Morfismos, Vol. 16, No. 2, 2012, pp. 1–28

Procesos markovianos en la toma de decisiones: contribuciones de Onésimo Hernández-Lerma

Francisco Venegas-Martínez

Resumen

En esta investigación se lleva a cabo una revisión de algunas de las aportaciones de Onésimo Hernández-Lerma a las matemáticas y, específicamente, a la teoría y práctica de los procesos markovianos en la toma de decisiones de agentes racionales. Se destacan, particularmente, sus extensiones, reformulaciones y nuevos planteamientos en los procesos markovianos de decisión, juegos estocásticos, optimalidad de Blackwell, optimalidad en sesgo, optimalidad rebasante y control óptimo estocástico.

2010 Mathematics Subject Classification: 91A15, 91A35, 49J20, 60J25, 93C05.

Palabras y frases clave: Procesos markovianos de decisión, juegos estocásticos, control óptimo estocástico.

Abstract

This paper conducts a review of some of the contributions of Onésimo Hernández-Lerma to the mathematics and specifically to the theory and practice of Markov processes in the decision making of rational agents. It highlights, in particular, his extensions, reformulations and new approaches to Markov decision processes, stochastic games, Blackwell optimality, bias optimality, overtaking optimality, and stochastic optimal control.

Keywords: Markov decision procesess, stochastic games, stochastic optimal control.

1 Introducción

En las últimas décadas, los modelos matemáticos de toma de decisiones de agentes racionales han experimentado una serie de cambios y transformaciones profundas. Estos cambios han abierto nuevos paradigmas

que resaltan la exposición de los agentes a diferentes tipos de riesgos. Esto, por supuesto, es una necesidad irremisible para modelar apropiadamente la realidad contingente y no sólo una sofisticación más en el tratamiento matemático. En este sentido, las investigaciones de Onésimo Hernández-Lerma han abierto un sinnúmero de horizontes a la teoría de los procesos markovianos y, como consecuencia, han conducido a la utilización de herramientas matemáticas más robustas, las cuales permiten una mejor comprensión de qué hacen y porqué lo hacen los agentes al tomar decisiones en ambientes de riesgo e incertidumbre. En este contexto, los procesos markovianos ocupan un lugar privilegiado, ya sea por sus bondades técnicas en el modelado del proceso de toma de decisiones o bien por su potencial riqueza en diversas aplicaciones.

Es también importante destacar que la crisis mundial de 2007-2009 y sus múltiples lecciones han sido un factor determinante en las diversas reformulaciones recientes de la teoría de los procesos markovianos, haciendo ineludible replantear el modelado del riesgo y la incertidumbre en los procesos de toma de decisiones de los agentes. Las exigencias mismas de la realidad contingente han motivado diversas extensiones de las teorías existentes y, en muchas ocasiones, ha sido necesaria la reformulación de nuevos paradigmas teóricos. En particular, los procesos markovianos de decisión, los juegos estocásticos, la optimalidad de Blackwell, la optimalidad en sesgo, la optimalidad rebasante y el control óptimo estocástico, han tenido un desarrollo notable en los últimos años, y esto en gran parte por las contribuciones de Onésimo Hernández-Lerma y a sus colaboradores en todo el mundo. Esto se puede apreciar en: Prieto-Rumeau y Hernández-Lerma (2012) en juegos markovianos y cadenas de Markov en tiempo continuo; Guo y Hernández-Lerma (2009) en procesos markovianos de decisión en tiempo continuo; Hernández-Lerma y Lasserre (1996) sobre criterios de optimalidad de proceso markovianos controlados en tiempo discreto; Hernández-Lerma y Lasserre (1999) en sus investigaciones en procesos markovianos controlados en tiempo discreto; Hernández-Lerma y Lasserre (2003) en cuanto a cadenas de Markov y probabilidades invariantes[1]; Hernández-Lerma (1989) por sus aportaciones en procesos markovianos adaptativos; y Hernández-Lerma (1990) y (1994) por sus contribuciones a los procesos markovianos en tiempo discreto y tiempo continuo, respectivamente.

[1]Véase también el trabajo de Hernández-Lerma y Lasserre (2001b) sobre cadenas de Markov y procesos de Harris.

La formación y trayectoria de Onésimo Hernández-Lerma se sustentan en su ardua y constante labor en el desarrollo de las matemáticas. Obtuvo la Licenciatura en Física y Matemáticas en la Escuela Superior de Física y Matemáticas (ESFM) del Instituto Politécnico Nacional (IPN) en 1970. Posteriormente, obtuvo la maestría, en 1976, y el doctorado, en 1978, en la División de Matemáticas Aplicadas de *Brown University* en los Estados Unidos de Norteamérica. Actualmente es Profesor Emérito y Jefe del Departamento de Matemáticas del Centro de Investigación y de Estudios Avanzados (CINVESTAV) del IPN.

Por su reconocida trayectoria y múltiples e importantes contribuciones a las matemáticas a través de la investigación, la formación de recursos humanos y la divulgación científica, Onésimo Hernández-Lerma recibió, en 2009, el Premio Thomson Reuters por ser el matemático mexicano con el mayor número de citas, (alrededor de 1200); en 2008, obtuvo el Premio Scopus de la Editorial Elsevier por el alto impacto de sus publicaciones; en 2008, la Presea Lázaro Cárdenas del IPN, la máxima distinción que otorga esta Institución a un investigador; en 2003, el Doctor Honoris Causa por la Universidad de Sonora (UNISON); y, en 2001, el Premio Nacional de Ciencias y Artes otorgado por el Gobierno de los Estados Unidos Mexicanos a través de la Secretaría de Educación Pública. Asimismo ha sido distinguido como miembro del Consejo Consultivo de Ciencias de la Presidencia de la República. Es también miembro del Sistema Nacional de Investigadores (SNI) con el nivel III desde 1993. Las áreas de investigación de Onésimo Hernández-Lerma son muchas y entre ellas destacan: el control óptimo de sistemas estocásticos, la teoría de juegos estocásticos, la programación lineal infinita y los procesos markovianos. Sobre estos temas, ha publicado alrededor de 140 artículos de investigación en revistas especializadas con estricto arbitraje y pertenecientes a índices internacionales de gran prestigio, así como 11 libros y monografías en casas editoriales de amplio reconocimiento. Su labor en la formación de recursos humanos a nivel de posgrado ha sido excepcional, habiendo graduado 18 estudiantes de doctorado y 39 de maestría. Además, por su liderazgo en el ámbito mundial ha recibido visitantes posdoctorales de Francia, España y la República Popular China, entre otros países.

