis 3.2.1, 3.2.3 y 3.2.4, una política $\pi \in \mathbb{P}$ es BO en x si y sólo si $\pi \in \mathbb{P}_\infty(x)$*

En otras palabras, una política π es BO en x si y sólo si satisface la EOCP P_x^π–c.s. para cada modelo $\mathfrak{M}_n$. El resto de este capítulo está dedicado a la demostración de este teorema.

3.4 Resultados preliminares

Recordemos que para cada factor de descuento $\alpha \in (0,1)$, la tasa de interés correspondiente $\mathbf{r}(\alpha)$ está dada por

$$\mathbf{r}(\alpha) := \frac{1-\alpha}{\alpha} \tag{3.4.1}$$

La demostración del Teorema 3.3.5 requiere probar que si $\pi^* \in \mathbb{P}_\infty(x)$ y $\pi \in \mathbb{P}_n(x)$ para alguna $n < \infty$,

$$\lim_{\alpha \uparrow 1} \frac{V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x)}{\mathbf{r}(\alpha)^n} > 0.$$

De hecho, como $\mathbf{r}(\alpha)$ es siempre positiva, la desigualdad anterior implica que $V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x) > 0$ para toda α cercana a 1, es decir, implica que π^* es Blackwell óptima. Por esta razón, los primeros 3 lemas que veremos a continuación tienen como objetivo dar una expresión para el cociente

$$\frac{V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x)}{\mathbf{r}(\alpha)^n}$$

en términos de $V_\alpha(\pi, \Phi_n, x)$ con la intención de usar la afirmación (3.3.9) en el análisis de la convergencia de dicho cociente cuando $\alpha \uparrow 1$.

Observación 3.4.1 *La Hipótesis 3.2.3(a) implica que $l(x,a)\,\nu(X) \leq 1$, por lo que la parte (c) de la misma Hipótesis nos conduce a*

$$\int_X \omega(y)\, Q(dy \,|\, x, a) \leq \mathcal{K}\omega(x) + b \tag{3.4.2}$$

donde $b = \nu(\omega)/\nu(X)$.

Para empezar, introducimos aquí una función auxiliar $\widetilde{h} : \mathbb{K} \to \mathbb{R}$.

Lema 3.4.2 *Para cada $h \in \mathbb{B}_w(X)$, la función*

$$\widetilde{h}(x,a) := \int_X h(y) Q(dy|x,a) \tag{3.4.3}$$

satisface la Hipótesis 3.2.1(b) y (d_1) y, por tanto, $\widetilde{h}$ puede ser usada como una función costo por etapa.

Demostración. Tomemos $h \in \mathbb{B}_w(X)$. Entonces

$$\begin{aligned} |\widetilde{h}(x,a)| &\leq \int_X |h(y)| Q(dy|x,a) \\ &\leq \|h\|_\omega \int_X \omega(y) Q(dy|x,a) \\ &\leq \|h\|_\omega (\mathcal{K}\omega(x) + b) \quad \text{por (3.4.2).} \end{aligned}$$

De aquí que

$$\sup_{a \in A(x)} |\widetilde{h}(x,a)| \leq (\mathcal{K} + b)\|h\|_\omega \omega(x),$$

y se cumple la parte (d_1) de la Hipótesis 3.2.1. Por otro lado, por la observación 3.3.1 tenemos que $\widetilde{h}(x,a)$ es continua en $A(x)$ para cada $x \in X$. ■

Lema 3.4.3 *Sea $x \in X$ un estado inicial fijo. Para cada $\alpha \in (0,1)$, $\pi \in \mathbb{P}$ y $h \in \mathbb{B}_w(X)$,*

$$V_\alpha(\pi, \widetilde{h}, x) = \frac{1}{\alpha}[V_\alpha(\pi, h, x) - h(x)].$$

Demostración. De la definición de $\widetilde{h}$ en (3.4.3) se tiene que

$$\widetilde{h}(x,a) = E_x^\pi[h(x_{t+1})|x_t = x,\ a_t = a],$$

y por lo tanto

$$E_x^\pi[h(x_{t+1})] = E_x^\pi[\widetilde{h}(x_t, a_t)], \qquad \forall\, t \in \mathbb{N}_0.$$

Usando $\widetilde{h}$ como función de costo por etapa en el costo $\alpha-$descontado esperado, obtenemos

$$\begin{aligned} V_\alpha(\pi, \widetilde{h}, x) &= \sum_{t=0}^{\infty} \alpha^t E_x^\pi[\widetilde{h}(x_t, a_t)] = \sum_{t=0}^{\infty} \alpha^t E_x^\pi[h(x_{t+1})] \\ &= \frac{1}{\alpha} \sum_{t=1}^{\infty} \alpha^t E_x^\pi[h(x_t)] = \frac{1}{\alpha}\left[\sum_{t=0}^{\infty} \alpha^t E_x^\pi[h(x_t)] - h(x)\right], \end{aligned}$$

donde usamos el hecho de que $E_x^\pi[h(x_0)] = h(x)$ en la última igualdad. ■

Observación 3.4.4 *Nótese que*

$$E_x^\pi\left[\omega(x_t) \mid h_{t-1}, a_{t-1}\right] = \int_X \omega(y)\, Q(dy \mid x_{t-1}, a_{t-1}) \leq \mathcal{K}\omega(x_{t-1}) + b$$

por 3.4.2. Aplicando E_x^π obtenemos $E_x^\pi \omega(x_t) \leq \mathcal{K} E_x^\pi \omega(x_{t-1}) + b$. Iterando esta última desigualdad se llega a

$$E_x^\pi \omega(x_t) \leq \mathcal{K}^t \omega(x) + b \sum_{j=0}^{t-1} \mathcal{K}^j \leq \left(1 + \frac{b}{1-\mathcal{K}}\right) \omega(x).$$

En la primera desigualdad usamos que $E_x^\pi \omega(x_0) = \omega(x)$ *y la segunda se obtiene trivialmente ya que* $0 < K^t < 1$ *y* $\omega(x) \geq 1$. *Por otro lado, para cualquier* $u \in \mathbb{B}_\omega(X)$ *se tiene que* $E_x^\pi |u(x_t)| \leq \|u\|_\omega E_x^\pi \omega(x_t)$, *de donde se concluye*

$$E_x^\pi |u(x_t)| \leq \|u\|_\omega \left(1 + \frac{b}{1-K}\right) \omega(x). \tag{3.4.4}$$

(Ver Lema 10.4.1 en [14]). Esta última desigualdad nos será útil en el siguiente lema.

Lema 3.4.5 *Sean* $x \in X$ *y* $\pi^* \in \mathbb{P}_\infty(x)$ *arbitrarias pero fijas. Entonces, para cualquier* $\pi \in \mathbb{P}$ *se cumple:*

$$V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x) = V_\alpha(\pi, \Phi_0, x) + \mathbf{r}(\alpha)\left[V_\alpha(\pi, h_1, x) - V_\alpha(\pi^*, h_1, x)\right]. \tag{3.4.5}$$

Análogamente, para $\pi \in \mathbb{P}_n(x)$, $n \in \mathbb{N}$, *tenemos*

$$\begin{aligned} V_\alpha(\pi, h_n, x) - V_\alpha(\pi^*, h_n, x) &= V_\alpha(\pi, \Phi_n, x) + \\ &+ \mathbf{r}(\alpha)\left[V_\alpha(\pi, h_{n+1}, x) - V_\alpha(\pi^*, h_{n+1}, x)\right]. \end{aligned} \tag{3.4.6}$$

Así que, para $n \in \mathbb{N}_0$ *y* $\pi \in \mathbb{P}_n(x)$

$$\begin{aligned} &\frac{V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x)}{\mathbf{r}(\alpha)^n} = \\ &= V_\alpha(\pi, \Phi_n, x) + \mathbf{r}(\alpha)\left[V_\alpha(\pi, h_{n+1}, x) - V_\alpha(\pi^*, h_{n+1}, x)\right]. \end{aligned} \tag{3.4.7}$$

Demostración. Consideremos el modelo $\mathfrak{M}_0$ y recordemos la definición de Φ_0:

$$\Phi_0(x, a) := c(x, a) - \int_X h_1(y) Q(dy \,|\, x, a) - \rho_0 + h_1(x), \quad \text{con} \quad (x, a) \in \mathbb{K}_0.$$

