Boletín de la Sociedad Matemática Mexicana Vol. 32 No. 1, 1987

CONTROL DE CADENAS DE MÁRKOV CON PARÁMETROS DESCONOCIDOS Y ESPACIO DE ESTADOS MÉTRICO

Por ROBERTO S. ACOSTA ABREU[1]

Resumen

En este trabajo consideramos el problema de determinar políticas (o estrategias) para el control óptimo de sistemas markovianos que dependen de parámetros desconocidos, cuando se usa el criterio de ganancia promedio y cuando el espacio de estados es un espacio métrico separable y completo. Demostramos que la política de control basada en un método de iteracíon de valores es óptima en ganancia promedio, y obtenemos un esquema iterativo para calcular el valor óptimo de esta política.

1. Introducción

En la referencia [1] consideramos el problema de determinar políticas o estrategias adaptables para el control óptimo de sistemas markovianos que dependen de parámetros desconocidos, con un espacio de estados numerable cuando se usa el criterio de ganancia promedio. En el presente trabajo extenderemos los resultados de [1] al caso en que el espacio de estados es un espacio polaco (i.e. métrico separable y completo). Usaremos aquí como en [1], un esquema de control adaptable que combina el proceso de estimación del parámetro desconocido con las ecuaciones de iteración de valores. A este esquema lo llamaremos política IV, y veremos que es una estrategia óptima con el criterio considerado. Kurano [9] y Mandl [11] introdujeron el "método de substituir las estimaciones en controles estacionarios óptimos", conocido también como "principio de estimación y control" (PEC) para espacios finitos y Georgin en [3] generalizó este método al caso de espacio de estados polaco. En el método PEC hay que conocer de antemano una política estacionaria óptima para todos los valores posibles del parámetro; la política IV no requiere de este conocimiento y procede en forma recursiva con las ecuaciones de iteración de valores para determinar los controles óptimos del sistema en cada etapa.

Hernández-Lerma en [6] ha usado políticas del tipo de la IV en problemas de programación dinámica descontada.

Este trabajo lo organizamos como sigue. En esta sección 1 damos la notación y terminología que usaremos; en la sección 2, presentamos el modelo de decisión markoviano con parámetros desconocidos y planteamos los problemas que queremos resolver; en la sección 3 introducimos hipótesis adicionales que nos permiten tratar con la ecuación de optimalidad; en la sección 4 consideramos las ecuaciones de iteración de valores; en la sección 5 definimos las políticas adaptables óptimas, damos nuestros resultados principales y un

[1] Becario de COFAA-IPN

ejemplo de control de inventarios; finalmente en la sección 6 demostramos los resultados de la sección 5.

Damos ahora la notación que usaremos en todo el trabajo.

Notación y terminología

En un espacio métrico X consideramos siempre la σ-álgebra $\mathcal{B}(X)$ de sus subconjuntos de Borel. Denotaremos por $B(X)$ [y $C(X)$, respectivamente] al conjunto de todas las funciones reales, acotadas, definidas sobre X y Borel medibles [continuas, respectivamente] con la norma del supremo $\| g \| = \sup_x | g(x) |$, para g en $B(X)$ ó $C(X)$. El producto cartesiano de X e Y lo denotaremos por XY. Sea X un espacio polaco (o sea un espacio métrico separable y completo); por $\mathbf{P}(X)$ indicaremos al conjunto de todas las medidas de probabilidad sobre el espacio $(X, \mathcal{B}(X))$; con la métrica correspondiente a la topología de la convergencia débil, $\mathbf{P}(X)$ es un espacio polaco (ver ref. [2]). Si Y es algún otro espacio polaco, por $\mathbf{P}(Y \mid X)$ denotaremos al espacio de todas las probabilidades de transición sobre Y dado X; esto es, $p(dy \mid x)$ pertenece a $\mathbf{P}(Y \mid X)$ si para cada $x \in X$ fijo, $p(\cdot \mid x)$ es una medida de probabilidad sobre Y y para cada $B \in \mathcal{B}(Y)$ fijo, $p(B \mid \cdot)$ es una función Borel medible sobre X. La probabilidad de transición $p(dy \mid x)$ es continua, si es continua en el sentido de la convergencia débil, es decir, si para cualquier función $g \in C(Y)$, se tiene que $\int g(y)p(dy \mid x)$ es una función continua sobre X. La norma $\| \cdot \|_v$ denotará la norma de variación total de las medidas con signo involucradas. Abreviaremos "casi seguramente" por c.s.

2. Procesos de decisión de Márkov con parámetros desconocidos

Consideremos un proceso de decisión de Márkov (PDM) en el que el conjunto de tiempos de decisión es $\mathbf{N} = \{0, 1, 2, \ldots\}$ y en el que las probabilidades de transición y la función de ganancia dependen de un parámetro desconocido. Nuestro modelo está dado por

$$(S, A, T, r(\theta), p(\theta))$$

para el cual supondremos lo siguiente.

Hipótesis (2.1).

a) S es el espacio de estados el cual supondremos que es un espacio polaco.

b) A es el espacio polaco de acciones o controles. Para cada $x \in S$, denotamos por $A(x)$ al conjunto no vacío de acciones admisibles en el estado x. Supondremos que el conjunto

$$K := \{(x, a) \mid x \in S,\ a \in A(x)\},$$

es medible en SA.

c) T es el espacio polaco de parámetros.

d) La función $r(\theta) = r(x, a, \theta)$ es la función de ganancia esperada en una etapa; supondremos que $r : KT \to \mathbf{R}$ es medible y acotada.

e) $p(\theta) = p(\cdot \mid x, a, \theta) \in \mathbf{P}(S \mid KT)$ es la probabilidad de transición del proceso.

Para cada valor $\theta \in T$ fijo del parámetro el proceso evoluciona en el tiempo como sigue. Se observa el sistema en los valores del tiempo $n = 0, 1, 2, \ldots$. Si al tiempo n el sistema está en el estado $x \in S$ y se escoge la acción $a \in A(x)$ se obtiene una ganancia esperada $r(x, a, \theta)$ y el sistema se mueve al tiempo $n+1$ al siguiente estado de acuerdo con la probabilidad de transición $p(\cdot \mid x, a, \theta)$. Después que ocurre la transición, se observa el nuevo estado, se escoge otra vez una acción y se repite el proceso.

Sea $\theta \in T$ fijo. Para $n \in \mathbf{N}$, sean x_n y a_n el estado y la acción al tiempo n y sea $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$ la historia del proceso hasta el tiempo n. Escribimos $H_0 = S$, $H_{n+1} = KH_n$, $H_\infty = KK\cdots$. Una política es una sucesión $\delta = \{\delta_n\}$ tal que $\delta_n \in \mathbf{P}(A \mid H_n)$ y $\delta_n(A(x_n) \mid h_n) = 1$ para todo $h_n \in H_n$ y $n \in \mathbf{N}$. Una política se llama markoviana si $\delta_n(\cdot \mid h_n) = \delta_n(\cdot \mid x_n)$, $n \geq 0$. Una política markoviana se llama estacionaria (determinista) si $\delta_n(\cdot \mid x) \equiv \delta(\cdot \mid x)$ para todo $n \in \mathbf{N}$, y $x \in S$ y si existe una función Borel medible $f : S \to A$ tal que $f(x) \in A(x)$ para todo $x \in S$ y $\delta(\{f(x)\} \mid x) = 1$, para todo $x \in S$.