El propósito del presente trabajo consiste en realizar una revisión, la cual no pretende ser exhaustiva, sobre las contribuciones y reformulaciones de Onésimo Hernández-Lerma en procesos markovianos. Muchas de las contribuciones recientes en procesos markovianos de decisión,

juegos estocásticos y control óptimo estocástico se deben a Onésimo Hernández-Lerma y a sus coautores en todo el mundo. Una ventaja didáctica de las aportaciones de sus investigaciones en el modelado del proceso de la toma secuencial de decisiones de agentes racionales, en tiempo discreto o continuo y en ambientes con riesgo e incertidumbre, es que todas sus investigaciones proporcionan una visión unificada y consistente.

Este trabajo está organizado de la siguiente manera: en la próxima sección se revisan los procesos markovianos de decisión; en la tercera sección se estudian los juegos markovianos; a través de la cuarta sección se analiza la optimalidad en sesgo y optimalidad rebasante; en el quinto apartado se examina la optimalidad de Blackwell para procesos markovianos de difusión controlados; en el transcurso de la sexta sección se revisa la teoría de control óptimo estocástico con procesos markovianos de difusión; por último, en la séptima sección se proporcionan las conclusiones, destacando las áreas de oportunidad para extender la teoría de los procesos markovianos.

2 Procesos markovianos de decisión

Existen muchos sistemas en ciencias naturales y sociales en donde los eventos futuros tienen asociada una distribución de probabilidad que depende sólo del presente, en cuyo caso podría ser idóneo modelarlos con cadenas de Markov. Varias preguntas surgen en el comportamiento de una cadena de Markov: ¿Cómo evoluciona un proceso de este tipo? ¿Converge, en algún sentido, a un estado estacionario? ¿Qué tan rápido converge? Estas preguntas han sido ampliamente contestadas en la literatura cuando la cadena de Markov tiene un número finito de estados ¿Pero qué sucede cuando hay un número infinito de estados, numerable o continuo? Al respecto, Hernández-Lerma y Lasserre (2003) se ocupan de las cadenas de Markov homogéneas en tiempo discreto con espacios arbitrarios de estados y con un comportamiento ergódico descrito con medidas de probabilidad invariantes. En particular, esta sección se concentra en procesos markovianos controlados en tiempo discreto y con horizonte de planeación finito o infinito. Muchos fenómenos y situaciones de interés son susceptibles de ser modelados bajo este esquema[2];

[2]Este esquema es también conocido como programación dinámica estocástica en tiempo discreto.

por ejemplo, la toma de decisiones de consumo, producción, inversión, y la valuación de proyectos de inversión, ya sea en el corto o largo plazo.

Una clase relevante de procesos de control la constituyen los procesos de control markovianos. La evolución de estos procesos en tiempo discreto se puede describir como sigue. El sistema se encuentra, inicialmente, en el estado $i_0 = x$ entonces el controlador elige una acción (o control) $a_0 = a$, lo que genera un costo, $r(x, a)$, que depende del estado y el control. Posteriormente, el sistema se mueve a un nuevo estado $i_1 = y$ de acuerdo con una ley de transición en la que el futuro sólo está determinado por el presente. Este procedimiento se repite y los costos se acumulan. Se dice que $\{i_n : \ n = 1, 2, \ldots\}$ es un proceso de control markoviano, en tiempo discreto, si para cualquier estrategia π (una función de las sucesiones de estados y acciones) y cualquier $n = 0, 1, \ldots$, la distribución en $n + 1$, dada toda la historia del proceso hasta n, depende sólo del estado y la acción en n. Los estados y las acciones son colecciones de variables aleatorias, definidas en un espacio de probabilidad adecuado, y el objetivo es encontrar una política de control que optimice un criterio de desempeño (en términos de valores esperados).

A continuación se presentan, en forma breve, los elementos que integran un proceso markoviano de decisión, abreviado mediante

$$\{S, A, K, q, r\}.$$

Para simplificar la exposición se supone que el espacio de estados, S, es discreto. Considere una cadena de Markov controlada en tiempo discreto con:

i) Un espacio de estados, finito o numerable S.

ii) Un espacio medible de acciones, A, equipado con una σ-álgebra $\mathcal{A}$ de A. En este caso, el conjunto de restricciones se representa mediante $K = S \times A$.

iii) Para cada estado $i \in S$ existe un conjunto de acciones $A(i)$ disponibles. Estos conjuntos se suponen elementos de $\mathcal{A}$.

iv) Una matriz de probabilidades de transición $[q(j \mid i, a)]$. Para cada $i, j \in S$ y $a \in A(i)$ la función $q(j \mid i, a)$ es no negativa y medible, y $\sum_{j \in S} q(j \mid i, a) = 1$ para cada $i \in S$ y $a \in A(i)$.

v) Una función $r : K \to \mathbb{R}$, llamada la utilidad, ganancia o costo, dependiendo del contexto.

Sea $H_n = S \times (S \times A)^n$ el espacio de historias hasta el tiempo $n = 0, 1, \ldots$. Sea $H = \bigcup_{0 \leq n < \infty} H_n$ el espacio de todas las historias finitas. Los espacios H_n y H están equipados con las σ-álgebras generadas por 2^s y $\mathcal{A}$. Una estrategia π es una función que asigna a cada historia de estados y acciones $h_n = (i_0, a_0, i_1, \ldots, i_{n-1}, a_{n-1}, i_n) \in H_n$, $n = 0, 1, \ldots$, una medida de probabilidad $\pi(\cdot \mid h_n)$ definida en $(A, \mathcal{A})$ que satisface las siguientes condiciones:

a) $\pi(A(i_n) \mid h_n) = 1$,

b) Para cualquier $B \in \mathcal{A}$ la función $\pi(B \mid \cdot)$ es medible en H.