Reescribiendo esta igualdad obtenemos:

$$\begin{aligned} -h_1(x) = c(x, a) - \Phi_0(x, a) - \rho_0 &- (1-\alpha) \int_X h_1(y) Q(dy \,|\, x, a) \\ &- \alpha \int_X h_1(y) Q(dy \,|\, x, a), \quad \text{con} \quad (x, a) \in \mathbb{K}_0, \end{aligned}$$

o, equivalentemente,

$$-h_1(x) = K(x, a) - \alpha \int_X h_1(y) Q(dy \,|\, x, a), \qquad (x, a) \in \mathbb{K}_0, \tag{3.4.8}$$

donde $K(x, a) = c(x, a) - \Phi_0(x, a) - \rho_0 - (1-\alpha)\widetilde{h}_1(x, a)$ y $\widetilde{h}_1$ se obtiene reemplazando $h(y)$ por $h_1(y)$ en (3.4.3). Observemos que $K(x, a)$ satisface las partes (b) y (d) de la Hipótesis 3.2.1. Iterando la expresión (3.4.8) llegamos a

$$-h_1(x) = \sum_{t=0}^{N-1} \alpha^t E_x^\pi K(x_t, a_t) - \alpha^N E_x^\pi h_1(x_N).$$

Por (3.4.4)

$$\alpha^N E_x^\pi h_1(x_N) \to 0 \quad \text{cuando} \quad N \to \infty.$$

Entonces

$$-h_1(x) = V_\alpha(\pi, K, x).$$

Usando la linealidad del mapeo $K \mapsto V_\alpha(\pi, K, x)$ obtenemos

$$-h_1(x) = V_\alpha(\pi, c, x) - V_\alpha(\pi, \Phi_0, x) - \rho_0/(1-\alpha) - (1-\alpha)V_\alpha(\pi, \widetilde{h}_1, x).$$

Combinando el Lema 3.4.3 con la definición de $\mathbf{r}(\alpha)$ en (3.4.1), obtenemos

$$-\frac{h_1(x)}{\alpha} = V_\alpha(\pi, c, x) - V_\alpha(\pi, \Phi_0, x) - \frac{\rho_0}{1-\alpha} - \mathbf{r}(\alpha)\, V_\alpha(\pi, h_1, x). \tag{3.4.9}$$

Reemplazando π con π^* en esta ecuación y recordando que $V_\alpha(\pi^*, \Phi_0, x) = 0$ (por la observación 3.3.4), se sigue que

$$-\frac{h_1(x)}{\alpha} = V_\alpha(\pi^*, c, x) - \frac{\rho_0}{1-\alpha} - \mathbf{r}(\alpha)\, V_\alpha(\pi^*, h_1, x).$$

Combinando esta igualdad con (3.4.9) se deduce (3.4.5).

La ecuación (3.4.6) se obtiene reemplazando $\mathfrak{M}_0$ con $\mathfrak{M}_n$ en el desarrollo anterior.

Finalmente, la expresión (3.4.7) se sigue fácilmente por inducción. De hecho, para $n = 0$ la afirmación se reduce a (3.4.5). Supóngase ahora que el resultado es verdadero para alguna $n \in \mathbb{N}$ y que $\pi \in \mathbb{P}_{n+1}(x) \subset \mathbb{P}_n(x)$. En este caso, por (3.3.9), $V_\alpha(\pi, \Phi_n, x) = 0$ para toda $\alpha \in (0,1)$. De manera que la hipótesis de inducción (3.4.7) conduce a

$$V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x) = \mathbf{r}(\alpha)^{n+1} [V_\alpha(\pi, h_{n+1}, x) - V_\alpha(\pi^*, h_{n+1}, x)].$$

Usando (3.4.6) con $n+1$ en vez de n se sigue que

$$\begin{aligned} V_\alpha(\pi, c, x) \;-\; & V_\alpha(\pi^*, c, x) = \\ = \; & \mathbf{r}(\alpha)^{n+1} [V_\alpha(\pi, \Phi_{n+1}, x) + \mathbf{r}(\alpha)\,(V_\alpha(\pi, h_{n+2}, x) - V_\alpha(\pi^*, h_{n+2}, x))], \end{aligned}$$

y se obtiene la conclusión deseada. ■

Observación 3.4.6 *En los siguientes dos lemas usaremos una pareja de igualdades dadas en la página 86 de [13]. Aquí transcribimos estos resultados para facilitar la referencia a ellos. Sea $\{c_t\}$ una sucesión de números y $\alpha \in (0,1)$. Definimos $s_N = c_0 + c_1 + ... + c_{N-1}$. Para cualquier número S, se cumple que*

$$\sum_{t=0}^{\infty} \alpha^t c_t = \frac{S}{1-\alpha} + (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1}(s_t - tS). \tag{3.4.10}$$

En particular, sea π una política con un costo promedio finito $S := J(\pi, c, x)$, y sea $c_t := E_x^\pi[c(x_t, a_t)]$, de manera que $s_N = V_{N,1}(\pi, c, x)$. Entonces (3.4.10) se transforma en

$$V_\alpha(\pi, c, x) = \frac{J(\pi, c, x)}{1-\alpha} + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1}\left[V_{t,1}(\pi, c, x) - tJ(\pi, c, x)\right]. \tag{3.4.11}$$

Por ejemplo, si $\pi \in \mathbb{P}_1(x)$, entonces

$$\begin{aligned} V_\alpha(\pi, c, x) &= \frac{\rho_0}{1-\alpha} + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1}[-h_1(x) + E_x^\pi h_1(x_t)] \\ &= \frac{\rho_0}{1-\alpha} - h_1(x) + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1} E_x^\pi h_1(x_t) \\ &= \frac{\rho_0}{1-\alpha} - \frac{h_1(x)}{\alpha} + \mathrm{r}(\alpha)\, V_\alpha(\pi, h_1, x) \end{aligned}$$

La primera parte del siguiente lema, establece que si una política π está en $\mathbb{P}_{n+1}(x)$, entonces es promedio óptima para el modelo $\mathfrak{M}_n$ en un sentido más fuerte que en (2.4.2)–(2.4.3) ya que se tiene un límite y no un límite superior.

Lema 3.4.7 *Sean $x \in X$ y $n \in \mathbb{N}$ valores fijos. Para cada $\pi \in \mathbb{P}_{n+1}(x)$*

$$J(\pi, h_n, x) = \rho_n = \lim_{N\to\infty} \frac{1}{N} V_{N,1}(\pi, h_n, x), \tag{3.4.12}$$

y

$$(1-\alpha)V_\alpha(\pi, h_n, x) \to \rho_n \quad \textit{cuando} \quad \alpha \uparrow 1. \tag{3.4.13}$$

Demostración. Como π está en $\mathbb{P}_{n+1}(x)$, tenemos que $\Phi_n(x_t, a_t) = 0$ P_x^π–c.s. para toda t , i.e.,

$$\rho_n - h_{n+1}(x_t) = h_n(x_t, a_t) - E_x^\pi[h_{n+1}(x_{t+1})|x_t, a_t] \quad \forall\, t \in \mathbb{N}_0 \quad P_x^\pi - \text{c.s.}, \tag{3.4.14}$$

Aplicando el operador esperanza $E_x^\pi(\,\cdot\,)$ obtenemos

$$E_x^\pi h_{n+1}(x_{t+1}) - E_x^\pi h_{n+1}(x_t) = E_x^\pi h_n(x_t, a_t) - \rho_n \qquad \forall\, t \in \mathbb{N}_0. \tag{3.4.15}$$

Ahora, sumando sobre $t = 0, 1 \ldots N-1$ obtenemos

$$E_x^\pi h_{n+1}(x_N) - h_{n+1}(x) = V_{N,1}(\pi, h_n, x) - N\rho_n \qquad \forall\ N \in \mathbb{N}_0. \tag{3.4.16}$$

Por (3.4.4) $|E_x^\pi h_{n+1}(x_N)| \le \|h_{n+1}\|_w C w(x)$ donde $C = 1 + b/(1-K)$. Luego, multiplicando (3.4.16) por $1/N$ y haciendo $N \to \infty$ obtenemos (3.4.12).