Sea $\mathbf{F} := \{f : S \to A \mid f$ es Borel medible y $f(x) \in A(x)$ para todo $x \in S\}$. Identificaremos con $\mathbf{F}$ al conjunto de todas las políticas estacionarias deterministas y llamaremos D al conjunto de todas las políticas.

Para cada valor fijo $\theta \in T$ del parámetro cada política $\delta \in D$ y cada estado inicial $x_0 = x$ existe una medida de probabilidad $P_x^{\delta,\theta}$ sobre el espacio $\Omega := H_\infty$ de todas las realizaciones posibles del proceso tal que para todo $x \in S$, $h_n \in H_n$, $a_n \in A(x_n)$, $n \in \mathbf{N}$ y para todo $A_1 \in \mathcal{B}(A)$, $B \in \mathcal{B}(S)$,

$$P_x^{\delta,\theta}(x_0 = x) = 1,$$
$$P_x^{\delta,\theta}(a_n \in A_1 \mid h_n) = \delta_n(A_1 \mid h_n)$$

y

$$P_x^{\delta,\theta}(x_{n+1} \in B \mid h_n, a_n) = p(B \mid x_n, a_n, \theta).$$

(Ver por ej. [2] u [8]).

Denotaremos por $E_x^{\delta,\theta}$ a la esperanza con respecto a $P_x^{\delta,\theta}$.

Definimos ahora, para cualquier política $\delta \in D$, cualquier estado inicial $x \in S$ y cualquier $\theta \in T$ fijo,

$$J(\delta, x, \theta) := \liminf_{n\to\infty} n^{-1} E_x^{\delta,\theta}[\sum_{i=0}^{n-1} r(x_i, a_i, \theta)]$$

La función $J(\delta, x, \theta)$ es la ganancia esperada promedio a la larga por unidad de tiempo, cuando se usa δ, $x_0 = x$ y $\theta \in T$. Una política δ^* se llama óptima (en ganancia promedio) si

$$J(\delta^*, x, \theta) = \sup_{\delta \in D} J(\delta, x, \theta), \quad x \in S, \quad \theta \in T.$$

Nuestro problema consiste en determinar políticas óptimas en ganancia promedio en el caso en el que el valor verdadero del parámetro $\theta^* \in T$ es una constante desconocida. Queremos también determinar un método para calcular el valor óptimo $J(\delta^*, x, \theta^*)$.

3. Hipótesis adicionales y la ecuación de optimalidad

Las condiciones que introducimos en esta sección nos aseguran la existencia de políticas estacionarias óptimas para cada valor fijo del parámetro.

Hipótesis (3.1).
a) El conjunto de acciones A es compacto; para cada $x \in S$ el conjunto $A(x) \subset A$ es cerrado; el conjunto K es cerrado en SA, lo que es equivalente (véase [10]) a suponer que la correspondencia $x \to A(x)$ es semicontinua superiormente; también supondremos que $x \to A(x)$ es semicontinua inferiormente, es decir: si $\{x_n\}$ es una sucesión en S tal que $x_n \to x$ y si $a \in A(x)$, entonces existe una sucesión $\{a_n\}$ en A con $a_n \in A(x_n)$ tal que $a_n \to a$. Estas condiciones implican que $x \to A(x)$ es continua (véase [10]).
b) La probabilidad de transición $p(\cdot \mid x, a, \theta)$ es débilmente continua en KT.
c) La función r es continua en KT y existe una constante L tal que $\mid r(x, a, \theta) \mid \leq L$ para todo $(x, a, \theta) \in KT$.
d) Para cada $x \in S$, $\theta \in T$ y $h \in B(S)$, se tiene que $\int_S h(y)p(dy \mid x, a, \theta)$ es una función continua de a en $A(x)$.

El estudio de políticas óptimas lo haremos usando la llamada ecuación de optimalidad dada en la proposición siguiente.

PROPOSICIÓN (3.2). *Supongamos que se cumplen las hipótesis* (2.1) *y* (3.1), *y que existen funciones* $g \in B(T)$ *y* $v \in B(ST)$ *que satisfacen la ecuación*

$$g(\theta) + v(x, \theta) = \max_{a \in A(x)} \{r(x, a, \theta) + \int_S v(y, \theta)p(dy \mid x, a, \theta)\} \tag{3.3}$$

para todo $x \in S$ *y* $\theta \in T$. *Entonces tenemos*:

i) $\sup_{\delta \in D} J(\delta, x, \theta) \leq g(\theta)$, *para todo* $\theta \in T$ *y todo* $x \in S$.
ii) *Si para cada* $\theta \in T$, $f_\theta \in \mathbf{F}$ *es una política determinista tal que*

$$g(\theta) + v(x, \theta) = r(x, f_\theta(x), \theta) + \int_S v(y, \theta)p(dy \mid x, f_\theta(x), \theta), \quad x \in S.$$

entonces f_θ *es óptima y* $g(\theta) = J(f_\theta, x, \theta)$ *para todo* $x \in S$ y $\theta \in T$.
iii) *Sea*

$$\phi(x, a, \theta) := r(x, a, \theta) + \int_S v(y, \theta)p(dy \mid x, a, \theta) - v(x, \theta) - g(\theta)$$

definida para todo $(x, a) \in K$ *y* $\theta \in T$. *Entonces para cualquier política* $\delta \in D$, $x \in S$ y $\theta \in T$,

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} r(x_i, a_i, \theta) = g(\theta) \qquad P_x^{\delta,\theta}\text{-c.s.},$$

si y sólo si

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \phi(x_i, a_i, \theta) = 0 \qquad P_x^{\delta,\theta}\text{-c.s.}$$

La ecuación (3.3) se llama *ecuación de optimalidad* (EO) para el criterio de ganancia promedio.

Si $g \in B(T)$ y $v \in B(ST)$ son funciones que satisfacen la EO, se dice que $\{g, v\}$ es una solución de la EO.
La demostración de la proposición (3.2) puede verse en [3] ó [4].

Introducimos ahora algunas condiciones de ergodicidad bajo las cuales veremos que hay una solución de la EO y por lo tanto, por la proposición (3.2), también hay políticas estacionarias óptimas.

Hipótesis (3.4)

Existe un número $\alpha < 1$ tal que para todo $\theta \in T$,

$$\sup\{\| p(\cdot \mid x, a, \theta) - p(\cdot \mid y, b, \theta) \|_v : \ (x, a) \in K, \ (y, b) \in K\} = 2\alpha$$

donde $\| \cdot \|_v$ denota la norma de variación total para medidas con signo.