Una estrategia de Markov ϕ es una sucesión de funciones $\phi_n : S \to A$, $n = 0, 1, \ldots$, tal que $\phi_n(i) \in A(i)$ para cualquier $i \in S$. Se dice que una estrategia de Markov ϕ es (N, ∞)-estacionaria, donde $N = 0, 1, \ldots$, si $\phi_n(i) = \phi_N(i)$ para cualquier $n = N+1, N+2$ y cualquier $i \in S$. A una estrategia $(0, \infty)$-estacionaria se le llama, simplemente, estacionaria. De esta manera, una estrategia estacionaria se determina por una función $\phi : S \to A$ tal que $\phi(i) \in A(i)$, $i \in S$.

Una estrategia estacionaria aleatorizada ϕ es definida por sus distribuciones condicionales $\phi(\cdot \mid i)$, $i \in S$, sobre $(A, \mathcal{A})$ de tal forma que $\phi(A(i) \mid i) = 1$ para cualquier $i \in S$. Observe que en esta construcción "canónica", los procesos de estados y de acciones son colecciones de variables aleatorias. El conjunto H_∞ de todas las sucesiones de estados-acciones

$$(i_0, a_0, i_1, a_1, ..., i_{n-1}, a_{n-1}, i_n, a_n, ...)$$

y su correspondiente sigma-álgebra producto, F, forman un espacio medible (H_∞, F). Así, cada estrategia π y estado inicial $i_0 = x$ inducen una única medida de probabilidad P_x^π sobre H_∞, en cuyo caso se denota al operador de esperanza por E_x^π. Así, la utilidad total descontada[3] cuando el estado inicial es i y la estrategia utilizada es π está dada por:

$$V(i \mid \pi) = E_i^\pi \left[\sum_{t=0}^{\infty} \beta^t r(i_t, a_t) \right]$$

donde $\beta \in (0, 1)$ es el factor de descuento. La función de valor del problema planteado se define mediante

$$V(i) = \sup_\pi V(i \mid \pi),$$

[3]Se pueden utilizar diversos criterios de desempeño; véase, por ejemplo, Feinberg(1982).

Sea ϵ una constante no negativa. Una estrategia π^* se llama ε-óptima si para todo estado inicial i

$$V(i \mid \pi) \geq V(i) - \varepsilon.$$

Una estrategia 0-óptima se llama, simplemente, óptima.

Con respecto del esquema anterior, Hernández-Lerma (1989) considera sistemas de control estocástico parcialmente observables en tiempo discreto. El autor estudia el problema de control adaptativo no paramétrico, en un horizonte infinito, con el criterio de ganancia total descontada y proporciona las condiciones para que una política adaptativa sea asintóticamente óptima, así mismo establece condiciones para aproximar uniformemente, casi seguramente, la función de ganancia óptima. Su trabajo combina resultados de convergencia con problemas de control estocástico adaptativo y paramétrico.

Asimismo, Hernández-Lerma (1986) proporciona procedimientos de discretización de procesos markovianos de control adaptativo, en tiempo discreto, con un número finito de estados y en un horizonte de planeación infinito, los cuales dependen de parámetros desconocidos. En su investigación las discretizaciones se combinan con un esquema coherente de estimación de parámetros para obtener aproximaciones uniformes a la función de valor óptimo, así como para determinar políticas de control adaptativas asintóticamente óptimas.

Por otro lado, Hernández-Lerma (1985), bajo el criterio de ganancia descontada y con un espacio de estados numerable, estudia los procesos semi-markovianos de decisión que dependen de parámetros desconocidos. Y dado que los valores verdaderos de los parámetros son inciertos, el autor proporciona un esquema iterativo para determinar asintóticamente la máxima ganancia total descontada. Las soluciones toman el esquema iterativo de valor no estacionario de Federgruen y Schweitzer (1981) y se combinan con el principio de estimación y control para el control adaptativo de procesos semi-markovianos de Schäl (1987).[4]

Asimismo, Jasso-Fuentes y Hernández-Lerma (2007) estudian una clase general de los procesos markovianos de difusión con ganancia media esperada (también conocido como ganancia ergódica) y proporcionan algunos criterios "sensibles" al descuento. Estos autores dan las

[4]Véase también Hernández-Lerma y Marcus (1987).

condiciones bajo las cuales varios criterios de optimalidad son equivalentes. Otros trabajos relacionados se encuentran en: Guo y Hernández-Lerma (2003a) al estudiar cadenas de Markov controladas en tiempo continuo; Guo y Hernández-Lerma (2003b) al proporcionar condiciones de tendencia y monotonicidad para procesos markovianos de control en tiempo continuo bajo el criterio de pago promedio; Guo y Hernández-Lerma (2003c) que analizan cadenas de Markov controladas en tiempo continuo con el criterio de pagos descontados; y Hernández-Lerma y Govindan (2001) quienes investigan el caso de procesos markovianos de control no estacionarios con pagos descontados en un horizonte infinito.

Por último es importante destacar que Hernández-Lerma (1986) extiende el esquema iterativo introducido por White (1980) para un número finito de estados con el propósito de aproximar la función de valor de un proceso markoviano con un conjunto numerable de estados a un espacio multidimensional numerable de estados. Bajo los mismos supuestos de White (1980), el autor proporciona un esquema iterativo para determinar asintóticamente una política óptima descontada, la cual, a su vez, se puede utilizar para obtener una política óptima estacionaria.[5]

3 Juegos markovianos en tiempo continuo

En esta sección se formaliza un juego estocástico de suma cero con dos jugadores en tiempo continuo y homogéneo[6]. Los elementos que conforman dicho juego se expresan en forma abreviada como

$$\{S, A, B, K_A, K_B, q, r\}.$$

Aquí, S es el espacio de estados, el cual se supone numerable, y A y B son los espacios de acciones para los jugadores 1 y 2, respectivamente. Estos espacios se suponen espacios polacos (*i.e.*, espacios métricos, separables y completos). Los conjuntos $K_A \subset S \times A$ y $K_B \subset S \times B$ son espacios de Borel que representan conjuntos de restricciones. Es decir, para cada estado $i \in S$, la i-sección en K_A, a saber, $A(i) :=$

[5]Otros trabajos relacionados con el tema son Hernández-Lerma (1985) y Hernández-Lerma y Marcus (1984) y (1985).

[6]En Hernández-Lerma (1994) se encuentra una introducción a los procesos markovianos de control en tiempo continuo. Asimismo, el caso discreto es tratado en Hernández-Lerma y Lasserre (1999).

