Boletín de la Sociedad Matemática Mexicana Vol. 32 No. 2, 1987

CONTROL DE PROCESOS DE MARKOV PARCIALMENTE OBSERVABLES Y CON PARAMETROS DESCONOCIDOS: CRITERIO DE GANACIA PROMEDIO[1)]

Por ROBERTO S. ACOSTA ABREU

Resumen

En este trabajo consideramos procesos de decisión markovianos parcialmente observables, con parámetros desconocidos y con espacio de estados polaco. Usamos el criterio de ganancia promedio. Siguiendo un procedimiento usual, transformamos el proceso a uno nuevo con información completa. Suponemos en el proceso original condiciones de continuidad y de ergodicidad adecuadas y vemos cómo se preservan éstas en el proceso transformado. Usamos luego resultados recientes sobre procesos de decisión de Markov con parámetros desconocidos, para obtener políticas adaptables óptimas y esquemas de aproximación del valor óptimo.

1. Introducción

Los procesos de decisión markovianos (PDM) parcialmente observables (PO) o con información incompleta, han sido investigados ampliamente: véase por ejemplo el artículo [17] en el que se describe la literatura sobre este tema.

Los PDM con parámetros desconocidos también han recibido considerable atención, [13]. En este trabajo consideramos PDM parcialmente observables que dependen de parámetros desconocidos. Tratamos el caso en el que el espacio de estados es un espacio polaco (i.e., métrico separable y completo) y usamos el criterio de ganacia promedio. Primero seguimos el enfoque usual de transformar el PDM-PO en un PDM equivalente (en el sentido de que sus valores óptimos son iguales) pero con información completa. Luego combinamos esto con los resultados de [1] para obtener: (a) políticas adaptables óptimas en ganacia promedio y (b) esquemas de aproximación del valor óptimo. En [8] se consideran PDM-PO con parámetros desconocidos, espacio de estados numerable y con el criterio de costo descontado, y en [9] sistemas en tiempo discreto con distribución desconocida.

En la sección 2 del presente trabajo, definimos los PDM-PO sin parámetros desconocidos. En la sección 3 describimos brevemente siguiendo [5,19], la transformación de un PDM-PO a un PDM con información completa. En la sección 4, introducimos en el sistema PO condiciones de continuidad y de ergodicidad y vemos como el sistema transformado hereda estas propiedades. Luego usamos resultados de [7] para obtener políticas estacionarias óptimas. En la sección 5 se consideran los PDM-PO con parámetros desconocidos, se definen algunas políticas adaptables óptimas y establecemos los resultados

1) Este trabajo se realizó con apoyo de la COFAA del I.P.N.

principales de este trabajo. En la sección 6 se considera un ejemplo de un sistema de inventario PO. La última sección es la 7 en la que resumimos nuestras conclusiones.

Terminamos esta seción introduciendo 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. El producto cartesiano de X e Y lo denotaremos por XY. 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 $C(X)$. Consideramos en todo lo que sigue a X e Y como espacios polacos. 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, y $\mathbf{P}(X)$ es un espacio compacto si y sólo si X es compacto [4,5,19]. Denotaremos por $\mathbf{P}(Y \mid X)$ 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 $\int_Y g(y)p(dy \mid x)$ es una función continua sobre X. La norma $\| \cdot \|_v$ denotará la norma de variación total de las medidas con signo involucradas.

2. PDM-PO: El caso de parámetros conocidos

En esta sección definimos un PDM-PO en el caso en que no hay parámetros desconocidos.

Un PDM-PO está definido por

$$(X, Y, A, p, p_0, q, r) \tag{2.1}$$

donde:

A1: Hipótesis

(a) X es el espacio de estados, que supondremos un espacio polaco.

(b) Y es el conjunto de señales (o mediciones) de observación, espacio polaco.

(c) A es el conjunto de controles (o acciones) que será un espacio métrico compacto; para cada $y \in Y$, denotamos por $A(y)$ al conjunto de controles admisibles cuando la observación es y. Supondremos que $A(y)$ es un subconjunto cerrado de A (y por tanto compacto), no vacío, para cada y en Y; supondremos que el conjunto

$$K := \{(y, a) : y \in Y,\ a \in A(y)\}$$

es *cerrado* en YA; esto significa que la correspondencia $y \to A(y)$ de Y en 2^A (donde 2^A son todos los subconjuntos cerrados no vacios de A) es *semicontinua superiormente*; supondremos también que $y \to A(y)$ es *semicontinua inferiormente*; estas condiciones implican que $y \to A(y)$ es *continua*; (ver [3], pg. 117, [10], pg. 113, [14] pg. 58).

(d) La probabilidad condicional $p = p(dx' \mid x, a) \in \mathbf{P}(X \mid XA)$ es la probabilidad de transición de los estados del sistema.

(e) $p_0 \in \mathbf{P}(X)$ es la distribución de probabilidad del estado inicial del sistema.

(f) $q = q(dy \mid x, a) \in \mathbf{P}(Y \mid XA)$ es la llamada característica del sistema de observaciones.

(g) $r = r(x, y, a) \in B(XYA)$ es la función de ganacia.

El proceso evoluciona en el tiempo de la siguiente manera. En la etapa n, $n = 0, 1, 2, \ldots$, sean x_n , y_n y a_n el estado del proceso, la observación hecha y la acción tomada. En el instante $n = 0$, la distribución (inicial) de x_0 es p_0 y la observación inicial es y_0. Si en el instante $n (n = 0, 1, \ldots)$ el estado del sistema es $x_n = x$, se observa $y_n = y$ y se escoge el control $a_n = a$, entonces se recibe una ganacia (esperada) $r(x, y, a)$ y el sistema se mueve a un nuevo estado x_{n+1} de acuerdo con la ley de transición $p(dx' \mid x, a)$. El estado al tiempo $n + 1$, digamos $x_{n+1} = x'$ no se puede observar directamente; en lugar de esto se obtiene una observación o medición y_{n+1} generada de acuerdo con el núcleo $q(dy \mid x', a)$. Después que ocurre la transición a x', se escoge otra vez un control y se repite el proceso.

Para $n \geq 0$, sea $H_n = (YA)^n Y$ el conjunto de todas las historias observables h_n

$$h_n = (y_0, a_0, \ldots, y_{n-1}, a_{n-1}, y_n) \tag{2.2}$$

con $H_0 = Y$. Una política (ó estrategia) para el PDM-PO es una sucesión $\delta = \{\delta_n\}_{n \geq 0}$ donde $\delta_n \in \mathbf{P}(A \mid H_n)$, δ_n concentrada en $A(y_n)$ y el control al tiempo n se escoge de acuerdo a $\delta_n(\cdot \mid h_n)$. Sea D el conjunto de todas las políticas. Una política se llama markoviana si es de la forma $\delta_n(\cdot \mid h_n) := \delta_n(\cdot \mid y_n)$, $n \geq 0$. Una polítita markoviana δ se llama estacionaria si $\delta_n(\cdot \mid y_n) = \delta(\cdot \mid y_n)$, $n \geq 0$; una política estacionaria se llama determinista si la medida $\delta(\cdot \mid y)$ es degenerada para cada $y \in Y$. Una política determinista se puede identificar con una función Borel medible $f : Y \to A$ tal que $f(y) \in A(y)$ para cada $y \in Y$. Denotaremos por F al conjunto de todas las políticas deterministas.

Cada política $\delta \in D$ y cada distribución inicial p_0 junto con las probabilidades condicionales p y q determinan una medida de probabilidad P_0^δ sobre el espacio de todas las realizaciones posibles $\Omega := (XYA)^\infty$ del PDM-PO; ver [4,19,20,21]. Denotaremos por E_0^δ a la esperanza con respecto a esta medida de probabilidad.