PROPOSICIÓN (3.5). *Cada una de las siguientes condiciones implica que se cumple la hipótesis* (3.4).
(3.4)$'$: *Existe un estado* $x^* \in S$ *y un número real* $\alpha_1 > 0$ *tal que para todo* $(x, a, \theta) \in KT$,

$$p(\{x^*\} \mid x, a, \theta) \geq \alpha_1$$

(3.4)$''$: *Existe una medida de transición* $\mu(\cdot \mid \theta)$ *sobre* S *dado* T, *débilmente continua en* T *y un número real* $\alpha_2 > 0$ *tal que* $\mu(S \mid \theta) \geq \alpha_2$ *para todo* $\theta \in T$ *y*

$$p(\cdot \mid x, a, \theta) \geq \mu(\cdot \mid \theta) \text{ para todo } (x, a, \theta) \in KT.$$

(3.4)$'''$: *Existe una medida de transición* $\nu(\cdot \mid \theta)$ *sobre* S *dado* T *débilmente continua en* T *y un número real* $\alpha_3 < 2$ *tal que* $\nu(S \mid \theta) \leq \alpha_3$ *para todo* $\theta \in T$ *y*

$$p(\cdot \mid x, a, \theta) \leq \nu(\cdot \mid \theta) \text{ para todo } (x, a, \theta) \in KT.$$

La demostración de la proposición anterior se puede ver en [3].

En la sección 6 demostraremos el siguiente resultado.

PROPOSICIÓN (3.6). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) y (3.4). *Entonces existen funciones* $g \in C(T)$ *y* $v \in C(ST)$ *tales que* g *y* v *satisfacen la ecuación de optimalidad.*

En las referencias [3] y [14] se dan otras condiciones de ergodicidad bajo las cuales la ecuación de optimalidad tiene solución.

Cuando se usa una política $f \in \mathbf{F}$ el proceso $\{x_n\}$ es una cadena de Márkov con probabilidad de transición $p_f(\cdot \mid x, \theta) = p(\cdot \mid x, f(x), \theta)$. La probabilidad de transición de n pasos de este proceso se define recursivamente por

$$p_f^n(B \mid x, \theta) = \int_S p_f^{n-1}(B \mid y, \theta) p_f(dy \mid x, \theta), \quad B \in \mathcal{B}(S), \quad n \geq 1,$$

donde $p_f^0(B \mid x, \theta) = 1_B(x)$ es la medida de probabilidad concentrada en x, es decir, $1_B(x) = 1$ si $x \in B$, $1_B(x) = 0$ si $x \notin B$.

La demostración de la siguiente proposición se puede ver en [3] ó [4].

PROPOSICIÓN (3.7). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Para cualquier política estacionaria* $f \in \mathbf{F}$, *existe una medida de probabilidad* $q_f(\cdot \mid \theta)$ *sobre* S *tal que*

$$\| p_f^n(\cdot \mid x, \theta) - q_f(\cdot \mid x, \theta) \|_v \leq 2\alpha^n, \text{ para todo } x \in S, \quad n \in \mathbf{N}.$$

La cadena de Márkov con probabilidad de transición $p_f(\cdot \mid x, \theta)$ es aperiódica y recurrente en el sentido de Doeblin.

Haremos uso también de las siguientes hipótesis de continuidad uniforme en el parámetro.

Hipótesis (3.8)

Supongamos que r y p satisfacen que, para todo $\theta \in T$

(a) $\sup_{(x,a) \in K} | r(x, a, \theta) - r(x, a, \theta') | \to 0$, cuando $\theta' \to \theta$;

(b) $\sup_{(x,a) \in K} \| p(\cdot \mid x, a, \theta) - p(\cdot \mid x, a, \theta') \|_v \to 0$, cuando $\theta' \to \theta$.

4. Las ecuaciones de iteración de valores

Definimos ahora el operador de valor óptimo U sobre $B(ST)$ por

$$Uv(x, \theta) := \max_{a \in A(x)} \{ r(x, a, \theta) + \int_S v(y, \theta) p(dy \mid x, a, \theta) \}, \tag{4.1}$$

para todo $(x, \theta) \in ST$, y para cualquier función $v \in B(ST)$.

Usando los Lemas 1 y 2 de [5] podemos ver que bajo las hipótesis (2.1) y (3.1), si $v \in B(ST)$ [ó $v \in C(ST)$] entonces $Uv \in B(ST)$ [ó $Uv \in C(ST)$, respectivamente].

Definimos la sucesión de funciones $\{V_n\}$ sobre ST por:

$$V_0 \in C(ST) \text{ arbitraria}$$
$$V_n = UV_{n-1}, \quad n \geq 1. \tag{4.2}$$

A las ecuaciones en (4.2) las llamamos *ecuaciones de iteración de valores*. De la definición de V_n y de las propiedades de $P_x^{\delta,\theta}$ obtenemos que:

$$V_n(x,\theta) = \sup_{\delta \in \mathbf{F}^\infty} E_x^{\delta,\theta}[\sum_{j=0}^{n-1} r(x_j, a_j, \theta) + V_0(x_n, \theta)] \tag{4.3}$$

donde $\mathbf{F}^\infty = \mathbf{FF}\cdots$. Así, $V_n(x,\theta)$ se puede interpretar como la ganancia máxima total esperada para un sistema que se mueve durante n etapas cuando el estado inicial es x y dado que se obtiene una ganancia terminal $V_0(x_n,\theta)$, al terminar en el estado x_n.

Observación (4.4). Bajo las hipótesis (2.1) y (3.1), por inducción obtenemos que las funciones V_n pertenecen a $C(ST)$ para cada $n \in \mathbf{N}$.

5. Políticas adaptables óptimas

En esta sección definimos algunas políticas adaptables en las que combinamos el proceso de estimación del parámetro desconocido con las ecuaciones de iteración de valores ó la ecuación de optimalidad.

Para cada $n \in \mathbf{N}$ sea H_n el conjunto de historias observadas hasta el tiempo n. Para estimar el parámetro desconocido se requiere de un esquema de estimación suficientemente "robusto" en el sentido de la siguiente hipótesis.

Hipótesis (5.1)

Para cualquier $\theta \in T$ existe una sucesión $\{\hat{\theta}_n\}$ de funciones medibles $\hat{\theta}_n : H_n \to T$ tal que $\hat{\theta}_n \to \theta$, $P_x^{\delta,\theta}$-c.s. cuando $n \to \infty$, para cualquier $x \in S$, y cualquier política δ. La sucesión $\{\hat{\theta}_n\}$ se llama una sucesión *fuertemente consistente* (FC) de estimadores de θ.

Nota (5.2). En las referencias [3,9,11] se obtiene una sucesión de estimadores FC de θ usando el método de mínimo contraste.

En lo que sigue supondremos que θ^* es el valor verdadero del parámetro desconocido y que $\{\hat{\theta}_n\}$ es una sucesión FC de estimadores de θ^*.