$\{a \in A \mid (i,a) \in K_A\}$ representa el conjunto de acciones admisibles para el jugador 1 en el estado i; similarmente, la i-sección en K_B, $B(i) := \{b \in B \mid (i,b) \in K_B\}$, representa la familia de acciones admisibles para el jugador 2 en el estado i. Considere ahora el subconjunto de Borel dado $S \times A \times B$ y defina

$$K := \{(i,a,b) \mid i \in S,\ a \in A(i),\ b \in B(i)\}.$$

La componente q denota la matriz $[q(j|i,a,b)]$ de tasas de transición del juego, la cual satisface $[q(j|i,a,b)] \geq 0$ para toda $(i,a,b) \in K$, $i \neq j$, y se supone conservativa, es decir,

$$\sum_{j \in S} [q(j|i,a,b)] = 0 \qquad \forall\ (i,a,b \in K)$$

y estable, esto es,

$$q(i) := \sup_{a \in A(i),\ b \in B(i)} q_i(a,b) < \infty,\ \forall\ i \in S,$$

donde $q_i(a,b) = -q(i \mid i,a,b)$ para toda $a \in A(i)$ y $b \in B(i)$. Además, $q(j \mid i,a,b)$ es una función medible en $A \times B$ para $i,j \in S$ fijas. Por último, $r : K \to \mathbb{R}$ es la tasa de ganancia (o utilidad) del jugador 1 (o la tasa de costo para el jugador 2).

Los jugadores 1 y 2 observan continuamente el estado presente del sistema. Siempre que el sistema esté en el estado $i \in S$ en el tiempo $t \geq 0$, los jugadores eligen de manera independiente las acciones $a \in A(i)$ y $b \in B(i)$, conforme a algunas estrategias admisibles introducidas más adelante. Como una consecuencia de esto, ocurre lo siguiente: (1) el jugador 1 recibe una ganancia $r(i,a_t,b_t)$; (2) el jugador 2 incurre en una pérdida $r(i,a_t,b_t)$ (se dice que el juego es de suma cero porque lo que un jugador gana, el otro irremediablemente lo pierde) y (3) el sistema se mueve a un nuevo estado $j \neq i$ con una función de transición posiblemente no homogénea determinada por las tasas de transición $[q(j \mid i,a,b)]$. El objetivo del jugador 1 es maximizar su ganancia, mientras que para el jugador 2 es minimizar su costo o pérdida con respecto a algún criterio de desempeño, V_α, el cual definirá posteriormente.

Sea X un espacio polaco. Denótese por $B(X)$ su σ-álgebra de Borel, y por $P(X)$ el espacio de Borel de medidas de probabilidad definidas sobre X, equipado con la topología de convergencia débil. Una estrategia markoviana para el jugador 1, denotada por π^1, es una familia $\{\pi_t^1,\ t \geq 0\}$ de núcleos estocásticos que satisfacen:

(1) Para cada $t \geq 0$ e $i \in S$, $\pi_t^1(\cdot \mid i)$ es una medida de probabilidad sobre A tal que $\pi_t^1(A(i) \mid i) = 1$;

(2) Para cada $E \in B(A)$ e $i \in S$, la función $t \mapsto \pi_t^1(E \mid i)$ es Borel medible para $t \geq 0$.

Sin pérdida de generalidad, en virtud de (1), también se puede ver a $\pi_t^1(\cdot \mid i)$ como una medida de probabilidad sobre $A(i)$. Asimismo, se denotará por Π_1^m a la familia de todas las estrategias markovianas del jugador 1. Una estrategia markoviana $\pi^1 = \{\pi_t^1(\cdot \mid i),\ t \geq 0\} \in \Pi_1^m$ es llamada estacionaria si para cada $i \in S$ existe una medida de probabilidad $\pi_t^1(\cdot \mid i) \in P(A(i))$ tal que $\pi_t^1(\cdot i) = \pi^1(\cdot \mid i)$ para toda $t \geq 0$; esta política se denota mediante $\{\pi^1(\cdot \mid i),\ i \in E\}$. El conjunto de todas las estrategias estacionarias del jugador 1 es representada por Π_1^s. La misma notación es utilizada para el jugador 2, con $P(B(i))$ en lugar de $P(A(i))$. Para cada par de estrategias, $(\pi^1, \pi^2) := \{(\pi_t^1, \pi_t^2),\ t \geq 0\} \in \Pi_1^m \times \Pi_2^m$, las tasas de ganancia y de transición se definen, respectivamente, para cada $i, j \in S$ y $t \geq 0$, como:

$$q(j \mid i, t, \pi^1, \pi^2) := \int_{B(i)} \int_{A(i)} q(j \mid i, a, b) \pi_t^1(\mathrm{d}a \mid i) \pi_t^2(\mathrm{d}b \mid i)$$

y

$$r(t, i, \pi^1 \pi^2) := \int_{B(i)} \int_{A(i)} r(i, a, b) \pi_t^1(\mathrm{d}a \mid i) \pi_t^2(\mathrm{d}b \mid i).$$

En particular, cuando π^1 y π^2 son ambas estacionarias, las expresiones anteriores se escriben, usualmente, como $q(j \mid i, \pi^1, \pi^2)$ y $r(i, \pi^1, \pi^2)$, respectivamente. Considere ahora la matriz $Q(t, \pi^1, \pi^2) = [q(j \mid i, t, \pi^1, \pi^2)]$ una función de transición (tal vez subestocástica) $\overline{p}(s, i, t, j, \pi^1, \pi^2)$ para la cual $Q(t, \pi^1, \pi^2)$ es su matriz de tasas de transición, es decir,

$$\left.\frac{\partial \overline{p}(s, i, t, j, \pi^1, \pi^2)}{\partial t}\right|_{t=s} = q(j \mid i, s, \pi^1, \pi^2)$$

para todo $i, j \in S$ y $s \geq 0$ es llamada un proceso de tipo Q. Un proceso de tipo Q, $\overline{p}(s, i, t, j, \pi^1, \pi^2)$, es llamado honesto si $\sum_{j \in S} \overline{p}(s, i, t, j, \pi^1, \pi^2) = 1$ para toda $i \in S$ y $t \geq s \geq 0$.