Para probar (3.4.13), observemos que (3.4.11) y (3.4.16) conducen a

$$V_\alpha(\pi, h_n, x) = \frac{\rho_n}{1-\alpha} + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1}\left[V_{t,1}(\pi, h_n, x) - t\rho_n\right] \tag{3.4.17}$$

$$= \frac{\rho_n}{1-\alpha} + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1}\left[E_x^\pi h_{n+1}(x_t) - h_{n+1}(x)\right] \tag{3.4.18}$$

Por (3.3.8) y (3.4.4)

$$|E_x^\pi h_{n+1}(x_t) - h_{n+1}(x)| \leq k w(x), \tag{3.4.19}$$

donde $k = D^{n+1}\bar{c}(1+C)$. Usando este hecho y la ecuación (3.4.18), tenemos

$$\begin{aligned} |(1-\alpha)V_\alpha(\pi, h_n, x) - \rho_n| &\leq (1-\alpha)^2 \sum_{t=1}^{\infty} \alpha^{t-1} k w(x) \\ &= (1-\alpha) k w(x) \to 0 \quad \text{cuando} \quad \alpha \uparrow 1, \end{aligned}$$

y se obtiene la conclusión buscada. ■

Lema 3.4.8 *Sean $x \in X$, $n \in \mathbb{N}_0$ y $\pi \in \mathbb{P}_n(x)$ valores fijos. Entonces*

$$\liminf_{\alpha \uparrow 1} (1-\alpha) V_\alpha(\pi, h_{n+1}, x) \geq \rho_{n+1}$$

Demostración. Fijemos $\pi \in \mathbb{P}_n(x)$ y consideremos el modelo

$$\mathfrak{M}_{n+1} = (X, A, \{A_{n+1}(x) : x \in X\}, Q, h_{n+1})$$

Por la Proposición 3.2.6 existen $\rho_{n+1} \in \mathbb{R}$ y $h_{n+2} \in \mathbb{B}_w(X)$ que satisfacen

$$\rho_{n+1} - h_{n+2}(x_t) \leq h_{n+1}(x_t) - E_x^\pi[h_{n+2}(x_{t+1})|x_t, a_t)] \quad \forall\, t \in \mathbb{N}_0.$$

Entonces, siguiendo los mismos pasos que nos condujeron de (3.4.14) a (3.4.16) llegamos a

$$V_{N,1}(\pi, h_{n+1}, x) - N\rho_{n+1} \geq E_x^\pi[h_{n+2}(x_N)] - h_{n+2}(x) \quad \forall\, N \in \mathbb{N}_0. \tag{3.4.20}$$

Por otro lado, de la expresión (3.4.10) [o (3.4.11) reemplazando c por h_{n+1}], obtenemos

$$V_\alpha(\pi, h_{n+1}, x) = \frac{\rho_{n+1}}{1-\alpha} + (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1}\left[V_{t,1}(\pi, h_{n+1}, x) - t\rho_{n+1}\right].$$

Por lo tanto, por (3.4.20),

$$\begin{aligned} (1-\alpha)V_\alpha(\pi, h_{n+1}, x) &\geq \rho_{n+1} + (1-\alpha)^2 \sum_{t=1}^{\infty} \alpha^{t-1}\left[E_x^\pi[h_{n+2}(x_t)] - h_{n+2}(x)\right] \\ &= \rho_{n+1} - (1-\alpha)h_{n+2}(x) + (1-\alpha)^2 \sum_{t=1}^{\infty} \alpha^{t-1} E_x^\pi[h_{n+2}(x_t)]. \end{aligned}$$

Como $|E_x^\pi [h_{n+2}(x_N)]| \leq \|h_{n+2}\|_w Cw(x)$ por (3.4.4), se sigue la conclusión. ■

El último resultado preliminar que necesitamos para nuestra demostración del teorema principal de este trabajo es:

Lema 3.4.9 *Sean $x \in X$, $n \in \mathbb{N}_0$ y $\pi \in \mathbb{P}_n(x)$ arbitrarios pero fijos. Si $\pi \notin \mathbb{P}_{n+1}(x)$, entonces*

$$E_x^\pi \left[\sum_{t=0}^{\infty} \Phi_n(x_t, a_t) \right] > 0.$$

Demostración. Sea $\pi \in \mathbb{P}_n(x)$. Por (2.4.1), $V_\alpha(\pi, \Phi_n, x) = E_x^\pi[\sum_{t=0}^{\infty} \alpha^t \Phi_n(x_t, a_t)]$, y como $\Phi_n(x_t, a_t) \geq 0$ el teorema de convergencia monótona conduce a

$$\lim_{\alpha \uparrow 1} V_\alpha(\pi, \Phi_n, x) = E_x^\pi \left[\sum_{t=0}^{\infty} \Phi_n(x_t, a_t) \right].$$

Mas aun, como Φ_n es no negativa

$$E_x^\pi \left[\sum_{t=0}^{\infty} \Phi_n(x_t, a_t) \right] = 0$$

implica que $\Phi_n(x_t, a_t) = 0 \quad P_x^\pi$−c.s. para toda $t \in \mathbb{N}_0$, es decir, como en la Observación 3.3.4(b),

$$P_x^\pi[a_t \in A_{n+1}(x_t) \ \forall \ t \in \mathbb{N}_0] = 1,$$

así que $\pi \in \mathbb{P}_{n+1}(x)$. Por lo tanto $\pi \notin \mathbb{P}_{n+1}(x)$ implica que $E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(x_t, a_t)] > 0$. ■

3.5 Demostración del Teorema 3.3.5

Observación 3.5.1 *Usando la desigualdad (3.4.4) para $c(x_t, a_t)$ obtenemos*

$$E_x^\pi |c(x_t, a_t)| \leq \overline{c} C \omega(x) \qquad \forall \ \pi \in \mathbb{P}, \ x \in X, \ t \in \mathbb{N}_0$$

donde $C = 1 + b/(1 - K)$ y $\overline{c}$ es la constante dada en la Hipótesis 3.2.1(d_1). Se sigue entonces que

$$|V_\alpha(\pi, c, x)| \leq \frac{\overline{c} C w(x)}{1 - \alpha} \qquad \forall \ \alpha \in (0, 1). \tag{3.5.1}$$

Demostración. del Teorema 3.3.5 Sea $x \in X$ fija y elijamos una política arbitraria $\pi^* \in \mathbb{P}_\infty(x)$. Demostraremos que π^* es Blackwell óptima en x. Seleccionemos $\pi \in \mathbb{P}$ y consideremos los siguientes dos casos.

Caso 1: $\pi \notin \mathbb{P}_\infty(x)$.

En este caso existe $n \in \mathbb{N}_0$ tal que $\pi \in \mathbb{P}_n(x)$ pero $\pi \notin \mathbb{P}_{n+1}(x)$; recuérdese que $\pi \in \mathbb{P}_0(x) = \mathbb{P}$. Por (3.4.7), para toda $\alpha \in (0,1)$ tenemos

$$\begin{aligned} \frac{V_\alpha(\pi,c,x) - V_\alpha(\pi^*,c,x)}{\mathbf{r}(\alpha)^n} &= V_\alpha(\pi,\Phi_n,x) \\ &+ \mathbf{r}(\alpha)\left[V_\alpha(\pi,h_{n+1},x) - V_\alpha(\pi^*,h_{n+1},x)\right], \quad (3.5.2) \end{aligned}$$

donde $\mathbf{r}(\alpha) = (1-\alpha)/\alpha$. Ahora mostraremos que

$$\liminf_{\alpha\uparrow 1} \frac{V_\alpha(\pi,c,x) - V_\alpha(\pi^*,c,x)}{\mathbf{r}(\alpha)^n} > 0. \quad (3.5.3)$$

Para verificar esta desigualdad nótese que la política π satisface una de las siguientes condiciones *(a)* o *(b)*.

(a) $\lim_{\alpha\uparrow 1} V_\alpha(\pi,\Phi_n,x) = E_x^\pi\left[\sum_{t=0}^\infty \Phi_n(x_t,a_t)\right] = \infty$.
En esta situación, como tenemos que $|(1-\alpha)V_\alpha(\,\cdot\,,h_{n+1},x)| \le \bar{c}Cw(x)$, por (3.5.1), la ecuación (3.5.2) conduce a

$$\liminf_{\alpha\uparrow 1} \frac{V_\alpha(\pi,c,x) - V_\alpha(\pi^*,c,x)}{\mathbf{r}(\alpha)^n} = E_x^\pi\left[\sum_{t=0}^\infty \Phi_n(x_t,a_t)\right] = \infty,$$

y se cumple (3.5.3).