Definimos la ganacia esperada por unidad de tiempo cuando se usa la política δ con la observación inicial y_0 y la distribución inicial p_0 por

$$J(\delta, y_0, p_0) = \liminf_{n\to\infty} (n+1)^{-1} E_0^\delta[\sum_{k=0}^{n} r(x_k, y_k, a_k)] \tag{2.3}$$

Brevemente, a (2.3) le llamaremos el criterio J, y a una política δ^* para la cual se cumple que

$$J(\delta^*, y_0, p_0) = \sup_{\delta\in D} J(\delta, y_0, p_0), \quad y_0 \in Y, \quad p_0 \in \mathbf{P}(X) \tag{2.4}$$

la llamaremos óptima en ganacia promedio.

3. Reducción a un PDM con información completa

En esta sección reduciremos el PDM-PO a un nuevo proceso de decisión de Markov (PDM) con información completa, al cual lo denotaremos por

$$(V, A, \overline{P}, \bar{r}) \tag{3.1}$$

Para describir el nuevo PDM seguiremos brevemente un procedimiento que ya es conocido ver, por ejemplo, [5,19]. En (3.1), el nuevo espacio de estados es $V := Y\mathbf{Z}$, donde $\mathbf{Z} := \mathbf{P}(X)$. El conjunto de acciones A es el mismo que para el PDM-PO y como conjuntos de control admisibles en el estado $v = (y, z)$ tomamos a $A(v) \equiv A(y)$ para $y \in Y$. La probabilidad de transición $\overline{P}$ la definimos de la siguiente manera:
Definimos primero una probabilidad condicional $\tilde{p} \in \mathbf{P}(YX \mid \mathbf{Z}A)$, la cual para cada $(z_n, a_n) \in \mathbf{Z}\,A$ está dada por:

$$\tilde{p}(dy_{n+1}dx_{n+1} \mid z_n, a_n) := \int xq(dy_{n+1} \mid x_{n+1}, a_n)p(dx_{n+1} \mid x_n, a_n)z_n(dx_n) \tag{3.2}$$

Usando el teorema de factorización dado en [4], Corolario 7.27.1 ó en [19,22] se obtiene que existe una probabilidad condicional $\bar{p} \in \mathbf{P}(X \mid \mathbf{Z}AY)$ tal que

$$\tilde{p}(dy_{n+1}dx_{n+1} \mid z_n, a_n) = R(dy_{n+1} \mid z_n, a_n)\bar{p}(dx_{n+1} \mid z_n, a_n, y_{n+1}) \tag{3.3}$$

donde

$$(3.4)\quad R(dy_{n+1} \mid z_n, a_n) = \int_X \int_X q(dy_{n+1} \mid x_{n+1}, a_n) p(dx_{n+1} \mid x_n, a_n) z_n(dx_n),$$

es la marginal de $\tilde{p}(dy_{n+1}dx_{n+1} \mid z_n, a_n)$ sobre Y.
Usando la función medible $s : \mathbf{Z}\,AY \to \mathbf{Z}$ dada por $s(z_n, a_n, y_{n+1}) := \overline{p}(dx_{n+1} \mid z_n, a_n, y_{n+1})$, definimos ahora la probabilidad de transición $t \in \mathbf{P}(\mathbf{Z} \mid \mathbf{Z}AY)$ por $t(B \mid z_n, a_n, y_{n+1}) := I_B[s(z_n, a_n, y_{n+1})]$, para cualquier $B \in \mathcal{B}(\mathbf{Z})$, donde $I_B(\omega) = 1$ si $\omega \in B$, e $I_B(\omega) = 0$, si $\omega \notin B$, es la función indicadora de B.
Finalmente, definimos la probabilidad de transición $\overline{P} \in \mathbf{P}(YZ \mid YZA)$ por

$$(3.5)\quad \overline{P}(dy_{n+1}dz_{n+1} \mid y_n, z_n, a_n) = R(dy_{n+1} \mid z_n, a_n) t(dz_{n+1} \mid z_n, a_n, y_{n+1})$$

De (3.5) vemos que para $A \in \mathcal{B}(Y)$ y $B \in \mathcal{B}(\mathbf{Z})$ arbitrarios:

$$(3.6)\quad \overline{P}(AB \mid y_n, z_n, a_n) = \int_A I_B[s(z_n, a_n, y_{n+1})] R(dy_{n+1} \mid z_n, a_n)$$

En el nuevo modelo definimos la función de ganacia esperada $\overline{r} : VA \to \mathbf{R}$ por

$$(3.7)\quad \overline{r}(v, a) = \overline{r}(y, z, a) := \int_X r(x, y, a) z(dx)$$

para $(v, a) \in \overline{K}$, donde

$$(3.8)\quad \overline{K} = \{(v, a) \mid v \in V,\ a \in A(v)\}$$

El conjunto $\overline{K}$ es cerrado en VA por ser K un conjunto cerrado en YA.
Definamos ahora el nuevo conjunto de políticas $\overline{D}$. Para esto consideramos la sucesión $\{z_n\}_{n \geq 0}$ con valores en $\mathbf{Z}$ definida, usando la función s, por

$$(3.9)\quad \left.\begin{aligned} z_0 &= p_0 \\ z_{n+1} &= s(z_n, a_n, y_{n+1}),\ n \geq 0. \end{aligned}\right\}$$

Usando (3.9) para cada historia observable $h_n \in H_n$, podemos determinar un vector de información

$$(3.10)\quad \begin{aligned} \overline{h}_n &= (y_0, z_0, a_0, \ldots, z_{n-1}, a_{n-1}, y_n, z_n) \\ &\equiv (v_0, a_0, v_1, \ldots, v_{n-1}, a_{n-1}, v_n) \in \overline{H}_n \end{aligned}$$

donde $\overline{H}_n := V(AV)^n$, $n = 0,1,\ldots$; $H_0 := V$. Definimos una I-política (ó política de información) como una sucesión $\bar{\delta} = \{\bar{\delta}_n\}$, donde $\bar{\delta} \in \mathbf{P}(A \mid \overline{H}_n)$. Sea $\overline{D}$ el conjunto de todas las I-políticas. Se puede considerar a $\overline{D}$ como un subconjunto de D, porque para cualquier I-política $\bar{\delta} \in \overline{D}$ podemos definir una política $\delta = \{\delta_n\} \in D$ por medio de $\delta_n(h_n) := \bar{\delta}_n(\bar{h}_n)$, donde para cada $h_n \in H_n$ el vector de información $\bar{h}_n \in \overline{H}_n$ está determinado por (3.10). Además $\overline{D}$ es completo, en el sentido de que para cualquier política $\delta \in D$, existe una I-política $\bar{\delta} \in \overline{D}$ tal que

$$J(\delta, y_0, p_0) = J(\bar{\delta}, y_0, p_0), \quad y_0 \in Y, \quad p_0 \in \mathbf{Z};$$

véase [5,19,21,22].

La probabilidad de trancisión $\overline{P}$ dada en (3.5) junto con una I-política $\bar{\delta}$ y una distribución a priori p_0 para el estado inicial, determinan una medida de probabilidad $\overline{P}_0^{\bar{\delta}}$ sobre el espacio de todas las sucesiones de información posibles $(v_0, a_0, v_1, a_1, \ldots)$. Denotemos por $\overline{E}_0^{\bar{\delta}}$ a la esperanza con respecto a esta probabilidad. Entonces la ganacia esperada por unidad de tiempo para el nuevo sistema es

$$(3.11) \qquad \overline{J}(\bar{\delta}, v_0) := \liminf_{n\to\infty} (n+1)^{-1} E_0^{\bar{\delta}}[\sum_{k=0}^{n} \bar{r}(v_k, a_k)], \bar{\delta} \in d, \quad v_0 \in V.$$

El proceso de decisión de Márkov dado en (3.1) es un proceso de decisión de Markov con información completa. Aquí las políticas son las I-políticas $\bar{\delta} \in \overline{D}$ y consideramos el criterio $\overline{J}$ dado en (3.11). El PDM (3.1) con el criterio (3.11) es equivalente al proceso original dado en (2.1) en el sentido de que para cualquier I-política $\bar{\delta}$, con $v_0 = (p_0, y_0)$, $y_0 \in Y$, $p_0 \in \mathbf{Z}$,

$$(3.12) \qquad \overline{J}(\bar{\delta}, v_0) = J(\bar{\delta}, y_0, p_0);$$

véase [5,19,21,22]

Así, del hecho de que $\overline{D}$ es completo, podemos en particular concluir que una I-política es óptima con el criterio $\overline{J}$ para el sistema con información completa (3.1), si y sólo si, es óptima con el criterio J para el sistema PDM-PO de (2.1). En otras palabras, los resultados para el sistema (3.1) se pueden traducir en resultados para el PDM-PO, reemplazando políticas por I-políticas. Ejemplos de esto se pueden ver en [2,19,20,22]. En la sección siguiente consideraremos condiciones sobre el PDM-PO bajo las cuales la función de ganacia óptima

$$(3.13) \qquad \overline{J}^*(v_0) = \sup_{\bar{\delta}\in\overline{D}} \overline{J}(\bar{\delta}, v_0), \quad v_0 \in V$$

satisface la ecuación de optimalidad dada en (4.9).