En lo que sigue se definen Π_1 y Π_2 como subconjuntos de estrategias markovianas que contienen a Π_1^s y Π_2^s y que satisfacen la condición de continuidad de las correspondientes tasas de transición para $t \geq 0$ y

para toda estrategia en Π_1 y Π_2. De esta manera, $q(j \mid t, i, \pi^1, \pi^2)$ es continua en $t \geq 0$, para todo $i, j \in S$, y $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$.

Para cada pareja de estrategias $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, los datos iniciales $(s, i) \in \overline{S} := [0, \infty) \times S$ y un factor de descuento $\alpha > 0$, el criterio de pago descontado $V_\alpha(s, i, \pi^1, \pi^2)$ se define como

$$V_\alpha(s, i, \pi^1, \pi^2) := \int_s^\infty \left[\sum_{j \in S} p(s, i, t, j, \pi^1, \pi^2) r(t, j, \pi^1, \pi^2) \right] e^{-\alpha(t-s)} \mathrm{d}t.$$

Las siguientes dos funciones:

$$L(s, i) := \sup_{\pi^1 \in \Pi_1} \inf_{\pi^2 \in \Pi_2} V_\alpha(s, i, \pi^1, \pi^2)$$

y

$$U(s, i) := \inf_{\pi^2 \in \Pi_2} \sup_{\pi^1 \Pi_1} V_\alpha(s, i, \pi^1, \pi^2)$$

definidas sobre $\overline{S}$ son llamadas el valor inferior y el valor superior respectivamente, del juego con pago descontado. Es claro que

$$L(s, i) \leq U(s, i) \ \ \forall \ (s, i) \in \overline{S}.$$

Si $L(s, i) = U(s, i)$ para toda $(s, i) \in \overline{S}$, entonces a la función común se le llama el valor del juego y es denotada por V.

Suponga que el juego tiene un valor V, entonces una estrategia π^{*1} en Π_1 se dice que es óptima para el jugador 1 si

$$\inf_{\pi^2 \in \Pi_2} V_\alpha(s, i, \pi^{*1}, \pi^2) = V(s, i) \ \ \forall \ (s, i) \in \overline{S}.$$

Similarmente, π^{*2} en Π_2 es óptima para el jugador 2 si

$$\sup_{\pi^1 \in \Pi^1} V_\alpha(s, i, \pi^1, \pi^{2*}) = V(s, i) \ \ \forall \ (s, i) \in \overline{S}.$$

Si $\pi^{*k} \in \Pi_k$ es óptima para el jugador k $(k = 1, 2)$, entonces el par (π^{*1}, π^{*2}) es llamado una estrategia óptima.

Con respecto del planteamiento anterior es importante destacar que Guo y Hernández-Lerma (2005a) estudian juegos de suma cero para cadenas de Markov en tiempo continuo con la posibilidad de que las utilidades y las tasas de transición sean no acotadas, esto bajo el criterio de

utilidad total descontada.[7] Estos autores proporcionan las condiciones bajo las cuales se garantiza la existencia del valor del juego y obtienen estrategias estacionarias óptimas mediante la ecuación de optimalidad de Shapley (1953). Asimismo, Guo y Hernández-Lerma (2005a) proponen un esquema de iteración de valores y demuestran su convergencia. El esquema converge hacia el valor del del juego y también hacia estrategias estacionarias óptimas. Por otra parte, cuando las tasas de transición son acotadas, se demuestra que la convergencia de esquema de iteración de valores es exponencial.

Otro trabajo relacionado es el de Hernández-Lerma y Lasserre (2001a) quienes analizan el caso de juegos estocásticos de suma cero con dos jugadores en espacios de Borel bajo el criterio de pago promedio en tiempo discreto. Este criterio de ganancia (esperada) promedio se precisa a continuación. Para cada política de control medible f y $x \in \mathbb{R}^n$ se define la ganancia promedio esperada de f dado el estado inicial $x \in \mathbb{R}^n$, bajo la tasa de ganancia $r(t, x(t), f)$, como

$$J(x, f, r) := \lim_{T\to\infty} \inf \frac{1}{T} E_x^f \left[\sum_{t=0}^{T-1} r(t, x(t), f) \right].$$

La función

$$J^*(x, r) := \sup_f J(x, f, r)$$

con $x \in \mathbb{R}^n$ es llamada la ganancia promedio óptima. Si existe una política f^* para la cual

$$J(x, f^*, r) = J^*(x, r)$$

para toda $x \in \mathbb{R}^n$, entonces se dice que f^* es una política promedio óptima.

Asimismo, Guo y Hernández-Lerma (2003d) han estudiado juegos de suma cero de dos personas para cadenas de Markov en tiempo continuo con un criterio de ganancia media (o promedio). Las tasas de transición pueden ser no acotadas, y las tasas de ganancia pueden no tener cotas superiores ni inferiores. Con respecto de las condiciones de tendencia y monotonicidad de los procesos de Markov en tiempo continuo, estos autores proporcionan las condiciones en los datos primitivos de un sistema controlado bajo las cuales se garantiza la existencia del

[7] Véase también Guo y Hernández-Lerma(2003b).

valor del juego y un par de estrategias estacionarias óptimas mediante el uso de la ecuación de optimalidad de Shapley (1953). Por último, presentan una caracterización de martingala de un par de estrategias óptimas estacionarias.

Por otro lado, Guo and Hernández-Lerma (2005b) realizan un estudio sobre juegos de suma no cero de dos personas para cadenas de Markov en tiempo continuo bajo el criterio de pago descontado en espacios de acción de Borel. Las tasas de transición son, posiblemente, no acotadas, y las funciones de pago podrían no tener cotas superiores ni inferiores. En este trabajo se proporcionan las condiciones que garantizan la existencia de equilibrios de Nash en estrategias estacionarias. Para el caso de juegos de suma cero, demuestran la existencia del valor del juego, y también proporcionan una forma recursiva de calcularlo, o al menos aproximarlo. Estos autores también demuestran que si las tasas de transición están uniformemente acotadas, entonces un juego de tiempo continuo es equivalente, en cierto sentido, a un juego markoviano en tiempo discreto.