(b) $\lim_{\alpha\uparrow 1} V_\alpha(\pi,\Phi_n,x) = E_x^\pi\left[\sum_{t=0}^\infty \Phi_n(x_t,a_t)\right] < \infty$.
En este caso el Lema 3.4.8 garantiza que

$$\liminf_{\alpha\uparrow 1}(1-\alpha)V_\alpha(\pi,h_{n+1},x) \ge \rho_{n+1}.$$

Por otro lado, como $\pi^* \in \mathbb{P}_\infty(x) \subset \mathbb{P}_{n+2}(x)$, la expresión (3.4.13) conduce a

$$\lim_{\alpha\uparrow 1}(1-\alpha)V_\alpha(\pi^*,h_{n+1},x) = \rho_{n+1}.$$

Combinando los dos últimos hechos obtenemos

$$\liminf_{\alpha\uparrow 1}(1-\alpha)[V_\alpha(\pi,h_{n+1},x) - V_\alpha(\pi^*,h_{n+1},x)] \ge \rho_{n+1} - \rho_{n+1} = 0,$$

expresión que junto con (3.5.2) y el Lema 3.4.9 implica que

$$\liminf_{\alpha\uparrow 1} \frac{V_\alpha(\pi,c,x) - V_\alpha(\pi^*,c,x)}{\mathbf{r}(\alpha)^n} = E_x^\pi\left[\sum_{t=0}^\infty \Phi_n(x_t,a_t)\right] > 0.$$

De manera que se cumple (3.5.3).

Finalmente, como $\mathbf{r}(\alpha) > 0$ para toda $\alpha \in (0,1)$, (3.5.3) implica que existe $\alpha^* = \alpha(\pi^*,\pi,x)$ en $(0,1)$ tal que $V_\alpha(\pi,c,x) - V_\alpha(\pi^*,c,x) > 0$ para toda $\alpha \in (\alpha^*,1)$. Así, se satisface (3.1.1) y, por tanto, π^* es BO en el Caso 1.

Caso 2. $\pi \in \mathbb{P}_\infty(x)$.

En este caso (3.5.2) se sigue cumpliendo y además $V_\alpha(\pi, \Phi_n, x) = 0$ para toda $\alpha \in (0,1)$ y $n \in \mathbb{N}$ (ver la Observación 3.5.1).Por lo tanto, por (3.5.1) y (3.5.2),

$$|V_\alpha(\pi, c, x) - V_\alpha(\pi^*, c, x)| \leq \mathrm{r}(\alpha)^{n+1} \frac{2\bar{c}Cw(x)}{\alpha}, \qquad \forall\, \alpha \in (0,1),\ n \in \mathbb{N} \tag{3.5.4}$$

Más aun, nótese que si $\mathrm{r}(\alpha) < 1$, entonces la parte derecha de la expresión (3.5.4) converge a cero cuando $n \to \infty$. Se sigue que para $\alpha \in (1/2, 1)$

$$V_\alpha(\pi, c, x) = V_\alpha(\pi^*, c, x),$$

así que se cumple la definición de optimalidad de Blackwell dada en (3.1.1) con $\alpha^* = 1/2$. De hecho, como dos series de potencias que coinciden en un intervalo necesariamente coinciden en todo su dominio, se sigue que $V_\alpha(\pi, c, x) = V_\alpha(\pi^*, c, x)$ para toda $\alpha \in (0,1)$, así que en este caso α^* puede ser tomada como cero.

Para terminar obsérvese que dada una política π necesariamente satisface uno de los dos casos previamente analizados, así que la discusión anterior puede resumirse en la afirmación: *una política $\pi^* \in \mathbb{P}_\infty(x)$ es Blackwell óptima en x.*

Inversamente, si $\pi \notin \mathbb{P}_\infty(x)$ el análisis del Caso 1 anterior muestra que para cualquier $\pi^* \in \mathbb{P}_\infty(x)$ la desigualdad $V_\alpha(\pi^*, c, x) - V_\alpha(\pi, c, x) < 0$ se cumple para toda α suficientemente cercana a 1, así que π no es Blackwell óptima en x. Esto concluye la demostración del Teorema 3.3.5. ■

3.6 Comparación con el trabajo de Hordijk y Yushkevich

A. Hordijk y A.A. Yushkevich (H-Y) publicaron en 1999 los artículos [18] y [19] en los que presentan condiciones bajo las cuales es posible asegurar la existencia de políticas BO en modelos cuyos espacios de estados y de acciones son espacios de Borel y con funciones de costo por etapa no acotadas. En virtud de que aquí hemos presentado otra vía para llegar al mismo resultado, en esta sección haremos una comparación de las hipótesis y de los procedimientos desarrollados en ambos trabajos.

Para facilitar la comparación, reescribimos a continuación las hipótesis de H-Y usando nuestra notación:

Hipótesis 3.6.1 *Para cada x en X :*

1. $A(x)$ *es compacto;*
2. (a) $\sup_{a \in A(x)} |r(x,a)| \leq \omega(x)$ y
   (b) $\int_X \omega(y)\, Q(dy \mid x, a) \leq C\omega(x)$ *para todo a en $A(x)$,*
   para alguna función de peso $\omega(\cdot) \geq 1$ y alguna constante C;

3. (*a*) *La función recompensa por etapa* $r(x,a)$ *es continua en* a *en* $A(x)$, *y*

   (*b*) *La función* $a \to \int_X v(y)\, Q(dy \mid x,a)$ *es continua en* $A(x)$ *para cada función* $v \in \mathbb{B}_\omega(X)$;

4. *Para cada política estacionaria y aleatorizada* σ, *existen* $C > 0$ *y* $\gamma \in (0,1)$ *tales que*

$$\left\| Q_\sigma^t - \overline{Q}_\sigma \right\|_\omega \leq C\gamma^t \qquad t = 0, 1, \ldots,$$

   en caso de que las cadenas de Markov generadas sean aperiódicas; si no es así, entonces se requiere la condición

$$\left\| \frac{1}{T} \sum_{k=1}^{T} Q_\sigma^{k+t} - \overline{Q}_\sigma \right\|_\omega \leq C\gamma^t \quad \textit{para alguna constante } T < \infty.$$

5. *En* X *hay una medida de referencia* m *que es* σ*−finita, y*

   (*a*) $Q(B \mid x,a) = \int_B p(y \mid x,a)\, m(dy)$ *con* $(x,a) \in \mathbb{K}$ *y* $B \in \mathcal{B}(X)$, *donde* p *es una densidad de transición,*

   (*b*) $p(y \mid x,a)$ *es continua en* a,

   (*c*) $\int_X \widehat{p}(y \mid x)\, \omega(y)\, m(dy) \leq C\omega(x)$, *donde* $\widehat{p}(y \mid x) = \max_{a \in A(x)} p(y \mid x,a)$.

6. *Para cada estado inicial* x_0 *y cada política Markoviana* $\pi = (\sigma_1, \sigma_2, \ldots)$, *la densidad de transición en* t *pasos satisface que, para cada* $\varepsilon > 0$, *existen un conjunto* $X' \subset X$ *con* $m(X') < \infty$ *y una constante positiva* L *tales que*

$$\begin{aligned} \int_{X \setminus X'} \omega(x)\, q_\pi^{(t)}(x \mid x_0)\, m(dx) &< \varepsilon \quad y \\ \omega(x)\, q_\pi^{(t)}(x \mid x_0) &\leq L \quad \textit{para } x \in X' \end{aligned}$$

   donde

$$\begin{aligned} q_\pi^{(1)}(x \mid x_0) &= p_{\sigma_1}(x \mid x_0) \quad y \\ q_\pi^{(t+1)}(x \mid x_0) &= \int_X q_\pi^{(t)}(z \mid x_0)\, p_{\sigma_{t+1}}(x \mid z)\, m(dz). \end{aligned}$$

Si se cumple la condición 5 de 3.6.1, se puede prescindir de las condiciones 2(*b*) y 3(*b*); los autores las incluyen para demostrar resultados preliminares que no requieren la existencia de densidades de transición.

Las condiciones 1, 2, 3 y 4 de la Hipótesis 3.6.1 están incluídas en nuestra Hipótesis 3.2.1 y en las conclusiones de la proposición 3.2.5.