Nota 3.1. Cuando el conjunto Y de observaciones es numerable, la probabilidad condicional $\bar{p}(dx_{n+1} \mid z_n, a_n, y_{n+1})$ obtenida en (3.3), está dada por

$$(3.14)\quad \bar{p}(A|z_n,a_n,y_{n+1})=R(y_{n+1}|z_n,a_n)^{-1}\int_X\int_A q(y_{n+1}|x_{n+1},a_n)p(dx_{n+1}|x_n,a_n)z_n(dx_n),$$

para cualquier $A \in \mathcal{B}(X)$, cuando $R(y_{n+1} \mid z_n, a_n) := R(\{y_{n+1}\} \mid z_n, a_n) \neq 0$. Si $R(y_{n+1} \mid z_n, a_n) = 0$, definimos $\bar{p}(A \mid z_n, a_n, y_{n+1})$ como una medida de probabilidad arbitraria sobre X.

4. Condiciones de optimalidad para el PDM con información completa

Una vez que se ha reducido el PDM-PO a un PDM con información completa hay que suponer algunas hipótesis sobre el PDM-PO, tales como continuidad de r, continuidad débil de las probabilidades de transición p y q y alguna condición de ergodicidad. Luego veremos que el modelo reducido hereda propiedades parecidas a éstas. Primero tenemos el resultado siguiente.

Consideremos sobre el PDM-PO dado en (2.1) las siguientes hipótesis.

A2: Hipótesis.

(a) El espacio de observaciones Y es *numerable* (consideramos en Y la topología discreta).

(b) La función $r \in C(XYA)$.

(c) La probabilidad de transición $p(dx' \mid x, a) \in \mathbf{P}(X \mid XA)$ es continua.

(d) La probabilidad de transición $g(dy \mid x, a) \in \mathbf{P}(Y \mid XA)$ es continua.

Nota (4.1). En la proposición siguiente necesitaremos usar el siguiente hecho.
Sean W y S espacios polacos arbitrarios y sean $h \in C(WS)$ y $t(ds \mid w) \in \mathbf{P}(S \mid W)$ continua. Entonces la función

$$\lambda(w) = \int_S h(s,w)t(ds \mid w), \quad w \in W$$

es continua. La demostración se puede ver p. ej. en la Proposición 7.30 de [4].

PROPOSICIÓN (4.1). *Bajo la hipótesis* A2 , *en el modelo* PDM (3.1) *tenemos:*
(i) *La función* $\bar{r}(v,a)$ *es continua y acotada para cada* $(v,a) \in \overline{K}$.
(ii) *Para cada* $y' \in Y$, *la probabilidad de transición* $\bar{p}(dx' \mid z, a, y')$ *dada en* (3.14) *es continua para cada* $(z,a) \in \mathbf{Z}A$.
(iii) *La probabilidad de transición* $\overline{P}(dv' \mid v, a) \in \mathbf{P}(V \mid VA)$ *es continua para cada* $(v,a) \in \overline{K}$.

Demostración. (i) Usando el resultado dado en la nota (4.1) con $W = VA = YZA$, $\mathbf{Z} = \mathbf{P}(X)$, $S = X$, con la probabilidad de transición continua $t(dx \mid y,z,a) \equiv z(dx)$ y con la función continua y acotada $h(x,y,z,a) \equiv r(x,y,a)$

obtenemos que $\bar{r}(y,z,a) = \bar{r}(v,a) = \int r(x,y,a)z(dx)$ es continua para cada $(y,z,a) \in \overline{K}$.

(ii) Usando de nuevo el resultado de la nota (4.1) y la hipótesis A2, podemos ver que la probabilidad de transición $\tilde{p}(y'dx' \mid z,a)$ definida como en (3.2) es continua en cada $(z,a) \in \mathbf{Z}A$ (ver pg. 157 de [22]). De aquí se sigue que la probabilidad marginal $R(y' \mid z,a)$ de $\tilde{p}(y'dx' \mid z,a)$ es continua para cada $(v,a) \in \mathbf{Z}A$. Como Y es numerable podemos usar (3.14) para escribir para $y' \in Y$,

$$\bar{p}(dx' \mid z,a,y') = R(y' \mid z,a)^{-1}\,\tilde{p}(y'dx' \mid z,a),$$

y de aquí vemos que $\bar{p}(dx' \mid z,a,y')$ es continua para cada $(z,a) \in \mathbf{Z}A$.

(iii) Para demostrar (iii) hay que ver que para cada $f \in C(V)$, $V = Y\mathbf{Z}$, la función

$$h(v,a) = \int f(z)\overline{P}(dz \mid v,a)$$

es continua para cada $(v,a) \in \overline{K}$. De (3.5) obtenemos la fórmula

$$h(v,a) = \sum_{y' \in Y} f[s(z,a,y')]R(y' \mid z,a) \tag{4.1}$$

La continuidad de $h(v,a)$ se sigue entonces de que $R(y' \mid v,a)$ es continua en (z,a), $f \in C(V)$, $s(z,a,y') = \bar{p}(dx' \mid z,a,y')$ es continua en $(z,a) \in \mathbf{Z}A$ y del lema 1 en [22]. Esto termina la demostración.

La siguiente hipótesis en el PDM-PO nos dará sobre el PDM (3.1) una propiedad de *ergodicidad*, la cual nos será útil al considerar la ecuación de optimalidad (4.9).

A3: Hipótesis.
Existen un estado $x^* \in X$, un valor de la señal de observación $y^* \in Y$ y números $\alpha_1 > 0$ y $\gamma_1 > 0$ tales que

$$p(\{x^*\} \mid x,a) \geq \alpha_1 \quad \text{para todo } (x,a) \in XA, \tag{4.2}$$

y

$$q(y^* \mid x',a) = \begin{cases} \gamma_1, & \text{si } x' = x^* \\ 0, & \text{si } x' \neq x^* \end{cases} \tag{4.3}$$

para todo $(x',a) \in XA$.

La hipótesis hecha sobre q implica que y^* sólo se puede observar con probabilidad γ_1 cuando el sistema está en x^*. Si $\gamma_1 = 1$, entonces x^* es completamente observable.

Las hipótesis A2 y A3 implican que en el PDM dado en (3.1) se satisface la siguiente condición de ergodicidad.