Por último, Guo y Hernández-Lerma (2007) extienden sus investigaciones de juegos de suma cero de dos personas para procesos de Markov de saltos en tiempo continuo con un criterio de pago con descuento. Los espacios de estados y de acciones son espacios polacos (espacios métricos, separables y completos), las tasas de transición pueden ser no acotadas, y las tasas de ganancia pueden no tener cotas superiores ni inferiores. En este trabajo, los autores extienden los resultados en Guo y Hernández-Lerma (2003d) a procesos markovianos de saltos en tiempo continuo.

Suponga que $J_T(f)$ denota la ganancia total esperada durante el intervalo de tiempo $[0, T]$ cuando se utiliza la política de control f. Sea

$$g(f) := \lim_{T\to\infty} \inf \frac{1}{T} J_T(f)$$

la ganancia promedio correspondiente. Si f y f' son dos políticas tales que

$$J_T(f) = J_T(f') + T^{\theta}$$

para toda $T > 0$ y algún $\theta \in (0, 1)$, entonces se tienen dos políticas que producen la misma ganancia promedio aunque sus ganancias en un horizonte finito son diferentes. Así, el criterio de ganancia promedio no

distingue entre las políticas f y f'. Para evitar este comportamiento, se imponen condiciones bajo las cuales las ganancias en un horizonte finito de políticas estacionarias son necesariamente de la forma

$$J_T(f) = Tg(f) + h_f(\cdot) + e(f, T),$$

donde $h_f(\cdot)$ es el sesgo de f y $e(f, T)$ es el término residual que tiende a 0 cuando $T \to \infty$. En consecuencia, si f y f' son dos políticas estacionarias con la misma ganancia promedio, entonces

$$J_T(f) - J_T(f') = h_f(\cdot) - h_{f'}(\cdot) + [e(f, T) - e(f', T)]$$

Si además se supone que $h_f(\cdot) \geq h_{f'}(\cdot)$, entonces la política f la cual tiene un sesgo mayor, eventualmente rebasará a f' en el sentido de que para cualquier $\varepsilon > 0$ dado

$$J_T(f) \geq J_T(f') - \varepsilon$$

para toda T suficientemente grande. En otras palabras, la maximización de la función de sesgo, dentro de la clase de políticas de ganancia óptima, permite obtener la política con mayor crecimiento. Al respecto, Escobedo-Trujillo, López-Barrientos y Hernández-Lerma (2012) tratan con juegos diferenciales estocásticos de suma cero con ganancias promedio en el largo plazo. Su principal objetivo es proporcionar las condiciones para la existencia y caracterización de equilibrios óptimos en sesgo y rebasantes. Primero caracterizan la familia de estrategias óptimas de ganancias promedio. Posteriormente, en esta familia, se imponen condiciones adecuadas para determinar las subfamilias de los equilibrios en sesgo y rebasantes. Un aspecto esencial para conseguir esto es demostrar la existencia de soluciones de las ecuaciones de optimalidad de ganancia promedio. Esto se hace mediante el enfoque usual del "descuento desvaneciente".

Por su parte, Álvarez-Mena y Hernández-Lerma (2006) consideran juegos estocásticos no cooperativos de N personas con los criterios de ganancias descontadas. El espacio de estados se supone que es numerable y los conjuntos de acción son espacios métricos compactos. Estos autores obtienen varios resultados importantes. El primero se refiere a la sensibilidad o la aproximación de juegos restringidos. El segundo muestra la existencia de equilibrios de Nash para juegos restringidos con un espacio de estados finito (y un espacio acciones compacto). El tercero extiende las condiciones para la existencia de una clase de juegos

restringidos que se pueden aproximar por juegos restringidos con un número finito de estados y espacios de acción compactos.

Otras contribuciones que son relevantes en juegos estocásticos son: Rincón-Zapatero (2004) y Rincón-Zapatero *et al.* (1998) y (2000) quienes caracterizan en juegos diferenciales equilibrios de Nash de subjuegos perfectos; Nowak (2003a) y (2003b), y Nowak y Szajowski (2003) y (2005) analizan equilibrios de Nash de juegos estocásticos de suma cero y no cero; y Neck (1985) y (1991) estudia juegos diferenciales entre la autoridad fiscal y el banco central.

4 Optimalidad en sesgo y optimalidad rebasante

Jasso-Fuentes y Hernández-Lerma (2008) proporcionan las condiciones para la existencia de políticas rebasantes óptimas para una clase general de procesos de difusión controlados. La caracterización es de tipo lexicográfico, es decir, en primer lugar se identifica la clase de las llamadas políticas canónicas y, posteriormente, dentro de esta clase se buscan políticas con alguna característica especial, por ejemplo, políticas canónicas que además maximizan el sesgo.[8]

Por otro lado, Escobedo-Trujillo y Hernández-Lerma (2011) estudian difusiones controladas moduladas con una cadena de Markov. Una difusión controlada modulada con una cadena de Markov es una ecuación diferencial estocástica de la forma

$$\mathrm{d}x(t) = b(x(t), \psi(t), u(t))\mathrm{d}t + \sigma(x(t), \psi(t))\mathrm{d}W_t,\ x(0) = 0,\ \psi(0) = i,$$

donde $\psi(t)$ es una cadena de Markov irreducible en tiempo continuo con un espacio de estados finito $S = \{1, 2, ..., N\}$ y probabilidades de transición

$$P\{\psi(s + \mathrm{d}t) = j \mid \psi(s) = i\} = q_{ij}\mathrm{d}t + o(\mathrm{d}t).$$

Para estados $i \neq j$ la cantidad q_{ij} es la tasa de transición de pasar de i a j, mientras que

$$q_{ii} = -\sum_{j \neq i} q_{ij}.$$

[8]La optimalidad rebasante fuerte es un concepto introducido inicialmente por Ramsey (1928). Una noción más débil se introdujo, de forma independiente, por Atsumi (1965) y von Weizsäcker (1965).

Estos autores proporcionan las condiciones para la existencia y la caracterización de políticas rebasantes óptimas. Para ello, primero, utilizan el hecho de que la ganancia promedio de la ecuación de Hamilton-Jacobi-Bellman asegura que la familia de las políticas de control canónicas es no vacío. Posteriormente, dentro de esta familia, se caracterizan las políticas canónicas que maximizan el sesgo y que son rebasantes óptimas. Otros resultados importantes sobre optimalidad en sesgo y optimalidad rebasante se encuentran en Prieto-Rumeau y Hernández-Lerma (2006) y (2009).