En el terreno de las hipótesis, una primera diferencia es que nuestro enfoque no requiere la existencia de densidades de transición (condición 5 de H-Y). Además, en

lugar de pedir directamente la ergodicidad geométrica para las cadenas de Markov generadas por políticas estacionarias (condición 4 de H-Y), nosotros incluímos otras condiciones más fácilmente verificables contenidas en las Hipótesis 3.2.3 y 3.2.4, que garantizan dicha ergodicidad (ver la Proposición 3.2.5). Es conveniente hacer notar que en nuestro trabajo se requiere que esta condición sea satisfecha por las cadenas generadas por políticas estacionarias y deterministas, mientras que H-Y lo requieren en las generadas por el conjunto más amplio de políticas estacionarias y aleatorizadas, es decir, políticas $\pi = (\sigma, \sigma, \sigma, \ldots)$ donde $\sigma(\cdot \mid \cdot)$ es una distribución de probabilidad en A dado X.

Más importante que lo anterior, es el hecho de que con las condiciones 1 a 5 de la Hipótesis 3.6.1, H-Y demuestran la existencia de políticas BO únicamente en la clase de las políticas estacionarias, y requieren la condición 6 para extender este resultado al conjunto de todas las políticas. En nuestro enfoque, con condiciones equiparables a sus primeras cinco hipótesis, se demuestra la existencia de políticas BO en el conjunto de todas las políticas, sin requerir ninguna condición adicional. La hipótesis de la cual prescindimos en este trabajo, requiere que para cualquier política estacionaria π y cualquier estado inicial x_0 la colección $\left\{\omega(x)\, q_\pi^{(t)}(x \mid x_0)\right\}$ cumpla una condición tipo "tensión" (*tightness*) para todo x en algún conjunto de medida finita, además de estar uniformemente acotada.

Las diferencias entre los conjuntos de hipótesis, están muy relacionadas con los distintos procedimientos que se utilizan en cada uno de los enfoques para demostrar el resultado principal. El trabajo de H-Y generaliza a espacios de Borel el procedimiento usado en los artículos [5], [6], [28] y [30]; es decir, definir un orden (parcial) lexicográfico en el espacio lineal de todas las series de Laurent de la forma

$$h = \sum_{n=-1}^{\infty} h^{(n)}(x)\, \mathrm{r}^n \quad \text{con} \quad h^{(n)} \in \mathbb{B}_\omega(X),\ x \in X$$

y construir los operadores necesarios para aplicar la técnica de iteración de políticas en ese espacio. H-Y introducen una topología *débil-fuerte* en la clase de las políticas estacionarias y aleatorizadas, basada en funciones de Caratheodory, en la cual dicho conjunto de políticas es compacto. Usando esta topología, demuestran la continuidad respecto a la política, de los coeficientes de las series de Laurent para la recompensa descontada esperada. Finalmente, aplican maximización lexicográfica en las recompensas descontadas, pero no de manera puntual (en cada estado inicial) sino para alguna distribución inicial absolutamente continua respecto a la medida de referencia m.

El enfoque que aquí utilizamos – encontrar políticas que satisfagan la ecuación de optimalidad para costo promedio en una sucesión de modelos encajados– resulta

mucho más directo y nos permite usar únicamente las hipótesis necesarias para la existencia de soluciones a dicha ecuación de optimalidad.

Capítulo 4

Ejemplos

4.1 Introducción

En los artículos [10] y [19] se presentan ejemplos que satisfacen las hipótesis que requerimos en este trabajo; de hecho, estos ejemplos satisfacen hipótesis más restrictivas que las que aquí utilizamos. El primero de ellos consiste en un sistema controlado cuya dinámica se describe por

$$x_{t+1} = (x_t + a_t\eta_t - \xi_t)^+, \qquad t \in \mathbb{N}_0,$$

donde $\{\eta_t\}$ y $\{\xi_t\}$ son perturbaciones estocásticas, es decir, sucesiones independientes de variables aleatorias independientes e identicamente distribuidas. Este tipo de sistemas tienen aplicaciones en modelos de inventarios y en modelos de colas con un solo servidor. Las condiciones impuestas al modelo son:

1. $X = [0, \infty)$ y todos los conjuntos de acciones son iguales: $A(x) = [0, \kappa]$ para algún $\kappa > 0$ finito.

2. η_0 y ξ_0 tienen densidades acotadas y continuas en $[0, \infty)$ y la variable aleatoria $\zeta = \kappa\eta_0 - \xi_0$ satisface

   $$(i) \quad E(\zeta) < 0 \qquad \text{y} \qquad (ii) \quad E\left(e^{q\zeta}\right) < \infty$$

   para algún número $q > 0$.

3. La función costo por etapa cumple que: $\sup_A c(x,a) \leq \overline{c}\, e^{qx}$ para alguna $\overline{c} > 0$, y para cada estado x existen funciones ψ^c, ψ^Q en $\mathbb{B}_\omega(X)$ que satisfacen que $\psi^c(y) \to 0$ y $\psi^Q(y) \to 0$ cuando $y \downarrow 0$, tales que, para toda $x' \in X$,

   $$\begin{aligned} \sup_A \left|c(x,a) - c(x',a)\right| &\leq \psi^c\left(d(x,x')\right) \\ \sup_A \left\|Q(\cdot\,|\,x,a) - Q(\cdot\,|\,x',a)\right\|_\omega &\leq \psi^Q\left(d(x,x')\right) \end{aligned}$$

donde $d(\cdot,\cdot)$ es una métrica en X.

El ejemplo incluido en el artículo [19] se refiere a un sistema cuya evolución está descrita por ecuaciones lineales con una perturbación estocástica de la forma:

$$x_t = x_{t-1} + a_t + \xi_t \qquad t \in \mathbb{N}_0,$$

en el que los costos son funciones cuadráticas en x y en a. A un sistema de este estilo se le conoce como lineal-cuadrático. Entre las condiciones del modelo se incluye la restricción de que todos los conjuntos de acciones viables $A(x)$ estén contenidos en un compacto $[-M, M]$. Además, se usa una densidad Gaussiana como distribución común de las perturbaciones.

En este trabajo vamos a analizar un sistema lineal–cuadrático con restricciones distintas a las que mencionamos anteriormente: los conjuntos de acciones serán subconjuntos compactos de $\mathbb{R}$ y la distribución de las perturbaciones será una densidad con soporte en un compacto. Las demás condiciones que le impondremos al modelo *son equivalentes a las usadas por Hordijk y Yushkevich en [19].*

4.2 Un sistema lineal-cuadrático.

4.2.1 Elementos del modelo

Consideremos el sistema cuya evolución está dada por:

$$x_{t+1} = k_1 x_t + k_2 a_t + \xi_t, \qquad t \in \mathbb{N}_0, \tag{4.2.1}$$

donde $x_t \in \mathbb{R}$ para toda t y los coeficientes k_1 y k_2 son positivos. Las perturbaciones estocásticas ξ_t son variables aleatorias independientes con distribución común $g(\cdot)$ que tienen media cero y varianza finita, i.e.

$$E(\xi_t) = 0 \qquad \text{y} \qquad \sigma^2 = E(\xi_t^2) < \infty.$$

El conjunto de los números reales es tanto el espacio de estados como el de acciones, i.e.

$$X = \mathbb{R} = A$$

y los conjuntos de acciones viables en cada estado son de la forma

$$A(x) = [-\psi_1(x), \psi_2(x)] \tag{4.2.2}$$

donde ψ_1 y ψ_2 son funciones continuas no-negativas.

La función de costo por etapa es la cuadrática

$$c(x,a) := c_1 x^2 + c_2 a^2 \quad \text{para toda } (x,a) \in \mathbb{K}, \tag{4.2.3}$$

con coeficientes no−negativos c_1 y c_2

La probabilidad de transición está determinada por la densidad común de las perturbaciones $g(\cdot)$:

$$\begin{aligned} Q(B \mid x, a) &= \Pr[x_{t+1} \in B \mid x_t = x, a_t = a] \\ &= \Pr[k_1 x + k_2 a + \xi_0 \in B] \\ &= \int_B g(y - k_1 x - k_2 a)\, dy, \end{aligned}$$

y con esto terminamos de describir todos los elementos de un modelo de control Markoviano

$$\mathfrak{M} := (X, A, \{A(x) \mid x \in X\}, Q, c).$$

4.2.2 Condiciones para que se cumplan las hipótesis

Las condiciones que requeriremos en el modelo que acabamos de describir para garantizar que las hipótesis del Teorema 3.3.5 se satisfagan, son las siguientes:

Hipótesis 4.2.1 *1.* $0 < k_1 < 1/2$ *donde* k_1 *es el coeficiente en (4.2.1).*

2. *Las funciones* ψ_1 *y* ψ_2 *en (4.2.2) satisfacen que* $\psi_i(x) \geq k_1 |x| / k_2$, $i = 1, 2$.