PROPOSICIÓN (4.2) *Bajo las hipótesis* A2 *y* A3 *existe un estado* $v^* \in V$ *y un número* $\alpha > 0$ *tal que*

$$\overline{P}(\{v^*\} \mid v, a) \geq \alpha \quad \text{para todo } (v, a) \in \overline{K} \tag{4.4}$$

Demostración. Sea $z^* \in \mathbf{Z}$ la medida de probabilidad sobre X concentrada en x^*, donde x^* cumple (4.2), y sea $v^* = (y^*, z^*) \in V$, donde y^* cumple (4.3). Entonces por (3.4):

$$\begin{aligned} R(y^* \mid z, a) &= \int_X \int_X q(y^* \mid x', a) p(dx' \mid x, a) z(dx) \\ &= \int_X \gamma_1 p(\{x^*\} \mid x, a) z(dx) \geq \gamma_1 \alpha_1 =: \alpha. \end{aligned} \tag{4.5}$$

Además, de (3.14),

$$\overline{p}(\{x^*\} \mid z, a, y^*) = R(y^* \mid z, a)^{-1} \int_X \gamma_1 p(\{x^*\} \mid x, a) z(dx) = 1,$$

y

$$\overline{p}(C \mid z, a, y^*) = R(y^* \mid z, a)^{-1} \int_X \int_C q(y^* \mid x', a) p(dx' \mid x, a) z(dx) = 0,$$

si $x^* \notin C$, $C \in \mathcal{B}(X)$, es decir,

$$s(z, a, y^*) = \overline{p}(\cdot \mid z, a, y^*) = z^* \quad \text{para cualquier } (z, a) \in \mathbf{Z}A. \tag{4.6}$$

Finalmente de (3.6), usando (4.5) y (4.6) obtenemos que para cada $(y, z, a) = (v, a) \in \overline{K}$,

$$\begin{aligned} \overline{P}(\{v^*\} \mid v, a) &= \sum_{y' \in Y} I_{z^*}[s(z, a, y')] R(y' \mid z, a) \\ &\geq I_{z^*}[s(z, a, y^*)] \alpha = \alpha, \end{aligned}$$

lo cual termina la demostración.

Es nuestro propósito aplicar al PDM en (3.1) el resultado dado en el teorema 3 de [7]. Para esto necesitamos la proposición siguiente.

PROPOSICIÓN (4.3). *En el modelo* PDM (3.1), *la correspondencia* $v \to A(v) \equiv A(y)$ *de* V *en* 2^A *es continua.*

Demostración. Ya vimos que el conjunto $\overline{K}$ dado en (3.8) es cerrado; esto implica que la correspondencia $(y, z) \to A(y)$ es semicontinua superiormente [3,14]. La semicontinuidad inferior de dicha correspondencia se sigue de la hipótesis A1(c). Esto termina la demostración.

Con $v^* \in V$ definido como en la proposición (4.2), definimos ahora la probabilidad de transición $\overline{P}'(\cdot \mid v, a) \in \mathbf{P}(V \mid VA)$ para $(v, a) \in \overline{K}$ por

$$\overline{P}'(\cdot \mid v, a) := \overline{P}(\cdot \mid v, a) - \alpha\delta_{v^*}(\cdot) \tag{4.7}$$

donde $\delta_{v^*}(\cdot)$ es la medida de probabilidad concentrada en v^*.

Por la proposición (4.2) y el teorema 2 de [7], el operador $G : B(V) \to B(V)$ definido para $u \in B(V)$ por

$$(Gu)(v) := \sup_{a \in A(v)} \{\overline{r}(v, a) + \int u(v')\overline{P}'(dv' \mid v, a)\}, \quad v \in V \tag{4.8}$$

es un operador de contracción. Observe que el sup en el lado derecho de (4.8) se alcanza, pues por la proposición (4.1), $\overline{r}(v, a)$ y $\overline{P}'(dv' \mid v, a)$ son continuas en cada $(v, a) \in \overline{K}$.

Tenemos el siguiente resultado.

TEOREMA 4.1. *Supongamos que se cumplen las hipótesis* A1, A2 *y* A3. *Entonces existe una función* $h \in C(V)$ *y una constante* g *tales que*:

$$g + h(v) = \max_{a \in A(v)} \{\overline{r}(v, a) + \int h(v')\overline{P}(dv' \mid v, a)\}, \quad v \in V. \tag{4.9}$$

Además existe una I-política estacionaria f^* *donde* $f^* : V \to A$ *se puede escoger como cualquier fuunción medible que maximiza el lado derecho de* (4.9):

$$g + h(v) = \overline{r}(v, f^*(v)) + \int h(v')\overline{P}(dv' \mid v, f^*(v)\}, \quad v \in V; \tag{4.10}$$

también tenemos que:

$$\overline{J}(f^*, v_0) \equiv g = \sup_{\overline{\delta} \in \overline{D}} \overline{J}(\overline{\delta}, v_0) = \overline{J}^*(v_0), \quad v_0 \in V. \tag{4.11}$$

Demostración. Ya vimos que el operador G definido en (4.8) es un operador de contracción en $B(V)$. Por la proposición (4.3) y por los lemas 1 y 2 de [7], se tiene que $Gu \in C(V)$ para cada $u \in C(V)$. Por el teorema de punto fijo de Banach, existe $u^* \in C(V)$ tal que

$$u^*(v) = \max_{a \in A(v)} \{\bar{r}(v,a) + \int u^*(v')\overline{P}'(dv' \mid v,a)\}, \quad v \in V \tag{4.12}$$

Si definimos ahora la función $h \in C(V)$ y la constante g por

$$h(v) = u^*(v), \quad v \in V \quad \text{y} \quad g = \int u^*(v')\delta_{v^*}(dv') = u^*(v^*), \tag{4.13}$$

obtenemos que h y g satisfacen la ecuación (4.9). Podemos ahora aplicar las proposiciones 4.1, 4.2, 4.3 y el teorema 3 de [7], para obtener la existencia de la función medible f^* que satisface (4.10) y que g satisface (4.11). Esto termina la demostración.

Nota 4.1 La ecuación (4.9) se llama *ecuación de optimalidad* para el criterio $\overline{J}$.

5. PDM-PO con parámetros desconocidos

Consideremos ahora un PDM-PO de la forma

$$(X, Y, A, p(\theta), p_0(\theta), q(\theta), r(\theta)) \tag{5.1}$$

en el cual la ley de transición $p(\theta) = p(dx' \mid x, a, \theta)$, la distribución inicial $p_0(\theta)$, el núcleo de observaciones $q(\theta) = q(dy \mid x, a, \theta)$, y la función de ganancia $r(\theta) = r(x, y, a, \theta)$ dependen de un parámetro θ. Suponemos que no se da ninguna información a priori sobre θ excepto que pertenece a un conjunto de parámetros T, el cual se supone que es un espacio métrico compacto.

Para el PDM-PO(θ) se supone que se cumplen hipótesis análogas a A1-A3 en las que se toma en cuenta la dependencia sobre θ. Así, A1θ queda como sigue.

A1θ: Hipótesis.

(a) y (b) como en la hipótesis A1;

(c) El conjunto de controles A es un espacio métrico compacto; el conjunto de controles admisibles para la observación y es $A(y) \subseteq A$.

El conjunto

$$K' = \{(y, a, \theta) \mid y \in Y,\ a \in A(y),\ \theta \in T\}$$

es cerrado en YAT. La correspondencia $y \to A(y)$ de Y en 2^A es continua.

Las partes (**d**)–(**g**) se consideran como las análogas de A1 tomando en cuenta la dependencia sobre θ, es decir, por ejemplo $p(dx' \mid x, a, \theta)$ es una probabilidad de transición en $\mathbf{P}(X \mid XAT)$.

A2θ: Hipótesis.
(**a**) El espacio Y es numerable
(**b**) $r(x, y, a, \theta) \in C(XYAT)$;
(**c**) $p(dx' \mid x, a, \theta)$ y $q(dy \mid x, a, \theta)$ son continuas para todo $(x, a, \theta) \in XAT$.
A3θ: Hipótesis.
Existen un estado $x^* \in X$, una señal de observación $y^* \in Y$ y números positivos α_2 y γ_2 tales que $p(\{x^*\} \mid x, a, \theta) \geq \alpha_2$ para todo $(x, a, \theta) \in XAT$ y

$$q(y^* \mid x', a, \theta) = \begin{cases} \gamma_2, & \text{si } x' = x \\ 0, & \text{si } x' \neq x^* \end{cases}$$

para todo $(x', a, \theta) \in XAT$.