Asimismo, Prieto-Rumeau y Hernández-Lerma (2005) tratan con juegos markovianos de suma cero en tiempo continuo con un espacio de estados numerable, espacios de Borel arbitrarios de acciones y tasas de transición y de ganancia (o costo) posiblemente no acotadas. Analizan también la optimalidad en sesgo y los criterios de optimalidad rebasante.

5 Optimalidad de Blackwell para procesos de difusión controlados

Los criterios de optimalidad más comunes para problemas de control óptimo con horizonte infinito son los de utilidad descontada esperada y utilidad promedio esperada. Estos dos criterios tienen objetivos opuestos: el primero distingue el desempeño en el corto plazo, ya que se desvanece para intervalos de tiempo grandes, mientras que el segundo considera la conducta asintótica, ignorando simplemente lo que pasa en intervalos de tiempo finito. Como alternativa a estas dos situaciones extremas se consideran refinamientos del criterio de utilidad promedio tales como optimalidad rebasante, optimalidad en sesgo y los llamados criterios sensibles al descuento, los cuales incluyen optimalidad con m-descuentos para un entero $m \geq -1$ y optimalidad de Blackwell para $m = +\infty$. Se les llama "refinamientos" porque se refieren a políticas de control que optimizan la utilidad promedio y que, además, satisfacen alguna otra propiedad adicional. Al respecto, es importante resaltar que Jasso-Fuentes y Hernández-Lerma (2009) proporcionan algunos de estos refinamientos. Estos autores dan condiciones que garantizan la optimalidad con m-descuentos para cada entero $m \geq -1$ y también para la optimalidad de Blackwell cuando el sistema controlado es un proceso de difusión markoviano de la forma

$$\mathrm{d}x(t) = b(x(t), u(t))\mathrm{d}t + \sigma(x(t))\mathrm{d}B(t) \quad \text{para todo } t \geq 0 \text{ y } x(0) = x,$$

donde $b(\cdot,\cdot) : \mathbb{R}^n \times U \to \mathbb{R}^n$ y $\sigma(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times d}$ son funciones dadas que satisfacen un conjunto de condiciones estándar y $B(\cdot)$ es un movimiento browniano de dimensión d. El conjunto $U \subset \mathbb{R}^m$ es llamado el conjunto de control (o de acciones) y $u(\cdot)$ es un proceso estocástico con valores en el conjunto U, el cual representa la acción del controlador a cada tiempo $t \geq 0$.

6 Control óptimo con procesos markovianos de difusión

En esta sección se establece el problema general de control óptimo estocástico donde las restricciones son procesos markovianos de difusión y se formula la técnica de programación dinámica con la cual se obtiene la ecuación diferencial parcial no lineal de Hamilton-Jacobi-Bellman (HJB), cuya solución caracteriza el control óptimo y con ello las trayectorias de las variables que optimizan a la función objetivo[9]. De acuerdo con Hernández-Lerma (1994), el control óptimo estocástico es una técnica matemática utilizada para resolver problemas de optimización de sistemas dinámicos en ambientes de incertidumbre.

A continuación se establece el modelo matemático general del problema de control óptimo estocástico en tiempo continuo. Considere un sistema dinámico en tiempo continuo con un horizonte temporal finito, $[0, T]$. Se consideran, primero, funciones adecuadas $\mu(t, x, u)$ y $\sigma(t, x, u)$ que satisfacen

$$\mu : \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n,$$
$$\sigma : \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^{n+d}.$$

Para un punto $x_0 \in \mathbb{R}^n$ considere la ecuación diferencial estocástica

$$\mathrm{d}X_t = \mu(t, X_t, u_t)\mathrm{d}t + \sigma(t, X_t, u_t)\mathrm{d}W_t$$
$$X_0 = x_0,$$

en donde se considera al proceso n-dimensional X_t, como el proceso de variables de estado, que se requiere controlar, el proceso k-dimensional u_t como el proceso de control, cuya correcta elección controlará a X_t, y W_t es un proceso de Wiener d-dimensional, definido sobre un espacio fijo de probabilidad equipado con una filtración $(\Omega, \mathcal{F}, (\mathcal{F}_t^W)_{t\in[0,T]}, \mathbb{P})$.

[9]Para más detalles sobre el problema de control óptimo estocástico en tiempo continuo véase, por ejemplo, Björk, Myhrman y Persson (1987).

Se define a continuación una regla de control admisible. Para tal efecto se considera la clase de procesos de control admisible como procesos cuyo valor u_t en el tiempo t es adaptado al proceso de estado X_t, el cual se obtiene mediante una función $\mathbf{u}(t,x)$

$$\mathbf{u} : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^k,$$

de manera que

$$u_t = \mathbf{u}(t, X_t),$$

$\mathbf{u}$ así definida es llamada regla de control de retroalimentación o estrategia markoviana. Ahora se impone a $\mathbf{u}$ la restricción de que para cada $t, u_t \in U \subset \mathbb{R}^k$ donde U es la clase de controles admisibles. Una regla de control $\mathbf{u}(t,x)$ es admisible si:

1) $\mathbf{u}(t,x) \in U$, $\forall\, t \in \mathbb{R}_+$ y $\forall\, x \in \mathbb{R}^n$,

2) Para cualquier punto inicial (t,x) dado, la ecuación diferencial estocástica

$$\mathrm{d}X_s = \mu(s, X_s, \mathbf{u}(s, X_s))\mathrm{d}s + \sigma(s, X_s, \mathbf{u}(s, X_s))\mathrm{d}W_s$$

$$X_t = x$$

tiene una única solución.

Dado que el problema de control óptimo se establecerá en un esquema estocástico, y toda vez que el proceso de estados es n-dimensional, será necesario definir las siguientes funciones y establecer el teorema fundamental del cálculo estocástico, llamado el lema de Itô (para el caso de n variables). Para cualquier regla de control $\mathbf{u}$, las funciones $\mu^{\mathbf{u}}$ y $\sigma^{\mathbf{u}}$ son definidas por:

$$\mu^{\mathbf{u}}(t,x) = \mu(t, x, \mathbf{u}(t,x)),$$

$$\sigma^{\mathbf{u}}(t,x) = \sigma(t, x, \mathbf{u}(t,x)),$$

y se suponen con segundas derivadas continuas. El lema de Itô para n variables de estado se establece a continuación.