Consideremos ahora la sucesión de iteración de valores $\{V_n\}$ definida en (4.2). Definimos la sucesión de funciones medibles $\{f_n\}$, $f_n : ST \to A$, por: $f_0 \in \mathbf{F}$ arbitraria, y para $n \geq 1$, $f_n(x,\theta)$es cualquier acción que maximiza al lado derecho de (4.2), es decir,

$$V_n(x,\theta) = r(x, f_n(x,\theta), \theta) + \int_S V_{n-1}(y,\theta) p(dy \mid x, f_n(x,\theta), \theta), \tag{5.3}$$

para todo $(x, \theta) \in ST$.

Nota. Bajo las hipótesis (2.1) y (3.1), se puede ver, usando los resultados en [5] ó [7] que existen funciones medibles que satisfacen la ecuación (5.3).

Definición (5.4.). Sea $\{f_n\}$ la sucesión de funciones definida en la ecuación (5.3). A la política $\hat{d} = \{\hat{d}_n\}$ definida por $\hat{d}_n(h_n) = f_n(x_n, \hat{\theta}_n)$, $n \in \mathbf{N}$, la llamamos *política adaptable de iteración de valores* (IV).

Establecemos ahora nuestros resultados principales.

TEOREMA (5.5). *Supongamos que se cumplen las hipótesis* (2.1), (3.1), (3.4), (3.8) y (5.1) *Supongamos además que la función* v *dada en la proposición* (3.6) *satisface la siguiente condición*:

$$\sup_{x \in S} | v(x, \theta') - v(x, \theta) | \to 0 \quad \textit{cuando } \theta' \to \theta, \quad \textit{para todo } \theta \in T. \tag{5.6}$$

Entonces la política adaptable de iteración de valores $\hat{d}$ *es óptima en ganancia promedio. Además el valor óptimo* $g(\theta^*)$ *se puede obtener como el límite*

$$g(\theta^*) = \lim_{n \to \infty} g_n(\theta_n) \tag{5.7}$$

donde $\{g_n\}$ *es la sucesión definida por*

$$g_n(\theta) := V_n(x^*, \theta) - V_{n-1}(x^*, \theta), \quad \theta \in T, \quad n \geq 1,$$

en la que $\{V_n\}$ es la sucesión de iteración de valores (4.2), x^* es cualquier estado fijo en S y $\{\theta_n\}$ es cualquier sucesión en T convergente a θ^*.

Para el caso especial en que S y T son compactos tenemos el resultado siguiente.

TEOREMA (5.9). *Supongamos que* S *y* T *son compactos y que se cumplen las hipótesis* (2.1), (3.1), (3.4), *y* (5.1). *Entonces la política adaptable* $\hat{d}$ *es óptima y el valor óptimo* $g(\theta^*)$ *se puede obtener como en* (5.8).

La demostración de estos resultados la haremos en la sección 6. El teorema (5.5) resuelve los problemas planteados en la parte final de la sección 2.

Definamos ahora otras políticas adaptables.

Supongamos que se cumplen las hipótesis (2.1), (3.1) y la hipótesis (3.4)''. Definimos para todo $u \in B(ST)$ el operador U' por

$$U'u(x, \theta) := \max_{a \in A(x)} \{r(x, a, \theta) + \int_S u(y, \theta) p'(dy \mid x, a, \theta)\}, \quad (x, \theta) \in ST, \tag{5.10}$$

donde

$$p'(\cdot \mid x, a, \theta) := p(\cdot \mid x, a, \theta) - \mu(\cdot \mid \theta). \tag{5.11}$$

Si $u \in C(ST)$ entonces $U'u \in C(ST)$ (ver Lemas 1 y 2 en [5]). U' es un operador de contracción sobre $C(ST)$, con la norma uniforme (véase [5]). Definimos la sucesión de aproximaciones sucesivas $\{w'_n\}$ por: $w'_0 \in C(ST)$ arbitraria, y

$$w'_{n+1} = U'w'_n, \qquad n \geq 1. \tag{5.12}$$

Definición (5.13). Supongamos que se cumplen las hipótesis (2.1), (3.1) y (3.4)''. Definimos una sucesión de funciones $\{f'_n\}$, $f'_n : ST \to A$, tal que $f'_n(x, \theta)$ es cualquier acción que maximiza el lado derecho de (5.12). Definimos la política $d' = \{d'_n\}$ por $d'_n(h_n) = f'_n(x_n, \hat{\theta}_n)$, $n \in \mathbf{N}$ que llamaremos *política adaptable de aproximaciones sucesivas*.

Sea ahora $f : ST \to A$ una función medible tal que $f(x, \theta)$ es cualquier acción que maximiza el lado derecho de la ecuación de optimalidad, es decir,

$$g(\theta) + v(x, \theta) = r(x, f(x, \theta), \theta) + \int_S v(y, \theta) p(dy \mid x, f(x, \theta), \theta), \tag{5.14}$$

para todo $(x, \theta) \in ST$.

Nota. Bajo las hipótesis (2.1), (3.1) y (3.4) el Teorema 2 de [7] implica la existencia de f [véase también la proposición (3.6)].

Definición (5.15). Sea f la función definida en (5.14). La política $\bar{d} = \{\bar{d}_n\}$ definida por $\bar{d}_n(h_n) = f(x_n, \theta_n)$, $n \in \mathbf{N}$, se llama *política adaptable de estimación y control* (PEC).

Observación (5.16). Notemos que para definir la política PEC en (5.15) hay que conocer de antemano el valor óptimo $g(\theta)$ y $v(x, \theta)$ para cada $\theta \in T$ y $x \in S$. La política IV definida en (5.4) y la política de iteraciones sucesivas de (5.13) no requieren de esto y tienen además la ventaja de dar esquemas de control iterativos.

Observación (5.17). Para las políticas definidas en (5.13) y (5.15) se cumple también un resultado análogo al teorema (5.5).

Damos ahora un ejemplo de control de inventarios.

Ejemplo (5.18). Consideremos el problema de control de inventario de un producto que se trata en la referencia [3]. Se tiene una capacidad finita C para el nivel de inventario. En el tiempo $n = 0, 1, 2, \ldots$, sean x_n el nivel de inventario, a_n la cantidad ordenada al principio del intervalo $[n, n+1)$ y d_n la demanda durante dicho intervalo. Supongamos que $\{d_n\}$ es una sucesión de

variables aleatorias con esperanza finita, no negativas, independientes, todas con la misma función de densidad $h(\theta,\cdot) : \mathbf{R} \to \mathbf{R}$, la cual suponemos que depende continuamente de un parámetro desconocido $\theta \in T$ con T un conjunto compacto. El espacio de estados S es el intervalo compacto $[0, C]$. El conjunto de acciones es $A = [0, C]$, y el conjunto de acciones admisibles en el estado x es $A(X) = [0, C - x]$; la correspondencia $x \to A(x)$ es continua: en efecto el conjunto

$$K = \{(x, a) : 0 \leq x \leq C,\ 0 \leq a \leq C - x\}$$

es cerrado en SA y la correspondencia $x \to A(x)$ es semicontinua inferiormente (ver [10]). El producto que se vende durante $[n, n+1)$ lo tomaremos como el $\min(d_n, x_n + a_n)$, de manera que el estado al tiempo $n+1$ es:

$$x_{n+1} = (x_n + a_n - d_n)^+.$$

La ganancia durante $[n, n+1)$ es

$$r_n = b_1(x_n + a_n - x_{n+1}) - b_2 a_n - b_3(x_n + a_n)$$

donde b_1 es el precio unitario de venta, b_2 es el precio unitario de compra (ó de producción) y b_3 es un costo unitario de mantenimiento. Se supone que b_1, b_2 y b_3 son constantes, $b_1 > b_2 > 0$ y $b_3 > 0$. La ganancia esperada en cada etapa es: para $x \in S$, $a \in A(x)$ y $\theta \in T$,

$$\begin{aligned} r(x, a, \theta) &= E(r_n \mid x_n = x,\ a_n = a, \theta) \\ &= \int_0^\infty [b_1 \min(u, x + a) - b_2 a - b_3(x + a)] h(\theta, u)\, du. \end{aligned}$$