Para cada $\theta \in T$, se puede reducir el PDM-PO(θ) como en la sección 3, a un PDM(θ) con información completa

$$(V, A, \overline{P}(\theta), \overline{r}(\theta)) \tag{5.2}$$

donde $\overline{P}(\theta)$ está dada por

$$\overline{P}(BC \mid v, a, \theta) = \sum_{y' \in Y} I_C[s(z, a, \theta, y')] R(y' \mid z, a, \theta) \tag{5.3}$$

para $v = (y, z)$, $B \in \mathcal{B}(Y)$ y $C \in \mathcal{B}(\mathbf{Z})$, donde

$$R(y' \mid z, a, \theta) = \int_X \int_X q(y' \mid x', a, \theta) p(dx' \mid x, a, \theta) z(dx), \tag{5.4}$$

$$s(z, a, \theta, y') = R(y' \mid z, a, \theta)^{-1} \int_X q(y' \mid x', a, \theta) p(dx' \mid x, a, \theta) z(dx)$$

y

$$\overline{r}(v, a, \theta) = \int_X r(x, y, a, \theta) z(dx) \tag{5.5}$$

Sea

$$\overline{K}' = \{(v, a, \theta) \mid v \in V,\ a \in A(v) \text{ y } \theta \in T\} \tag{5.6}$$

El conjunto $\overline{K}'$ es un conjunto cerrado en VAT por ser K' un conjunto cerrado en YAT. Bajo las hipótesis A1θ–A3θ se cumplen las proposiciones análogas a las proposiciones (4.1)–(4.3):

PROPOSICION (5.1). *Supongamos que se cumplen las hipótesis* A1θ–A3θ. *Entonces*
(i) $\overline{r}(v,a,\theta) \in C(\overline{K}')$;
(ii) $\overline{P}(dv' \mid v,a,\theta)$ *es continua para cada* $(v,a,\theta) \in \overline{K}'$;
(iii) *Existen un estado* $v^* \in V$ *y un número* $\overline{\alpha} > 0$ *tales que* $\overline{P}(\{v^*\} \mid v,a,\theta) \geq \overline{\alpha}$, *para todo* $(v,a,\theta) \in \overline{K}'$;
(iv) *La correspondencia* $(v,\theta) \to A(v)$ *de* VT *en* 2^A *es continua.*

En la parte **(iii)** de la proposición 5.1 tenemos que en el sistema (3.1) se cumple una condición de ergodicidad.

Por la parte **(iv)** de la proposición 5.1 y usando los lemas 1 y 2 de [7], obtenemos que para cualquier función $w \in C(\overline{K}')$, la función

$$u^*(v,\theta) = \max_{a \in A(v)} w(v,a,\theta) \tag{5.7}$$

es continua y acotada sobre VT. De esto, usando los resultados dados en la proposición 5.1, obtenemos como en (4.8) que el operador $\overline{G} : B(VT) \to B(VT)$ definido para $u \in B(VT)$ por

$$(\overline{G}u)(v,\theta) = \max_{a \in A(v)} \{\overline{r}(v,a,\theta) + \int u(v',\theta)\overline{P}'(dv' \mid v,a,\theta)\}, \tag{5.8}$$

donde

$$\overline{P}'(\cdot \mid v,a,\theta) = \overline{P}(\cdot \mid v,a,\theta) - \overline{\alpha}\delta_{v^*}(\cdot), \tag{5.9}$$

es un operador de contracción sobre el subespacio $C(VT)$ de $B(VT)$.
Si $u^*(v,\theta) = h(v,\theta)$ en $C(VT)$ es el punto fijo de $\overline{G}$ y

$$g(\theta) = \int u^*(v',\theta)\delta_{v^*}(dv') = u^*(v^*,\theta),$$

entonces, $g \in C(T)$. De aquí resulta que las funciones g y h satisfacen una ecuación análoga a la (4.9). Podemos seguir ahora como en la demostración del teorema 4.1 para obtener el resultado siguiente.

PROPOSICIÓN (5.2). *Supongamos que se cumplen las hipótesis* A1θ–A3θ. *Entonces*:

(a) *Existen funciones* $g \in C(T)$ *y* $h \in C(VT)$ *tales que*

$$g(\theta) + h(v,\theta) = \max_{a\in A(v)}\{\bar{r}(v,a,\theta) + \int h(v',\theta)\overline{P}(dv' \mid v,a,\theta)\}. \tag{5.10}$$

(b) *Existe una I-política estacionaria* $f_\theta^* : V \to A$, *medible en* (v,θ) *que satisface*

$$g(\theta) + h(v,\theta) = \bar{r}(v, f_\theta^*(v),\theta) + \int h(v',\theta)\overline{P}(dv' \mid v, f_\theta^*(v),\theta)\} \tag{5.11}$$

Además $g(\theta)$ cumple que

$$g(\theta) \equiv \overline{J}_\theta^*(v_0) = \sup_{\bar{\delta}\in\overline{D}} \overline{J}_\theta(\bar{\delta}, v_0), \quad v_0 \in V \tag{5.12}$$

donde

$$\overline{J}_\theta(\bar{\delta}, v_0) = \liminf_{n\to\infty}(n+1)^{-1}\overline{E}_0^{\bar{\delta},\theta}[\sum_{k=0}^{n}\bar{r}(v_k,a_k,\theta)], \quad \bar{\delta}\in\overline{D}, \quad v_0 \in V \tag{5.13}$$

La ecuación (5.10) se llama ecuación de optimalidad para el sistema (5.2).

En el PDM-PO(θ) haremos uso de las siguientes hipótesis de "continuidad" sobre el parámetro:

A4θ: hipótesis.

Para cualquier $\theta \in T$ y cualquier sucesión $\{\theta_n\}$ en T convergente a θ cuando $n \to \infty$, tenemos:

(a) $R(n,\theta) := \sup_{x,y,a} \mid r(x,y,a,\theta_n) - r(x,y,a,\theta) \mid \to 0,$

(b) $P(n,\theta) := \sup_{x,a} \| p(\cdot \mid x,a,\theta_n) - p(\cdot \mid x,a,\theta) \|_v \to 0,$

(c) $Q(n,\theta) := \sup_{x,a} \| q(\cdot \mid x,a,\theta_n) - q(\cdot \mid x,a,\theta) \|_v \to 0,$

donde $\| \cdot \|_v$ es la norma de variación total.

Para el sistema reducido PDM(θ) con información completa tenemos el resultado siguiente.

PROPOSICIÓN (5.3). *Si* $\theta_n \to \theta$ *entonces:*

(a) $\overline{R}(n,\theta) := \sup_{v,a} \mid \bar{r}(v,a,\theta_n) - \bar{r}(v,a,\theta) \mid \to 0,$

(b) $\overline{P}(n,\theta) := \sup_{v,a} \| \overline{P}(\cdot \mid v,a,\theta_n) - \overline{P}(\cdot \mid v,a,\theta) \|_v \to 0.$