3. *La función de peso* $\omega(\cdot)$ *está dada por*

$$\omega(x) := \overline{\omega}\, e^{\gamma |x|}$$

donde $\overline{\omega} := \max\{1, c_1 + c_2 (k_1/k_2)^2\}$ *y* $\gamma \geq 2$. *(Nótese que* $\omega(\cdot) \geq 1$.)

4. *La densidad común* $g(\cdot)$ *de las perturbaciones* ξ_t, *es una función continua y acotada con soporte en el intervalo* $S := [-\widehat{s}, \widehat{s}]$ *donde* $\widehat{s}$ *satisface*

$$\gamma \widehat{s} < \log(\gamma/2 + 1). \tag{4.2.4}$$

5. *Existe* $\varepsilon > 0$ *tal que* $g(s) \geq \varepsilon$ *para toda* $s \in S$.

6. *Sea* λ *la medida de Lebesgue en* X *y* $S_0 := [0, \widehat{s}]$. *La función* $l(\cdot, \cdot)$ *y la medida* $\nu(\cdot)$ *requeridas en la Hipótesis 3.2.3, están dadas por*

$$l(x, a) \equiv I_{S_0}(x) \quad \forall (x, a) \in \mathbb{K}, \quad y \quad \nu(B) := \varepsilon \lambda(B \cap S_0) \quad \forall B \in \mathfrak{B}(X).$$

Observación 4.2.2 *En relación con la Hipótesis 4.2.1(2), tomaremos específicamente*

$$A(x) = \left[-\frac{k_1 |x|}{k_2}, \frac{k_1 |x|}{k_2} \right] \tag{4.2.5}$$

en (4.2.2) debido a que tal elección simplifica enormemente los cálculos.

4.2.3 Verificación de las hipótesis

Hipótesis 3.2.1

Por (4.2.5) y (4.2.3) se cumplen las partes (a) (compacidad de $A(x)$) y (b) (continuidad de $c(x,\cdot)$) de 3.2.1. De la Hipótesis 4.2.1(3) se desprende que la función costo por etapa satisface

$$\sup_{a\in A(x)} c(x,a) \leq \left[c_1 + c_2 (k_1/k_2)^2\right] x^2 \leq \varpi x^2 \leq \omega(x)$$

de manera que se cumple también la parte (d_1) de 3.2.1 con $\bar{c} = 1$. Para verificar la parte (c) de 3.2.1, sea $v(\cdot)$ una función medible y acotada arbitraria. Entonces, por 4.2.1(4)

$$\begin{aligned} \int_X v(y)\, Q(dy \mid x,a) &= E[v(x) \mid x_t = x, a_t = a] \\ &= E[v(k_1 x + k_2 a + \xi_0)] \\ &= \int_X v(y)\, q(y \mid x,a)\, dy \end{aligned} \tag{4.2.6}$$

siendo $q(y \mid x,a)$ la función de densidad

$$q(y \mid x,a) := I_{D(x,a)}(y)\, g(y - k_1 x - k_2 a),$$

donde $I_{D(x,a)}$ es la función indicadora del intervalo

$$D(x,a) := [k_1 x + k_2 a - \widehat{s}, k_1 x + k_2 a + \widehat{s}].$$

Nótese que y está en $D(x,a)$ si y sólo si $y - k_1 x - k_2 a$ está en S. La Hipótesis 3.2.1(c) se sigue de 4.2.6 y de la hipótesis 4.2.1(4).

La parte (d_2) no se requiere porque la función de costo es no-negativa (ver la Observación 3.3.1(b)).

Hipótesis 3.2.3 y 3.2.4

Para empezar, hay que notar que la forma de los conjuntos $A(x)$ dada en (4.2.5), lleva a

$$|k_1 x + k_2 a| \leq 2k_1 |x| < |x| \qquad \forall\, (x,a) \in \mathbb{K},$$

que a su vez implica que el intervalo $D(x,a)$ contiene a $S_0 = [0, \widehat{s}]$. Así que

$$I_{D(x,a)} \geq I_{S_0}. \tag{4.2.7}$$

Finalmente, en (4.2.6) reemplacemos $v(\cdot)$ por la función indicadora I_B de un conjunto de Borel arbitrario $B \subset X$ para obtener:

$$Q(B \mid x,a) = \int_B q(y \mid x,a)\, dy.$$

Entonces, las partes (5) y (6) de la Hipótesis 4.2.1 y la relación entre funciones indicadoras (4.2.7) conducen a

$$Q(B \mid x, a) \geq I_{S_0}(x)\, \varepsilon \int_X I_{B \cap S_0}(y)\, dy \qquad \text{para toda } B \in \mathfrak{B}(X).$$

lo que nos garantiza la Hipótesis 3.2.3(a). Es evidente que las definiciones de l y ν dadas en 4.2.1(6) satisfacen las partes (b) $\left(\int_{\mathbb{R}} \omega\, d\nu < \infty\right)$ y (d) $\left(\int_{\mathbb{R}} l_f\, d\nu > 0\right)$ de 3.2.3 y la Hipótesis 3.2.4.

Así que lo único que falta es verificar la desigualdad (3.2.1) dada en la Hipótesis 3.2.3(c). Para este fin, hay que observar que

$$\int_X \omega(y)\, Q(dy \mid x, a) = \int_{-\widehat{s}}^{\widehat{s}} \omega(k_1 x + k_2 a + s)\, g(s)\, ds$$

y la definición de ω dada en 4.2.1(3) conducen a

$$\begin{aligned} \int_X \omega(y)\, Q(dy \mid x, a) &\leq \overline{\omega} e^{\gamma |k_1 x + k_2 a|} \int_{-\widehat{s}}^{\widehat{s}} e^{\gamma |s|} ds \\ &\leq \omega(x)\, e^{-\gamma |x|} e^{\gamma |x|} 2 \left(e^{\gamma \widehat{s}} - 1\right) / \gamma \\ &= \mathcal{K}\, \omega(x) \qquad \forall x \in X, \end{aligned}$$

donde $\mathcal{K} := \frac{2}{\gamma}\left(e^{\gamma \widehat{s}} - 1\right)$ que es menor que 1 debido a la desigualdad (4.2.4).

Por lo tanto, se cumplen todas las hipótesis del Teorema 3.3.5 lo que garantiza la exitencia de políticas Blackwell óptimas para el sistema lineal–cuadrático descrito.

4.2.4 Política Blackwell óptima

Para empezar, debemos buscar políticas que satisfagan la ecuación de optimalidad para costo promedio en el modelo inicial

$$\mathfrak{M}_0 := (X, A, \{A_0(x) \mid x \in X\}, Q, h_0),$$

donde $A_0(x) := A(x) = [-k_1/k_2 x, k_1/k_2 x]$, y $h_0(x, a) := c(x, a) = c_1 x^2 + c_2 a^2$.

D. Blackwell fue también el iniciador de un método para obtener políticas promedio óptimas conocido como *descuento desvaneciente* (*vanishing discount approach*). Este método consiste en encontrar funciones de costo α–descontado $V_\alpha(\cdot)$ para valores variables de α, y hacer tender α a 1 para obtener políticas promedio óptimas partiendo de políticas α–descontadas óptimas. Para ver bajo qué condiciones es posible aplicar este procedimiento, se puede consultar por ejemplo [13].