La probabilidad de transición en este sistema está dada por

$$p(B \mid x, a, \theta) = \int_B h(\theta, x + a - u)\, du,$$

si B es cualquier conjunto de Borel en $(0, C]$ y

$$p(\{0\} \mid x, a, \theta) = \int_{x+a}^\infty h(\theta, u)\, du$$

para $(x, a,) \in K$, $\theta \in T$.

Verifiquemos la hipótesis (3.1). La función $r(x, a, \theta)$ es acotada:

$$\mid r(x, a, \theta) \mid \leq b_1 E d_n + (b_1 + b_2 + b_3) C,$$

donde $E d_n < \infty$ es la esperanza de la demanda. Como $h(\theta, u)$ depende continuamente de θ y es acotada, entonces

$$r(x, a, \theta) = \int_0^{x+a} b_1 u h(\theta, u)\, du + b_1(x + a) \int_{x+a}^\infty h(\theta, u)\, du - b_2 a - b_3(x + a)$$

es una función continua de x, a y θ, para $(x,a) \in K$ y $\theta \in T$. Si $g \in C(S)$

$$\int_S g(y)p(dy \mid x,a,\theta) = g(0)\int_{x+a}^{\infty} h(\theta,u)du + \int_{(0,C]} g(u)h(\theta,x+a-u)du$$

es una función continua de (x,a,θ) para $(x,a) \in K$ y $\theta \in T$ (ver [13]), por tanto $p(\cdot \mid x,a,\theta)$ es débilmente continua para $(x,a) \in K$ y $\theta \in T$. Además $p(\cdot \mid x,a,\theta)$ satisface la hipótesis de ergodicidad (3.4)': con $x^* = 0$ y para $(x,a,) \in K$ y $\theta \in T$,

$$\begin{aligned} p(\{0\} \mid x,a,\theta) &= \int_{x+a}^{\infty} h(\theta,u)du \\ &\geq \int_{C}^{\infty} h(\theta,u)du =: \alpha_1, \end{aligned}$$

donde suponemos que $h(\theta,u)$ es tal que α_1 es mayor que cero.

Sea θ^* el valor verdadero del parámetro y supongamos que se cumple la hipótesis (5.1); en los casos "usuales" (densidades gama, normal,...) la estimación de θ^* se puede hacer por algún método de mínimo contraste. Si $\{\hat{\theta}_n\}$ es una sucesión FC de estimadores de θ^*, podemos entonces aplicar cualquiera de las políticas iterativas de esta sección.

6. Demostración de los resultados principales

Introducimos la siguiente notación. Para cada $\theta \in T$ definimos la "seminorma de apertura" para cada $u \in B(ST)$ por

$$\| u(\cdot,\theta) \|_a := \sup_{x\in S} u(x,\theta) - \inf_{x\in S} u(x,\theta). \tag{6.1}$$

En el siguiente resultado veremos que el operador U definido en (4.1) tiene una propiedad de contracción con respecto a la seminorma (6.1).

PROPOSICION (6.2). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Sean* v *y* w *en* $B(ST)$. *Entonces*

$$\| (Uv - Uw)(\cdot,\theta) \|_a \leq \alpha \| (v-w)(\cdot,\theta) \|_a \text{ . para todo } \theta \in T. \tag{6.3}$$

Demostración. Sean $x_1, x_2 \in S$ y $\theta \in T$ fijo. De la definición de U obtenemos

$$\begin{aligned} Uv(x_1,\theta) - Uw(x_1,\theta) &\leq \max_{a\in A(x_1)} \{\int_S (v-w)(y,\theta)p(dy \mid x_1,a,\theta)\}, \\ -(Uv(x_2,\theta) - Uw(x_2,\theta)) &\leq \max_{a\in A(x)} \{\int_S (w-v)(y,\theta)p(dy \mid x_2,a,\theta)\}, \\ &= -\min_{a\in A(x_2)} \{\int_S (v-w)(y,\theta)p(dy \mid x_2,a,\theta)\}. \end{aligned}$$

De aquí obtenemos:

$$
\begin{aligned}
(6.4)\qquad & Uv(x_1,\theta) - Uw(x_1,\theta) - (Uv(x_2,\theta) - Uw(x_2,\theta)) \\
& \leq \max_{a\in A(x_1)} \{\int_S (w-v)(y,\theta)p(dy \mid x_1,a,\theta)\} \\
& \quad - \min_{a\in A(x_2)} \{\int_S (v-w)(y,\theta)p(dy \mid x_2,a,\theta)\} \\
= & \max_{a_1\in A(x_1),a_2\in A(x_2)} \mid \int_S (v-w)(y,\theta)[p(dy \mid x_1,a_1,\theta) - p(dy \mid x_2,a_2,\theta)] \mid
\end{aligned}
$$

Denotemos ahora por ψ a la medida con signo

$$\psi(\cdot) := p(\cdot \mid x_1,a_1,\theta) - p(\cdot \mid x_2,a_2,\theta)$$

Por la hipótesis (3.4) se tiene que $\| \psi \|_v \leq 2\alpha$. Por el Teorema de descomposición de Hahn-Jordan (ver [13]), existen conjuntos medibles ajenos S^+ y S^- tales que $S = S^+ \cup S^-$ y $\| \psi \|_v = \psi(S^+) - \psi(S^-)$. Como $\psi(S) = \psi(S^+) + \psi(S^-) = 0$, tenemos que $\| \psi \|_v = 2\psi(S^+) \leq 2\alpha$, de lo cual se obtiene que $\psi(S^+) \leq \alpha$. Además se tiene:

$$
\begin{aligned}
\int_S (v-w)(y,\theta)\psi(dy) &= \int_{S^+} (v-w)(y,\theta)\psi(dy) + \int_{S^-} (v-w)(y,\theta)\psi(dy) \\
&\leq \sup_x (v-w)(x,\theta)\psi(S^+) + \inf_x (v-w)(x,\theta)\psi(S^-) \\
&= [\sup_x (v-w)(x,\theta) - \inf_x (v-w)(x,\theta)]\psi(S^+) \\
&\leq \alpha \| (v-w)(\cdot,\theta) \|_a .
\end{aligned}
$$

Aplicando este resultado a la desigualdad (6.4) y tomando en cuenta que x_1 y x_2 son arbitrarios obtenemos la desigualdad (6.3). □

Usando las hipótesis (2.1) y (3.1) podemos definir inductivamente las potencias $U^n v(x,\theta)$ para cualquier $n \in \mathbf{N}$ y $v \in B(ST)$ por: $U^1 v = Uv$ y para $n \geq 2$, $U^n v = U(U^{n-1} v)$.