Demostración. **(a)** De (5.5) se sigue que $\overline{R}(n,\theta) \le R(n,\theta)$ y esto implica **(a)**. Para la parte **(b)**, recordemos que si p_1 y p_2 son medidas de probabilidad sobre V, entonces $p_1 - p_2$ es una medida con signo finita y

$$\| p_1(\cdot) - p_2(\cdot) \|_v = 2 \sup_C | p_1(C) - p_2(C) |$$

donde el sup se toma sobre todo $C \in \mathcal{B}(V)$. Sea $B_1 B_2$ cualquier rectángulo medible en V. Entonces, con $v = (y, z)$

$$\begin{aligned} \overline{P}(B_1 B_2 \mid v, a, \theta) &= \sum_{y' \in B_1} I_{B_2}[s(z, a, \theta, y')] R(y' \mid z, a, \theta) \\ &= R[\overline{B}_1(v, a, \theta; B_1) \mid z, a, \theta] \end{aligned} \tag{5.14}$$

$$= \int_X \int_X q(\overline{B}_1(z, a, \theta; B_1) \mid x', a, \theta) p(dx' \mid x, a, \theta) z(dx)$$

donde $\overline{B}_1(z, a, \theta; \mathcal{B}_1) = \{y' \in Y \mid s(z, a, \theta, y') \in B_2\} \in \mathcal{B}(Y)$. Ahora si $C_1, C_2 \in \mathcal{B}(Y)$,

$$\begin{aligned} &| R(C_1 \mid z, a, \theta_n) - R[C_2 \mid z, a, \theta) | \\ &= \left| \int_X \int_X [q(C_1 \mid x', a, \theta_n) p(dx' \mid x, a, \theta_n) - q(C_2 \mid x', a, \theta) p(dx' \mid x, a, \theta)] z(dx) \right| \\ &\leq \left| \int_X \int_X q(C_1 \mid x', a, \theta_n)[p(dx' \mid x, a, \theta_n) - p(dx' \mid x, a, \theta)] z(dx) \right| \\ &+ \left| \int_X \int_X p(dx' \mid x, a, \theta)[q(C_1 \mid x', a, \theta_n) - q(C_2 \mid x', a, \theta) z](dx) \right| \\ &\leq P(n, \theta) + Q(n, \theta). \end{aligned}$$

De (5.15) obtenemos usando A4θ que:

$$\sup_{v,a} \| R(\cdot \mid z, a, \theta_n) - R(\cdot \mid z, a, \theta) \|_v \to 0, \quad \text{cuando } n \to \infty \tag{5.16}$$

La parte (b) de la proposición se sigue ahora de (5.14) y (5.16). Esto termina la demostración.

Nota (5.1). En el caso en que $p(dx' \mid x, a, \theta)$ tenga una densidad $p_1(x' \mid x, a, \theta)$ con respecto a alguna medida finita λ sobre $(X, \mathcal{B}(X))$, la hipótesis A4$\theta(b)$ es equivalente a

$$\sup_{x,a} \int_X | p_1(x' \mid x, a, \theta_n) - p_1(x' \mid x, a, \theta) | \lambda(dx') \to 0.$$

5.1 Estimación de parámetros en PDM-PO(θ)

Para obtener I-políticas adaptables necesitamos un esquema de estimación consistente del valor verdadero del parámetro. Esto se puede conseguir de varias formas. Por ejemplo, se puede usar el método de estimación condicional por mínimos cuadrados (CMC) de [11]; este se relaciona estrechamente con el método de predición de error por mínimos cuadrados en [15]. Para describir el método CMC, fijemos una I-política $\bar{\delta}$ y denotemos por $\hat{y}_{n+1}(\theta)$ (si existe) el llamado "predictor de un paso hacia adelante":

$$\begin{aligned}\hat{y}_{n+1}(\theta) &= \overline{E}_0^{\bar{\delta},\theta}[y_{n+1} \mid \bar{h}_n, a_n] \\ &= \sum_{y \in Y} yR(y \mid z_n, a_n, \theta), \quad \bar{h}_n \in \overline{H}_n, \quad a_n \in A,\end{aligned}$$

donde $\{y_n\}$ es el proceso de observaciones. Consideremos la función

$$L_n(\theta) \equiv L_n(\theta; \bar{h}_n, a_n) := \sum_{k=0}^{n} [y_{k+1} - \hat{y}_{k+1}(\theta)]^2$$

Una función medible $\hat{\theta}_n : \overline{H}_n A \to T$ se llama un *estimador CMC* si $\hat{\theta}_n(\bar{h}_n, a_n)$ minimiza a la función $\theta \to L_n(\theta)$. En [11,15] se dan condiciones que aseguran la existencia de estimadores CMC que convergen casi seguramente al valor verdadero del parámetro. Otra forma de considerar el problema de estimación se puede basar en el método de contraste mínimo (c.m.), [6,12,16]. En [6] se han obtenido resultados sobre el método c.m. que se pueden aplicar a un PDM completamente observable con espacio de estados polaco. Bajo condiciones apropiadas de "*identificabilidad*", el método de estimación por c.m. incluye a los métodos de estimación de parámetros usuales, como máxima verosimilitud, mínimos cuadrados, etc.; ver [6,12,16]. Dependiendo de las situaciones específicas, se puede obtener una sucesión consistente de estimadores en varias formas y por lo tanto, esto nos permite definir el concepto siguiente [6,8,16].

Definición (5.1). Una sucesión de funciones $\hat{\theta}_n : \overline{H}_n A \to T$ se llama una sucesión de estimadores de $\theta \in T$ fuertemente consistente (FC) si $\hat{\theta}_n \to \theta$, $\overline{P}_0^{\bar{\delta},\theta}$ casi seguramnete cuando $n \to \infty$, para cualquier I-política $\bar{\delta}$.

5.2. I-políticas adaptables

Supongamos que θ^* es el valor verdadero del parámetro y que tenemos una sucesion $\{\hat{\theta}_n\}$ FC de estimadores de θ^*. Consideremos el sistema reducido PDM(θ): $(V, A, \overline{P}(\theta, \overline{r}(\theta))$ y los problemas de: obtener I-políticas adaptables $\overline{J}$-óptimas y aproximar el valor óptimo $g(\theta^*) \equiv \overline{J}^*_{\theta}(v_0)$. Siendo el PDM($\theta$) un PDM completamente observable con estados $v_n = (y_n, z_n)$, tenemos un PDM adaptable en el sentido usual [1,6,8,12,13] para el cual existen políticas adaptables y esquemas de aproximación. Resolveremos estos problemas usando los resultados de [1] para PDM adaptables.

Primero definimos la sucesión de funciones $u_n(v, \theta)$ para $v \in V$ y $\theta \in T$ por:

$$u_n(v,\theta) := \max_{a \in A(v)} \{\overline{r}(v,a,\theta) + \int u_{n-1}(v',\theta)\overline{P}(dv' \mid v,a,\theta)\}, \tag{5.17}$$

para $n = 0, 1, \ldots$, con $u_{-1} := 0$. La sucesión $\{u_n\}$ se llama *sucesión de iteración de valores.* Bajo las hipótesis A1θ–A3θ, usando el resultado dado en la parte (**iv**) de la proposición 5.1., los lemas 1 y 2 de [7] e inducción matemática, vemos que $u_n(v, \theta) \in C(VT)$ para $n \geq 0$. Podemos aplicar también los teoremas conocidos de selecciones medibles, e.g., en [6, ó 14, pg. 74], para obtener funciones medibles $f_n(v, \theta) : VT \to A$ tales que $f_n(v, \theta)$ es algún valor de a para el cual se alcanza el máximo en el lado derecho de la ecuación (5.17) para todo $n \geq 0$.

Definición 5.2. Sea $\hat{\delta} = \{\hat{\delta}_n\}$ la I-política definida por $\hat{\delta}_n(\overline{h}_n) = f_n(v_n, \hat{\theta}_n)$, para $\overline{h}_n \in \overline{H}_n$, $n \geq 0$. La política $\hat{\delta}$ se llama I-política *adaptable de iteración de valores.*

Otras políticas adaptables que consideramos son las siguientes.

Definimos la sucesión de funciones $w_n(v, \theta)$ para $(v, \theta) \in VT$ por : $w_0 \in C(VT)$, arbitraria, y para $n \geq 1$,

$$w_n \equiv \overline{G} w_{n-1}, \tag{5.18}$$

donde $\overline{G}$ es el operador de contracción definido en (5.8). La sucesión $\{w_n\}$ es la sucesión de *aproximaciones sucesivas* del operador $\overline{G}$, la cual converge al único punto fijo de $\overline{G}$. Sean $f'_n : VT \to A$ funciones medibles tales que $f'_n(v, \theta)$ es algún valor de a para el cual se alcanza el máximo en el lado derecho de la ecuación (5.18), $n \geq 0$ (aquí aplicamos de nuevo los resultados de [6,14]).