Siguiendo esa vía en nuestro ejemplo, para una política estacionaria f_α, una función u y un estado fijo z definimos:

$$\begin{aligned} \rho_\alpha &:= (1 - \alpha)\, V_\alpha(f_\alpha, u, z) \\ h_\alpha(x) &:= V_\alpha(f_\alpha, u, z) - V_\alpha(f_\alpha, u, x). \end{aligned}$$

Si f_α es una política estacionaria α–descontada óptima y $u = c(\cdot, f_\alpha(\cdot))$, las igualdades anteriores se transforman en:

$$\rho_\alpha = (1-\alpha) V_\alpha^*(z) \tag{4.2.8}$$

$$h_\alpha(x) = V_\alpha^*(z) - V_\alpha^*(x), \tag{4.2.9}$$

donde el costo óptimo esperado V_α^* es la única solución en $\mathbb{B}_\omega(X)$ de la ecuación de programación dinámica para costo α–descontado, dada por

$$V_\alpha^*(x) = \min_{a \in A_0(x)} \left\{ c_1 x^2 + c_2 a^2 + \alpha \int_X V_\alpha^*(y)\, Q(dy \mid x, a) \right\} \qquad \forall x \in X, \tag{4.2.10}$$

(ver el Teorema 8.3.6 en [14]).

Por la forma cuadrática de la función de costo, $V_\alpha^*(x)$ necesariamente es una función también cuadrática de la forma

$$V_\alpha^*(x) = v_1(\alpha) x^2 + v_2(\alpha) \qquad \forall x \in X. \tag{4.2.11}$$

Entonces, en lugar de resolver explícitamente la ecuación de programación dinámica, reemplazamos V_α por la expresión anterior y y por $k_1 x + k_2 a + s$ en (4.2.10) para obtener:

$$v_1(\alpha) x^2 + v_2(\alpha) =$$

$$= \min_{a \in A_0(x)} \left\{ c_1 x^2 + c_2 a^2 + \alpha \int_X \left[v_1(\alpha) (k_1 x + k_2 a + s)^2 + v_2(\alpha) \right] g(s)\, ds \right\},$$

de donde, recordando que g es una densidad con media cero y varianza σ^2, obtenemos:

$$\min_{a \in A_0(x)} \left\{ c_1 x^2 + c_2 a^2 - v_1(\alpha) x^2 + \alpha v_1(\alpha) \left[k_1^2 x^2 + k_2^2 a^2 + 2 k_1 k_2 a x + \sigma^2 \right] + (\alpha - 1) v_2(\alpha) \right\}$$

$$= 0.$$

Para encontrar el mínimo, tomando

$$v_2(\alpha) = \frac{\alpha v_1(\alpha) \sigma^2}{1-\alpha}$$

la expresión anterior queda en la forma:

$$0 = \min_{a \in A_0(x)} \left\{ c_1 x^2 + c_2 a^2 + v_1(\alpha) \left[k_1^2 x^2 - \frac{x^2}{\alpha} + k_2^2 a^2 + 2 k_1 k_2 a x \right] \right\}. \tag{4.2.12}$$

Derivando la función entre corchetes respecto a la acción a e igualando a cero, obtenemos:

$$a^* = -\frac{\alpha v_1(\alpha) k_1 k_2}{c_2 + \alpha v_1(\alpha) k_2^2} x.$$

Sustituyendo este valor en la función que queremos minimizar, llegamos a la ecuación:

$$-\alpha k_2^2 x^2 v_1(\alpha)^2 + \left(\alpha c_1 k_2^2 x^2 + \alpha c_2 k_1^2 x^2 - c_2 x^2\right) v_1(\alpha) + c_1 c_2 x^2 = 0.$$

De esta manera, podemos finalmente concluir que $V_\alpha^*(x)$ dada en la forma (4.2.11), es la única solución de (4.2.10) si sus coeficientes cumplen que:

$$v_2(\alpha) := (1-\alpha)^{-1} \alpha v_1(\alpha) \sigma^2$$

y $v_1(\alpha)$ es la única solución positiva de la ecuación cuadrática

$$\alpha k_2^2 v_1(\alpha)^2 + \left(c_2 - \alpha c_1 k_2^2 - \alpha c_2 k_1^2\right) v_1(\alpha) - c_1 c_2 = 0. \tag{4.2.13}$$

Además, la política estacionaria α–descontada óptima está dada por:

$$f_\alpha^*(x) := -f(\alpha)\, x \qquad \forall x \in X \tag{4.2.14}$$

con coeficiente

$$f(\alpha) := \left[c_2 + \alpha v_1(\alpha) k_2^2\right]^{-1} \alpha v_1(\alpha) k_1 k_2. \tag{4.2.15}$$

Obsérvese que como $c_2 \geq 0$, se tiene que $|f(\alpha)| \leq k_1/k_2$, lo que garantiza que $f_\alpha^*(x)$ está en $A_0(x)$ para toda $x \in X$.

Haciendo tender α a 1 en la expresión (4.2.14)–(4.2.15) llegamos a

$$f_\alpha^*(x) \to f_0^*(x) := -f_0\, x \qquad \forall x \in X, \tag{4.2.16}$$

donde

$$f_0 := \left(c_2 + v_0 k_2^2\right)^{-1} v_0 k_1 k_2 \tag{4.2.17}$$

y v_0 es la única solución positiva de la ecuación (4.2.13) cuando $\alpha \uparrow 1$, es decir,

$$k_2^2 v_0^2 + \left(c_2 - c_1 k_2^2 - c_2 k_1^2\right) v_0 - c_1 c_2 = 0. \tag{4.2.18}$$

Veremos más adelante que $f_0^*(x)$ es una política promedio óptima.

Por otro lado, tomando $z = 0$ en (4.2.8) y (4.2.9), obtenemos:

$$\rho_\alpha := (1-\alpha) V_\alpha^*(0) = (1-\alpha) v_2(\alpha)$$

y

$$h_\alpha(x) := V_\alpha^*(0) - V_\alpha^*(x) = -v_1(\alpha) x^2.$$

De estas últimas expresiones obtenemos ρ_0 y h_1 haciendo nuevamente tender α a 1:

$$\rho_\alpha \to \rho_0 := v_0 \sigma^2 \qquad \text{y} \qquad h_\alpha(x) \to h_1(x) := -v_0 x^2. \tag{4.2.19}$$

La constante ρ_0 y la función $-h_1(\cdot)$ son solución de la ecuación de optimalidad para costo promedio. Para verificarlo, sustituimos los valores dados en (4.2.19) en la ecuación:

$$\rho_0 - h_1(x) = \min_{A_0(x)} \left[c(x,a) - \int_X h_1(y)\, Q(dy \mid x,a) \right],$$

y obtenemos

$$\begin{aligned} v_0\sigma^2 + v_0 x^2 &= \min_{A_0(x)} \left[c_1 x^2 + c_2 a^2 + v_0 \int_X (k_1 x + k_2 a + s)^2\, g(s)\, ds \right] \\ &= \min_{A_0(x)} \left[c_1 x^2 + c_2 a^2 + v_0 \left(k_1^2 x^2 + k_2^2 a^2 + 2k_1 k_2 a x + \sigma^2 \right) \right]. \end{aligned}$$

Realizando cálculos directos, se observa que la función $f_0^*(x)$ dada en (4.2.16) minimiza el lado derecho de la ecuación anterior y que la igualdad se verifica.

Al aplicar la política promedio óptima, el sistema dinámico (4.2.1) toma la forma:

$$\begin{aligned} x_{t+1} &= k_1 x_t + k_2 a_t + \xi_t \\ &= k_1 x_t - \frac{v_0 k_1 k_2^2}{c_2 + v_0 k_2^2} x_t + \xi_t, \end{aligned}$$

es decir,

$$x_{t+1} = -\frac{c_2 k_1}{c_2 + v_0 k_2^2} x_t + \xi_t,$$

donde $c_2 \geq 0$ y v_0 satisface la ecuación (4.2.18).

Ahora *debemos encontrar una solución a la ecuación de optimalidad para costo* promedio correspondiente al modelo

$$\mathfrak{M}_1 := (X, A, \{A_1(x) \mid x \in X\}, Q, h_1),$$

donde el conjunto de acciones viables en cada estado x es

$$A_1(x) = \left\{ a = \left(c_2 + v_0 k_2^2\right)^{-1} v_0 k_1 k_2 x \right\}.$$

Obsérvese que en este modelo la función de costo $h_1(x) = -v_0 x^2$ no depende de la acción. De hecho, es una función de la forma $c(x,a) = c_1 x^2 + c_2 a^2$ con $c_2 = 0$. Así que, al repetir el procedimiento anterior en el modelo $\mathfrak{M}_1$, se obtiene una política promedio óptima del tipo de la política dada en (4.2.16)–(4.2.17) pero con $c_2 = 0$, es decir,

$$f_1^*(x) := -\frac{k_1}{k_2} x \qquad \forall x \in X. \tag{4.2.20}$$

Los valores para ρ_1 y $h_2(x)$ que se encuentran en este caso son:

$$\rho_1 = -v_0\sigma^2 \quad \text{y} \quad h_2(x) = v_0 x^2,$$

y el nuevo modelo a analizar es

$$\mathfrak{M}_2 := (X, A, \{A_2(x)|x \in X\}, Q, h_2),$$

con $A_2(x) = \{a = -(k_1/k_2)\, x\}$. Como $A_2(x)$ contiene una sola acción para cada $x \in X$ (acción que depende exclusivamente de los parámetros del sistema dinámico) y los conjuntos $A_3(x)$, $A_4(x), \ldots$ son subconjuntos no vacíos de $A_2(x)$, se tiene que todos los conjuntos $A_n(x)$ con $n = 3, 4, 5, \ldots$, son iguales a $A_2(x)$ y la política Blackwell óptima es $f_1^*(x)$ dada en (4.2.20).