Para las ecuaciones de iteración de valores, tenemos $V_n = U^n V_0$, $n \in \mathbf{N}$. De la proposición (6.2), por inducción obtenemos el siguiente resultado:

PROPOSICION (6.5). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Entonces para* v *y* $w \in B(ST)$ *arbitrarias, se cumple que*:

$$(6.6)\qquad \| (U^n v - U^n w)(\cdot,\theta) \|_\alpha \leq \alpha^n \| (v-w)(\cdot,\theta) \|_a, \text{ para todo } \theta \in T.$$

Consideremos ahora un estado fijo $x^* \in S$ y definamos para cada $n \in \mathbf{N}$,

$$(6.7)\qquad v_n(x,\theta) := V_n(x,\theta) - V_n(x^*,\theta) \qquad (x,\theta) \in ST,$$

donde $\{V_n\}$ es la sucesión de iteración de valores definida en (4.2). La función v_n se llama función de valor relativo en la etapa n.

El resultado siguiente se debe a Rieder [12]:

PROPOSICION (6.8). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Entonces la sucesión* $\{v_n\}$ *es uniformemente acotada, existe* $v := \lim_{n\to\infty} v_n$ *y*

$$\sup_{x\in S} |\, (v - v_n)(x,\theta) \,| \leq B\alpha^n, \text{ para todo } \theta \in T, \tag{6.9}$$

donde $B = \sup_{\theta\in T} \| (V_1 - V_0)(\cdot,\theta) \|_a /(1-\alpha) < \infty$.

Demostración. Sea $k \in \mathbf{N}$. Por la proposición (6.5) tenemos

$$\begin{aligned} \| (v_{n+k} - v_n)(\cdot,\theta) \|_a &= \| (V_{n+k} - V_n)(\cdot,\theta) \|_a \\ &= \| (U^n V_k - U^n V_0)(\cdot,\theta) \|_a \\ &\leq \alpha^n \| (V_k - V_0)(\cdot,\theta) \|_\alpha . \end{aligned}$$

Por otro lado,

$$\begin{aligned} \| (V_k - V_0)(\cdot,\theta) \|_a &\leq \sum_{n=0}^{k-1} \| (U^n V_1 - U^n V_0)(\cdot,\theta) \|_a \\ &\leq \| (V_1 - V_0)(\cdot,\theta) \|_a \sum_{n=0}^{\infty} \alpha^n \\ &= \| (V_1 - V_0)(\cdot,\theta) \|_a /(1-\alpha). \end{aligned}$$

Además, la sucesión $\{v_n\}$ satisface que,

$$\inf_{x\in S}(v_{n+k} - v_n)(x,\theta) \leq 0 \leq \sup_{x\in S}(v_{n+k} - v_n)(x,\theta),$$

por consiguiente:

$$\begin{aligned} \sup_{x\in S} |\, (v_{n+k} - v_n)(x,\theta) \,| &\leq \| (v_{n+k} - v_n)(\cdot,\theta) \|_\alpha \\ &\leq \alpha^n \| (V_1 - V_0)(\cdot,\theta) \|_\alpha /(1-\alpha) \\ &\leq B\alpha^n. \end{aligned} \tag{6.10}$$

De (6.10) obtenemos que $\{v_n(\cdot,\theta)\}$ converge para todo $\theta \in T$. Haciendo tender $k \to \infty$ en (6.10) obtenemos (6.9). □

Observación (6.11). De (6.9) obtenemos que la función v pertenece al espacio $C(ST)$ por ser límite uniforme de funciones en $C(ST)$ y además que v satisface la siguiente desigualdad:

$$\sup_{(x,\theta)} | (v - v_n)(x,\theta) | \leq B\alpha^n, \qquad n \in \mathbf{N}. \tag{6.12}$$

Consideremos ahora las sucesiones

$$\overline{w}_n(\theta) := \sup_{x \in S}(V_n(x,\theta) - V_{n-1}(x,\theta)), \qquad \theta \in T,$$

y

$$\underline{w}_n(\theta) = \inf_{x \in S}(V_n(x,\theta) - V_{n-1}(x,\theta)), \qquad \theta \in T,$$

donde $\{V_n\}$ es la sucesión de iteración de valores de (4.2)

LEMA (6.13). *Supongamos que se cumplen las hipótesis* (2.1) *y* (3.1) *Entonces para todo* $\theta \in T$, $\{\overline{w}_n(\theta)\}$ *es no creciente y* $\{\underline{w}_n(\theta)\}$ *es no decreciente.*

Demostración. De la definición de $V_n(x,\theta)$ tenemos:

$$\begin{aligned} V_{n+1}(x,\theta) - V_n(x,\theta) &= \sup_{a \in A(x)} \{r(x,a,\theta) + \int_S V_n(y,\theta)p(dy \mid x,a,\theta)\} \\ &\quad - \sup_{a \in A(x)} \{r(x,a,\theta) + \int_S V_{n-1}(y,\theta)p(dy \mid x,a,\theta)\} \\ &\leq \sup_{a \in A(x)} \{\int_S (V_n - V_{n-1})(y,\theta)p(dy \mid x,a,\theta)\} \\ &\leq \sup_{x \in S}\{(V_n - V_{n-1})(x,\theta)\} \\ &= \overline{w}_n(\theta), \quad \theta \in T. \end{aligned}$$

De esto obtenemos que $\overline{w}_{n+1}(\theta) \leq \overline{w}_n(\theta)$, para $n \in \mathbf{N}$, $\theta \in T$. De manera análoga obtenemos que $\{\underline{w}_n(\theta)\}$ es no decreciente. □

Supongamos ahora que se cumple también la hipótesis (3.4). Podemos aplicar la proposición (6.5) a la sucesión

$$w_n(x,\theta) := V_n(x,\theta) - V_{n-1}(x,\theta), \qquad (x,\theta) \in ST$$

para obtener

$$\begin{aligned} \| w_n(x,\theta) \|_a &= \| (U^n V_1 - U^{n-1}V_0)(\cdot,\theta) \|_a \\ &\leq \alpha^{n-1} \| (V_1 - V_0)(\cdot,\theta) \|_a \end{aligned}$$

de donde

$$\| w_n(x,\theta) \|_a \leq C\alpha^{n-1}, \tag{6.14}$$

en donde $C = \sup_{\theta \in T} \| (V_1 - V_0)(\cdot,\theta) \| < \infty$.