Definición (5.3). Sea $\hat{\delta}' = \{\hat{\delta}'_n\}$ la I-política definida por $\hat{\delta}'_n(\overline{h}_n) = f'_n(v_n, \hat{\theta}_n)$, para $\overline{h}_n \in \overline{H}_n$, $n \geq 0$. La política $\hat{\delta}'$ se llama *I-política adaptable de aproximaciones sucesivas.*

Finalmente consideremos la siguiente I-política. Para cada $\theta \in T$, sea $\psi(\cdot, \theta)$ una I-política estacionaria óptima, por ejemplo, como la obtenida en (5.11).

Definición (5.4). Sea $\hat{\delta}'' = \{\hat{\delta}_n''\}$ la I-política definida por $\hat{\delta}''(\overline{h}_n) = \psi(v_n, \hat{\theta}_n)$, para $\overline{h}_n \in \overline{H}_n$, $n \geq 0$. La política $\hat{\delta}''$ se llama *I-política adaptable de estimación y control* (PEC).

Vamos ahora a aplicar los resultados de [1] para establecer nuestros resultados principales sobre la optimalidad de las I-políticas adaptables definidas en esta sección.

TEOREMA (5.1). *Sea θ^* el valor verdadero del parámetro y sea $\{\hat{\theta}_n\}$ una sucesión FC de estimadores de θ^*. Supongamos además que:*
(i) *Se cumplen las hipótesis* A1θ–A4θ *y que la función $h \in C(VT)$ que satisface la ecuación de optimalidad* (5.10), *cumple la siguiente condición:*

$$\sup_{v \in V} | h(v, \theta') - h(v, \theta) | \to 0, \quad \text{cuando } \theta' \to \theta, \quad \text{para todo} \theta \in T. \tag{5.19}$$

ó bien,
(ii) *Se cumplen las hipótesis* A1θ–A3θ *y los espacios X e Y son compactos. Entonces:* (a) *la I-política adaptable $\hat{\delta}$ de iteración de valores dada en la definición* 5.2 *es j-óptima, y* (b) *para cualquier sucesión $\{\theta_n\}$ que converge a θ^*, el valor óptimo $g(\theta^*)$ se puede obtener como*

$$g(\theta^*) = \lim_{n \to \infty} [u_n(v_*, \theta_n) - u_{n-1}(v_*, \theta_n)], \tag{5.20}$$

donde v_* es cualquier estado fijo en V y la sucesión $\{u_n\}$ está definida en (5.17).

Demostracion. La demostración se obtiene directamente del teorema (5.9) de [1]. Sin embargo conviene dar aquí los hechos más importantes en los que se basa dicha demostración.

Definamos la sucesión $u_n' \in C(VT)$ por

$$u_n'(v, \theta) := u_n(v, \theta) - u_n(v^*, \theta), \quad (v, \theta) \in VT, \quad n \geq 1,$$

donde v^* es cualquier estado fijo en V. Como consecuencia de las hipótesis A1θ–A3θ, se obtiene que la sucesión $\{u_n'\}$ converge uniformemente a una función $h \in C(VT)$. Definamos también la sucesión $\{g_n\}$ en $C(T)$ por

$$g_n(\theta) := u_n(*, \theta) - u_{n-1}(v_*\theta), \quad \theta \in T, \quad n \geq 1,$$

donde v^* es cualquier estado fijo en V. Bajo A1θ–A3θ, la sucesión $\{g_n\}$ es acotada, y converge uniformemente a una función $g \in C(T)$. En [1] se obtiene que las funciones h y g satisfacen la ecuación de optimalidad (5.10). Esto demuestra que $g(\theta^*)$ se puede obtener como en (5.20).

Consideremos ahora la I-política adaptable $\hat{\delta}$ de iteración de valores. Usando la ecuación de optimalidad definamos la función ϕ sobre $\overline{K}'$ por

$$\phi(v,a,\theta);=\bar{r}(v,a,\theta)+\int h(v',\theta)\overline{P}(dv' \mid v,a,\theta)-h(v,\theta)-g(\theta)$$

La función ϕ permite comparar los controles obtenidos usando $\bar{\delta}$ con controles estacionarios óptimos (ver la proposición 5.2.). En [1] se demuestra, bajo las hipótesis de la parte (i) con la condición (5.19), ó bien bajo las hipótesis de la parte (ii), que:

$$\lim_{n\to\infty} \mid \phi(v_n, f_n(v_n,\hat{\theta}_n),\theta^*) \mid = 0, \quad \overline{P}_0^{\hat{\delta},\theta}\text{-casi seguramente,}$$

donde $(v_n, f_n(v_n,\hat{\theta}_n))$ son los estados y acciones obtenidos usando $\hat{\delta}$. La optimalidad de $\hat{\delta}$ se obtiene ahora usando la parte (iii) de la proposición (3.2) de [1]. Esto termina la demostración.

Observación (5.1). La parte (a) del teorema 5.1 nos da una forma de escoger los controles óptimos; la parte (b) nos da un esquema para aproximar el valor óptimo $g(\theta^*)$. De aquí obtenemos la solución a los problemas planteados al principio de la subsección 5.2.

Para las otras I-políticas adaptables definidas anteriormente tenemos el resultado siguiente.

TEOREMA (5.2). *Bajo las hipótesis del teorema* 5.1 *se tiene que la I-política de aproximaciones sucesivas y la I-política PEC, dadas en las definiciones* 5.3 *y* 5.4 *respectivamente, son* $\overline{J}$*-óptimas. Además, en el caso de la política* $\hat{\delta}'$ *de aproximaciones sucesivas, el valor óptimo* $g(\theta^*)$ *se puede obtener como*

$$g(\theta^*)=\bar{\alpha}\lim_{n\to\infty} w_{n-1}(v^*,\theta_n),$$

donde $\{w_n\}$ *está definida en* 5.18 *y el estado* $v^* \in V$ *y el número* $\bar{\alpha}$*, están dados en la parte (iii) de la proposición* 5.1.

La demostración del teorema 5.2 es análoga a la del teorema 5.1, y por lo tanto la omitimos.

En la sección siguiente ilustraremos los resultados de los teoremas anteriores con un ejemplo de un sistema de inventario parcialmente observable.

6. Ejemplo

Consideremos el ejemplo de control de inventario tratado en [1,6]. Sea x_n el nivel de inventario de un producto al tiempo n y supongase que x_n no es observable directamente (o que resulta muy costoso hacerlo) pero que se dispone de un proceso de observaciones $\{y_n\}$ con valores en un espacio finito Y. Supongamos que se tiene una capacidad finita C para x_n y que el espacio de estados es el intervalo $X=[0,C]$. Sean ρ_1 y ρ_2 los precios unitarios de compra y venta, respectivamente, de x_n, con ρ_1 y ρ_2 constantes $\rho_2 > \rho_1 > 0$. Supongamos que hay un costo de mantenimiento y que cuesta μx mantener un nivel de inventario x en cada etapa del proceso. Sea d_n la demanda en el intervalo

$[n, n+1)$. Supongamos que $\{d_n\}$ es una sucesión de variables aleatorias con esperanza finita, no negativas, independientes, todas con la misma distribución con densidad $f(\theta, \cdot) : \mathbf{R} \to \mathbf{R}$, la cual depende continuamente del parámetro $\theta \in T$, donde T es un conjunto compacto. La acción a_n es la cantidad de producto ordenada al principio de $[n, n+1)$, y suponemos que es entregada inmediatamente. Supongamos que el conjunto de acciones admisibles dada la observación y_n es $A(y_n) = [0, C - y_n]$ y que $A = [0, C]$. Con esto se cumple que la correspondencia $y \to A(y)$ es continua. Suponemos que la cantidad de producto que se vende durante $[n, n+1)$ es $v_n = \min(d_n, x_n + a_n)$. Entonces el estado al tiempo $n+1$ será $x_{n+1} = x_n + a_n - v_n$. La ganancia esperada en $[n, n+1)$ es

$$r(x, a, \theta) = \int_0^{x+a} \rho_2 s f(\theta, s) ds + \rho_2(x+a) \int_{x+a}^{\infty} f(\theta, s) ds - \rho_1 a - \mu(x+a).$$

Se tiene que $r \in C(VAT)$, y se cumple la hipótesis A4θ(a).