Obsérvese finalmente que al aplicar la política Blackwell óptima, el sistema dinámico se transforma en la sencilla expresión

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

Capítulo 5

Conclusiones y problemas abiertos

Este trabajo se enmarca en el contexto de los procesos de control Markovianos con costo promedio que se usan en una gran variedad de aplicaciones. Uno de los problemas en su aplicación consiste en que, por tratarse de un criterio límite, los costos asintóticos determinan completamente su valor sin diferenciar entre políticas que generan comportamientos muy distintos. El criterio de optimalidad en el sentido de Blackwell brinda una alternativa para resolver esta deficiencia generando políticas que contemplan tanto los primeros costos producidos como los asintóticos. El resultado principal de este trabajo consiste en encontrar condiciones bajo las cuales la optimalidad de Blackwell se puede obtener como resultado de aplicaciones sucesivas de optimalidad promedio en una sucesión infinita de modelos anidados en los conjuntos de acciones factibles para cada estado x.

Aun cuando este criterio surgió hace más de 40 años, hasta hace poco tiempo sólo se había podido demostrar la existencia de políticas Blackwell óptimas (BO) en modelos con espacios de estados y acciones finitos o numerable [2, 25, 5, 6, 4]. La primera demostración sobre existencia de políticas BO en modelos con espacio de estados de Borel, conjuntos de acciones compactos y costos no necesariamente acotados, se debe a Hordijk y Yushkevich (H-Y) y fue publicada en 1999. En trabajos previos a éste último, Yushkevich [28, 30] estudió el problema de existencia en modelos con espacio de estados de Borel, pero con espacio de acciones numerable o con costos acotados, bajo hipótesis aun más restrictivas que las usadas en [18, 19]. Aquí presentamos otra vía para asegurar la existencia de políticas BO en modelos tan generales como los usados por H-Y, que es mucho más directa y requiere hipótesis menos restrictivas. La comparación entre los procedimientos y entre las hipótesis de ambos trabajos, se presenta en la sección 3.6.

En particular, en este trabajo hemos resuelto el problema abierto que plantearon H-Y [18] relativo a la posibilidad de prescindir de una hipótesis utilizada por ellos para garantizar la existencia de políticas BO en el conjunto de todas las políticas, la hipótesis que referimos en 3.6.1(6).

Como en cualquier trabajo de investigación, un problema abierto es la posibilidad de debilitar las hipótesis que requerimos aquí para obtener el resultado principal.

Un segundo problema abierto es el de identificar las características generales de familias de modelos en los que es posible asegurar la existencia de políticas BO. En este trabajo desarrollamos un ejemplo de modelo lineal-cuadrático en el que se cumplen las condiciones para su existencia. Una de las características más útiles de este modelo es que la función de costo por etapa es una función convexa con un valor mínimo único. De aquí surge la pregunta de si es posible un tratamiento similar en todos los modelos Markovianos de control convexos.

Por último, un tercer problema es la extensión del resultado a modelos de control a tiempo continuo o a alguna clase particular de estos modelos, como los procesos de difusión.

Bibliografía

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

[2] D. Blackwell (1962): Discrete dynamic programming, *Ann. Math. Statist.* **33**, 719–726.

[3] E.V. Denardo (1971): Markov renewal programs with small interest rate, *Ann. Math. Stat.* **42**, 477–496.

[4] R. Cavazos-Cadena y J.B. Lasserre (1999): A direct approach to Blackwell optimality, *Morfismos* **3,** 9–13.

[5] R. Dekker y A. Hordijk (1988): Average, sensitive y Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards, *Math. Oper. Res.* **13**, 395–420.

[6] R. Dekker y A. Hordijk (1992): Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains, *Math. Oper. Res.* **17**, 271–289.

[7] E.B. Dynkin y A.A. Yushkevich : *Controlled Markov Processes*, Springer–Verlag, New York, 1979.

[8] A. Federgruen y P.J. Schweitzer (1984): Successive approximation methods for solving nested functional equations in Markov desicion problems, *Math. Oper. Res.* **9**, 319–344.

[9] E. Gordienko, O. Hernández-Lerma (1995): Average cost Markov control processes with weighted norms: existence of canonical policies, *Appl. Math* (Warsaw) **23**, 199-218.

[10] E. Gordienko, O. Hernández-Lerma (1995): Average cost Markov control processes with weighted norms: value iteration, *Appl. Math* (Warsaw) **23**, 219-237.

[11] O. Hernández–Lerma (1989): *Adaptive Markov Control Processes*, Springer–Verlag, New York.

[12] O. Hernández-Lerma, G. Carrasco y R. Pérez-Hernández (1999): Markov control processes with the expected total cost criterion: optimality, stability and transient models, *Acta Appl. Math.* **59**, 229-269.

[13] O. Hernández–Lerma y J.B. Lasserre (1996): *Discrete–Time Markov Control Processes: Basic Optimality Criteria*, Springer–Verlag, New York.

[14] O. Hernández–Lerma y J.B. Lasserre (1999): *Further Topics on Discrete–Time Markov Control Processes*, Springer–Verlag, New York.

[15] O. Hernández–Lerma, O. Vega–Amaya y G. Carrasco (1999) : Sample–path optimality and variance–minimization of average cost Markov control processes, *SIAM J. Control Optim.* **39**, 79-93.

[16] N Hilgert y O. Hernández-Lerma (2002). Bias optimality versus strong 0–discount optimality in Markov control processes, (sometido)

[17] A. Hordijk y K. Sladký (1977): Sensitive optimality criteria in countable state dynamic programming, *Math. Oper. Res.* **2**, 11–14.

[18] A. Hordijk y A.A. Yushkevich (1999): Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards, *Math. Meth. Oper. Res.* **49**, 1–39.

[19] A. Hordijk y A.A. Yushkevich (1999): Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards, *Math. Meth. Oper. Res.* **50**, 421-448.

[20] A. Hordijk y A.A. Yushkevich (2002): Blackwell optimality, Chapter 8 in *Handbook of Markov Decision Processes*, editado por E.A. Feinberg y A. Shwartz, Kluwer, Dordrecht, 231-267.

[21] J.B. Lasserre (1988): Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes, *J. Math. Anal Appl.* **136**, 479–489.

[22] B.L. Miller y A.F. Veinott (1969): Discrete dynamic programming with a small interest rate, *Ann. Math. Satist.* **40**, 366–370.

[23] P. Mandl (1974): Estimation and control in Markov chains, *Adv. Appl. Prob.* **6**, 40–60.

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

[25] A.F. Veinott (1969): Discrete dynamic programming with sensitive discount optimality criteria, *Ann. Math. Statist.* **40**, 1635-1660.

[26] O. Vega-Amaya y R. Montes de Oca (1998). Application of average dynamic programming to inventory systems, *Math. Meth. Oper. Res.* **47**, 451-471.

[27] O. Vega-Amaya (2002) The average cost optimality equation: a fixed point approach, (sometido).

[28] A.A. Yushkevich (1994): Blackwell optimal policies in a Markov decision process with a Borel state space, *Z. Oper. Res.* **40**, 253–288.

[29] A.A. Yushkevich (1995): Strong 0–discount optimal policies in a Markov decision process with a Borel state space, *Math. Meth. Oper. Res.* **42** , 93–108.

[30] A.A. Yushkevich (1997): Blackwell optimality in Borelian continuous–in–action Markov decision processes, *SIAM . Control Optim.* **35**, 2157–2182.