De la desigualdad (6.14) obtenemos:

$$\lim_{n\to\infty} \| w_n(x,\theta) \|_a = 0. \tag{6.15}$$

Fijemos ahora el mismo estado $x^* \in S$ que en (6.7) y definamos la sucesión

$$g_n(\theta) := V_n(x^*,\theta) - V_{n-1}(x^*,\theta), \qquad \theta \in T,\ n \geq 1. \tag{6.16}$$

Del lema (6.13) y de (6.15) obtenemos que existe el límite

$$g(\theta) := \lim_{n\to\infty} g_n(\theta), \qquad \theta \in T \tag{6.17}$$

Además, la sucesión $\{g_n(\theta)\}$ es uniformemente acotada y

$$| g_n(\theta) - g(\theta) | \leq C\alpha^{n-1}, \qquad \theta \in T \text{ y } n \geq 1. \tag{6.18}$$

De (6.18) obtenemos que $g \in C(T)$, pues cada $g_n \in C(T)$.
Podemos ahora resumir los resultados anteriores en la siguiente proposición.

PROPOSICION (6.19) *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Con el estado* $x^* \in S$ *fijo, sea* $\{g_n(\theta)\}$ *la sucesión definida en* (6.16). *Entonces existe el límite* $g(\theta) := \lim_{n\to\infty} g_n(\theta)$, $\theta \in T$. *Se tiene además que* $g \in C(T)$ *y se cumple la desigualdad* (6.18).

Demostraremos ahora que la pareja $\{g, v\}$ satisface la ecuación de optimalidad. Esto demostrará la proposición (3.6).

PROPOSICIÓN (6.20). *Supongamos que se cumplen las hipótesis* (2.1), (3.1) *y* (3.4). *Entonces la función* $g \in C(T)$ *dada en* (6.17) *y la función* $v \in C(ST)$ *definida en la proposición* (6.8) *satisfacen la ecuación de optimalidad*:

$$g(\theta) + v(x,\theta) = Uv(x,\theta), \qquad (x,\theta) \in ST. \tag{6.21}$$

Demostración. (cf. Rieder [12], Teorema 4.3) Sea $\{v_n\}$ la sucesión definida en (6.7). Para cualquier $x \in S$ y $\theta \in T$, la sucesión de funciones continuas definidas por

$$a \to r(x,a,\theta) + \int_S v_n(y,\theta)p(dy \mid x,a,\theta),$$

converge uniformemente sobre el conjunto compacto $A(x)$ a la función

$$r(x,a,\theta) + \int_S v(y,\theta)p(dy \mid x,a,\theta).$$

Además tenemos

$$\begin{aligned} Uv_n(x,\theta) &= UV_n(x,\theta) - V_n(x^*,\theta) \\ &= V_{n+1}(x,\theta) - V_n(x^*,\theta) \\ &= V_{n+1}(x,\theta) - V_{n+1}(x^*,\theta) + V_{n+1}(x^*,\theta) - V_n(x^*,\theta) \\ &= v_{n+1}(x,\theta) + g_{n+1}(\theta), \end{aligned}$$

es decir, se cumple la ecuación

$$v_{n+1}(x,\theta) + g_{n+1}(\theta) = \sup_{a\in A(x)} \{r(x,a,\theta) + \int_S v_n(y,\theta)p(dy \mid x,a,\theta)\}$$

Usando ahora los resultados (6.12) y (6.18), y pasando al límite en la ecuación anterior, obtenemos que g y v satisfacen la ecuación de optimalidad (6.21). □

Comentario (6.22). La ecuación de optimalidad escrita en la forma (6.21) significa que la función v es un punto fijo del operador U en la seminorma de apertura. La idea de usar una condición de ergodicidad para obtener puntos fijos de U con dicha seminorma ha sido también usada en [12] y [14], entre otros.

En el caso en que se cumple la hipótesis de ergodicidad (3.4)'', resulta que el operador U' definido en (5.10) es un operador de contracción en la norma uniforme ver ([5]), y en este caso podemos obtener una función v' como el punto fijo de U'. Podemos aproximar uniformemente a v' usando la sucesión (5.12):

$$\begin{aligned} w'_{n+1}(x,\theta) &= U'w'_n(x,\theta) \\ &= \max_{a\in A(x)} \{r(x,a,\theta) + \int_S w'_n(y,\theta)p(dy \mid x,a,\theta)\} - g'_n(\theta) \end{aligned}$$

donde

$$g'_n(\theta) = \int_S w'_n(y,\theta)\mu(dy \mid \theta), \qquad \theta \in T, \tag{6.23}$$

converge uniformemente a $g'(\theta)$ dada ahora por

$$g'(\theta) = \int_S v(y,\theta)\mu(dy \mid \theta).$$

Se tiene que $g' \in C(T)$, $v' \in C(ST)$ y la pareja $\{g', v'\}$ satisface la ecuación de optimalidad.

Una observación análoga se tiene también en el caso en que se cumpla la hipótesis de ergodicidad (3.4)'''.

(6.24) *Demostración del Teorema* (5.5).

Sea θ^* el valor verdadero del parámetro desconocido y consideremos la función ϕ definida en la proposición (3.2) por

$$\phi(x,a,\theta) := r(x,a,\theta) + \int_S v(y,\theta)p(dy \mid x,a,\theta) - v(x,\theta) - g(\theta),$$

para todo $(x,a) \in K$ y $\theta \in T$. Por las hipótesis (2.1), (3.1) y por los resultados en la proposición (6.20), se tiene que la función ϕ es acotada. Consideremos ahora la política $\hat{d}$ definida en (5.4). Demostraremos que

$$\lim_{n\to\infty} \mid \phi(x_n, f_n(x_n,\hat{\theta}_n),\theta^*) \mid = 0, \quad P_x^{\hat{d},\theta^*}\text{-c.s.} \tag{6.25}$$

Entonces por la proposición (3.2)(iii), (6.25) implicará que $\hat{d}$ es óptima.

Sea $\{\theta_n\}$ cualquier sucesión en T convergente a θ^*. Consideremos las sucesiones $\{v_n\}$ y $\{g_n\}$ definidas respectivamente en (6.7) y (6.16). Usamos ahora la sucesión $\{f_n\}$ definida en la ecuación de iteración de valores (5.3) para obtener:

$$v_n(x,\theta_n) = r(x, f_n(x,\theta_n),\theta_n) + \int v_{n-1}(y,\theta_n)p(dy \mid x, f_n(x,\theta_n),\theta_n) - g_n(\theta_n).$$

Escribimos ahora ϕ en la forma

$$\phi(x, f_n(x,\theta_n),\theta^*) = \phi(x, f_n(x,\theta_n),\theta^*) - v_n(x,\theta_n) + v_n(x,\theta_n),$$

y sumamos y restamos el término $\int v(y,\theta^*)p(dy \mid x, f_n(x,\theta_n),\theta_n)$ en la igualdad anterior para obtener:

$$\begin{aligned}\phi(x, f_n(x,\theta_n),\theta^*) &= r(x, f_n(x,\theta_n),\theta^*) - r(x, f_n(x,\theta_n),\theta_n) \\ &+ \int v(y,\theta^*)[p(dy \mid x, f_n(x,\theta_n),\theta^*) - p(dy \mid x, f_n(x,\theta_n),\theta_n)] \\ &+ \int [v(y,\theta^*) - v_{n-1}(y,\theta_n)]p(dy \mid x, f_n(x,\theta_n),\theta_n) \\ &+ [g_n(\theta_n) - g(\theta^*)] + v_n(x,\theta_n) - v(x,\theta^*)\end{aligned} \tag{6.26}$$