La probabilidad de transición está dada por

$$p(B \mid x, a, \theta) = \int_B f(\theta, s + a - y) dy,$$

si B es un subconjunto de Borel en $(0, C]$ y por

$$p(\{0\}) \mid x, a, \theta) = \int_{x+a}^{\infty} f(\theta, s) ds$$

para todo $(x, a, \theta) \in XAT$. Esta probabilidad de transición satisface las hipótesis A1θ–A3θ con $x^* = 0$ en A3θ y $\alpha_1 > 0$ tal que

$$\int_C^{\infty} f(\theta, s) ds \geq \alpha_1.$$

Supongamos que la densidad $f(\theta, \cdot)$ satisface la propiedad

$$\sup_{s \in \mathbf{R}} \mid f(\theta, s) - f(\theta_n, s) \mid \to 0, \quad \text{cuando } n \to \infty$$

para cualquier sucesión $\{\theta_n\}$ en T, convergente a θ. Entonces $p(\cdot \mid x, a, \theta)$ satisface que:

$$\sup_{x,a} \| p(\cdot \mid x, a, \theta) - p(\cdot \mid x, a, \theta_n) \|_v \leq (\sup_{s \in \mathbf{R}} \mid f(\theta, s) \mid)(1 + C) \to 0$$

cuando $n \to \infty$; por tanto p satisface la hipótesis A4θ(b).

Supongamos que el proceso de observaciones es tal que

$$y_n = g(x_n, a_{n-1}, s_n), \quad n \geq 1,$$

donde g es una función continua sobre XAX y las s_n son variables aleatorias independientes con la misma distribución discreta $q(\theta, \cdot)$, que depende del parámetro desconocido, concentradas en algún conjunto numerable $S = \{\beta_j\}$; supongamos además que $q(\theta, \cdot)$ satisface que:

$$\sum_j \mid q(\theta, \beta_j) - q(\theta_n, \beta_j) \mid \to 0, \quad \text{cuando } n \to \infty,$$

para cualquier sucesión $\{\theta_n\}$ en T convergente a θ. Supongamos también que para g existe un valor de la observación y^*, y un único $\beta^* \in S$ tales que

$$g(0, a, \beta^*) = y^* \quad \text{y} \quad g(x, a, \beta^*) \neq y^* \quad \text{si } x \neq 0, \quad \text{para todo } a,$$

y que $q(\theta, \beta^*) = \gamma > 0$ para todo $\theta \in T$. Entonces

$$q(y \mid x, a, \theta) = \sum_j I_y[g(x, a, \beta_j)] q(\theta, \beta_j),$$

para todo $y \in Y$ y $(x, a, \theta) \in XAT$, satisface que:

$$q(y^* \mid x', a, \theta) = \begin{cases} \gamma, & \text{si } x' = 0 \\ 0, & \text{si } x' \neq 0, \end{cases}$$

para todo $(a, \theta) \in AT$. Por lo tanto q satisface la hipótesis A3θ. También, $q(\cdot \mid x, a, \theta)$ es continua para todo $(x, a, \theta) \in XAT$ y satisface la condición:

$$\sup_{x,a} \| q(\cdot \mid x, a, \theta) - q(\cdot \mid x, a, \theta_n) \|_v \leq 2 \sum_j \mid q(\theta, \beta_j) - q(\theta_n, \beta_j) \mid \to 0,$$

cuando $n \to \infty$ para cualquier sucesión $\{\theta_n\}$ en T convergente a θ, y por lo tanto $q(\cdot \mid x, a, \theta)$ satisface la hipótesis A4θ(c). Con esto tenemos que se cumplen las hipótesis A1θ–A4θ. El parámetro θ se puede estimar por cualquier esquema consistente usual, como por ejemplo, máxima verosimilitud ó por el método de c.m. Podemos aplicar entonces os resultados de los teoremas 5.1 ó 5.2.

7. Conclusiones

En este trabajo hemos transformado un PDM-PO con parámetros desconocidos en un PDM completamente observable. Hemos visto como el PDM transformado hereda algunas propiedades deseables del sistema PO como por ejemplo la continuidad y las condiciones de ergodicidad. Indicamos como, bajo condiciones apropiadas, se pueden estimar consistentemente los parámetros desconocidos. También, combinamos la transformación del PDM-PO con resultados recientes sobre PDM con parámetros desconocidos y espacio de estados polaco, para obtener políticas adaptables $\overline{J}$-óptimas y esquemas de aproximación del valor óptimo.

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

REFERENCIAS

[1] R. S. ACOSTA ABREU, *Control de cadenas de Markov con parámetros desconocidos y espacio de estados métrico*, Bol. Soc. Mat. Mexicana

[2] K. J. ASTRÖM, *Optimal control of Markov processes with incomplete state information*. J. Math. Anal. Appl. **10**, 174-205, 1965.

[3] C. BERGE, *Espaces Topologiques*, Dunod, Paris, 1959.

[4] D. P. BERTSEKAS Y S. E. SHREVE, *Stochastical Optimal Control: Thet Discrete Time Case*, Academic Press, New York, 1978.

[5] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Proceses*, Springer, New York, 1979.

[6] J. P. GEORGIN, *Estimation et controle des chaines de Markov sur des espaces arbitraries*, Lecture Notes Math., **636**, 71-113, Springer, Berlín, 1978.

[7] L. G. GUBENKO, AND E. S. STATLAND *On controlled, discrete time Markov decision processes*, Theory Probab. Math. Statist., **7**, 47-61, 1975.

[8] O. HERNÁNDEZ-LERMA AND S. I. MARCUS, *Adaptative control of Markov processes with incomplete state information and unknown parameters*, J. Optim. Theory Appl., **52**, 227-241, 1987.

[9] ———, *Nonparametric adaptive control of discrete-time partially observable stochastic systems*, J. Math. Anal. Appl. **137**, 312-334, 1989.

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

[11] L. A. KLIMKO AND P. T. NELSON, *On conditional least squares estimation for stochastic processes*, Ann. Statist ., **6**, 629-642, 1978.

[12] M. KOLONKO, *Strongly consistent estimation in a controlled Markov renewal model*, J. Appl. Probab., **19**, 532-545, 1982.

[13] P. R. KUMAR, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., **23**, 329-380, 1985.

[14] K. KURATOWSKI, *Topology, volume II*, Academic Press, New York, 1968.

[15] L. LJUNG, *Analysis of a general recursive prediction error identification algorithm*, Automatica, **17**, 89-99, 1981.

[16] P. MANDL, *Estimation and control of Markov chains*, Adv. in Appl. Probab., **6** 40-60, 1974.

[17] G. E. MONAHAN, *A survey of partially observable Markov decision processes: Theory, models, and algorithms*, Man. Sci., **28**, 1-16, 1982.

[18] K. R. PARTHASARATHY, *Probability measures on metric spaces*, Academic Press, New York, 1967.

[19] D. RHENIUS *Incomplete information in Markovean desicion models*, Ann. Statist., **2**, 1327-1334, 1974.

[20] Y. SAWARAGI AND T. YOSHIKAWA, *Discrete-time Markovian decision processes with incomplete state observation*, An. Math. Statist., **41**, 78-86, 1970.

[21] K. WAKUTA, *Semi-Markov decision processes with incomplete state observation —average cost criterion—*, J. Oper. Res. Soc. Japan, **24**, 95-109, 1981.

[22] A. A. YUSHKEVICH, *Reduction of a controlled Markov model with complete data to a problem with incomplete information in the cae of Borel state and control spaces*, Theory Probab. Appl., **21**, 153-158, 1976.