De aquí que para todo $x \in S$,

$$
\begin{aligned}
(6.27) \qquad | \phi(x, f_n(x,\theta_n),\theta^*) | \leq & \sup_{(x,a)\in K} | r(x,a,\theta^*) - r(x,a,\theta_n) | \\
& + \sup_x | v(x,\theta^*) | \sup_{(x,a)\in K} \| p(\cdot \mid x,a,\theta^*) - p(\cdot \mid x,a,\theta_n) \|_v \\
& + \sup_x | v(x,\theta^*) - v_{n-1}(x,\theta_n) | \\
& + | g_n(\theta_n) - g(\theta^*) | + \sup_x | v_n(x,\theta_n) - v(x,\theta^*) | .
\end{aligned}
$$

Los dos primeros términos a la derecha de (6.27) tienden a cero, cuando $n \to \infty$ por la hipótesis (3.8) y por ser v acotada.

Por otro lado tenemos que para todo $x \in S$,

$$
\begin{aligned}
& | v(x,\theta^*) - v_{n-1}(x,\theta_n) | \\
\leq & | v(x,\theta^*) - v(x,\theta_n) | + | v(x,\theta_n) - v_{n-1}(v,\theta_n) | \\
\leq & \sup_x | v(x,\theta^*) - v(x,\theta_n) | + \sup_x | v(x,\theta_n) - v_{n-1}(x,\theta_n) | .
\end{aligned}
$$

Usando ahora la hipótesis (5.6) y la desigualdad (6.9) obtenemos que

$$
(6.28) \qquad \sup_x | v(x,\theta^*) - v_{n-1}(x,\theta_n) | \to 0, \text{ cuando } n \to \infty.
$$

Análogamente se cumple que

$$
(6.29) \qquad \sup_x | v_n(x,\theta_n) - v(x,\theta^*) | \to 0, \text{ cuando } n \to \infty.
$$

Por otro lado tenemos que

$$
| g_n(\theta_n) - g(\theta^*) | \leq | g_n(\theta_n) - g(\theta_n) | + | g(\theta_n) - g(\theta) |,
$$

de donde, por la desigualdad (6.18) y por la continuidad de g obtenemos que

$$
(6.30) \qquad | g_n(\theta_n) - g(\theta^*) | \to 0 \text{ cuando } n \to \infty.
$$

De lo obtenido en (6.28), (6.29) y (6.30), de (6.27) vemos que:

$$
\sup_x | \phi(x, f_n(x,\theta_n),\theta^*) | \to 0, \text{ cuando } n \to \infty.
$$

Por lo tanto, bajo la hipótesis (5.1) tenemos que para cualquier sucesión FC de estimadores $\{\hat{\theta}_n\}$ de θ^*, se cumple (6.24), y esto demuestra la optimalidad de $\hat{d}$.

Por otro lado, en (6.30) hemos demostrado que el valor óptimo se puede aproximar como en (5.8). Esto termina la demostración del teorema (5.5).

(6.31) *Demostración del Teorema* (5.9).

Procedemos como en la demostración del teorema (5.5) definiendo la función acotada ϕ y consideremos la igualdad (6.26); en esta igualdad, como S y A son compactos y r continua entonces se cumple la parte (a) de la hipótesis (3.8); es decir.

$$\lim_{n\to\infty} \sup_{(x,a)\in K} | r(x,a,\theta^*) - r(x,a,\theta_n) | = 0.$$

Por otro lado como $p(\cdot \mid x,a,\theta)$ es continua sobre KT, v es continua sobre ST y S y T con compactos, tenemos:

$$\lim_{n\to\infty} \sup_{(x,a)\in K} | \int_S v(y,\theta^*)[p(dy \mid x,a,\theta) - p(dy \mid x,a,\theta_n)] | = 0.$$

Verifiquemos ahora que la función v satisface la propiedad (5.6). Esta propiedad es consecuencia de que v es continua sobre ST y S es compacto.

Así, que de (6.26) se sigue también en este caso (6.25), y esto termina la demostración.

Observación (6.32). De manera análoga al teorema (5.5) podemos demostrar que la política d' definida en (5.13), y la política PEC de (5.15) son también óptimas.

DEPARTAMENTO DE MATEMÁTICAS
ESCUELA SUPERIOR DE FÍSICA Y MATEMÁTICAS
INSTITUTO POLITÉCNICO NACIONAL
MÉXICO, D.F. MÉXICO 07300

REFERENCIAS

[1] R.S. ACOSTA-ABREU y O. HERNÁNDEZ-LERMA, *Iterative adaptive control of denumerable state average-cost Markov systems*, Control and Cybernetics, **14** (1985), 313-322.

[2] D.P. BERTSEKAS and S.E. SHEREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[3] J.P. GEORGIN, *Estimation et controle des chaines de Markov sur des espaces arbitraires*, Lecture Notes Math. 636, Springer-Verlag, Berlin (1978), 71-113.

[4] ———, *Controle des chaines de Markov sur des espaces arbitraires*, Ann. Inst. Henri Poncaré, (B), Vol. XIV, No. 3, 1978, 255-277.

[5] L.G. GUBENKO and E.S. SHTATLAND, *On controlled, discrete-time Markov decision processes*, Theory Probab. Math. Statist., **7** (1975), 47-61.

[6] O. HERNÁNDEZ-LERMA, *Approximation and adaptive policies in discounted dynamic programming*, Bol. Soc. Mat. Mexicana, **30** (1985), 25-35.

[7] C.J. HIMMELBERG, T. PARTHASARATHY and F.S. VAN VLECK, *Optimal plans for dynamic programming problems*, Math. Oper, Res., **1** (1976), 390-394.

[8] K. HINDERER, *Foundations of non-stationary dynamic programming with discrete time parameter*, Lecture Notes in Oper. Res. and Math. Syst., **33**, Springer-Verlag, New York, 1970.

[9] M. KURANO, *Discrete-time markovian decision processes with an unknown parameter: average return criterion*, J. Oper. Res. Soc. Japan, **15** (1972), 65-76.

[10] K. KURATOWSKY, *Topology*, Vol. II, Academic Press, New York, 1968.

[11] P. MANDL, *Estimation and Control of Markov chains*, Adv. in Appl. Probab., **6** (1974), 40-60.

[12] U. RIEDER, *On non-discounted dynamic programming with arbitrary state space*, University of Ulm, RFA, 1979.

[13] H. L. ROYDEN, *Real Analysis, Macmillan*, New York, 1968.

[14] H. C. TIJMS, *On dynamics programming with arbitrary state space, compact action space and the average return as criterion*. Report BW 55/75 (1975), Mathematisch Centrum, Amsterdam.