Considere la función $y = f(t, \mathbf{x})$, $\mathbf{x} = (x_1, x_2, ..., x_n)$ y las ecuaciones diferenciales estocásticas

$$\mathrm{d}x_i = \mu_i(x_i, t)\mathrm{d}t + \sigma_i(x_i, t)\mathrm{d}W_{it},\ \ i = 1, 2, ..., n,$$

y cualquier vector fijo $u_t \in \mathbb{R}^k$, entonces para cualquier regla de control $\mathbf{u}$, se tiene

$$\mathrm{d}y = \left[\frac{\partial f\left(t,\mathbf{x}\right)}{\partial t} + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 f\left(t,\mathbf{x}\right)}{\partial x_j \partial x_i} \sigma_i^{\mathbf{u}}\left(x_i,t\right)\sigma_j^{\mathbf{u}}\left(x_i,t\right)\rho_{ij} + \sum_{i=1}^{n} \frac{\partial f\left(t,\mathbf{x}\right)}{\partial x_i}\mu_i^{\mathbf{u}}\left(x_i,t\right)\right]\mathrm{d}t + \sum_{i=1}^{n} \frac{\partial f\left(t,\mathbf{x}\right)}{\partial x_i}\sigma_i^{\mathbf{u}}\left(x_i,t\right)\mathrm{d}W_{it},$$

donde ρ_{ij} es el coeficiente de correlación entre $\mathrm{d}W_{j,t}$ y $\mathrm{d}W_{it}$, de tal forma que $\rho_{ij}\mathrm{d}t = \mathrm{Cov}(\mathrm{d}W_{it}, \mathrm{d}W_{jt})$. Ahora bien, dada una regla de control $\mathbf{u}_t = \mathbf{u}(t, X_t^{\mathbf{u}})$ con su correspondiente proceso controlado $X^{\mathbf{u}}$ se utilizará la notación

$$\mathrm{d}X_t^{\mathbf{u}} = \mu^{\mathbf{u}}\mathrm{d}t + \sigma^{\mathbf{u}}\mathrm{d}W_t.$$

Para definir la función objetivo del problema de control se consideran las funciones:

$$F : \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \ \text{ dada por } \ (t, X_t^{\mathbf{u}}, \mathbf{u}_t) \to F(t, X_t^{\mathbf{u}}, \mathbf{u}_t)$$

y

$$\Phi : \mathbb{R}^n \to \mathbb{R} \ \text{ dada por } \ X_t^{\mathbf{u}} \to \Phi(X_t^{\mathbf{u}}),$$

donde F evalúa el desempeño del sistema a través del tiempo y Φ evalúa el final. Se supone que tanto F como Φ son de clase C^2. Se define la funcional objetivo del problema de control como $J_0 : U \to \mathbb{R}$,

$$J_0(\mathbf{u}) = E\left[\int_0^T F(t, X_t^{\mathbf{u}}, \mathbf{u}_t)\mathrm{d}t + \Phi(X_T^{\mathbf{u}}) \mid \mathcal{F}_0\right],$$

donde $\mathcal{F}_0$ representa la información disponible al tiempo $t = 0$. El problema de control puede ser escrito como uno de maximización de la funcional $J_0(\mathbf{u})$, sobre $\mathbf{u} \in U$. Se define la funcional óptima por $\widehat{J}_0 = \max_{\mathbf{u}\in U} J_0(\mathbf{u})$. Si existe un control admisible $\widehat{\mathbf{u}}$ tal que $\widehat{J}_0 = J_0(\widehat{\mathbf{u}})$ se dice entonces que $\widehat{\mathbf{u}}$ es un control óptimo para el problema dado. Si se supone una pareja (t, x) fija donde $t \in [0, T]$ y $x \in \mathbb{R}^n$, el problema de control se puede definir como:

$$\max_{\mathbf{u}_s} E\left[\int_t^T f(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}t + \Phi(X_T^{\mathbf{u}}) \mid \mathcal{F}_t\right]$$

sujeto a

$$\mathrm{d}X_s^{\mathbf{u}} = \mu(s, X_s^{\mathbf{u}}, \mathbf{u}(s, X_s^{\mathbf{u}}))\mathrm{d}s + \sigma(s, X_s^{\mathbf{u}}, \mathbf{u}(s, X_s^{\mathbf{u}}))\mathrm{d}W_s, \quad X_t = x,$$

y a la restricción

$$\mathbf{u}(s,y) \in U \quad \text{para todo } (s,y) \in [t,T] \times \mathbb{R}^n.$$

La función de valor $J : \mathbb{R}_+ \times \mathbb{R}^n \times U \to \mathbb{R}$ es definida mediante

$$J(t,x,\mathbf{u}) = E\left[\int_t^T F(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}s + \Phi(X_T^{\mathbf{u}}) \mid \mathcal{F}_t\right].$$

La función de valor óptimo es $\widehat{J} : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ y está definida por

$$\widehat{J}(t,x) = \max_{\mathbf{u}\in U} J(t,x,\mathbf{u}).$$

El objetivo, ahora, es caracterizar la función de valor en el control óptimo mediante una ecuación diferencial parcial, mejor conocida como la ecuación diferencial parcial de HJB[10]. Es importante destacar que la derivación que, a continuación, se hace de la ecuación de HJB es informal, pero ilustrativa. Suponga que:

1) Existe una regla de control óptimo,

2) La función de valor óptimo $\widehat{J}$ es de clase C^2.

Considere el par $(t,x) \in (0,T) \times \mathbb{R}^n$ fijo pero arbitrario, y suponga un incremento muy pequeño, de hecho diferencial, $\mathrm{d}t \in \mathbb{R}$ tal que $t < t + \mathrm{d}t < T$. También se considera una regla de control $\mathbf{u}$ fija pero arbitraria. Por lo tanto, dada la definición de la función de valor óptimo y el incremento $\mathrm{d}t$, se tiene la relación recursiva temporal

$$\begin{aligned}
\widehat{J}(t,x) &= \max_{\mathbf{u}\in U} J(t,x,\mathbf{u}) = \max_{\mathbf{u}\in U} E\left[\int_t^T F(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}s + \Phi(X_T^{\mathbf{u}}) \mid \mathcal{F}_t\right] \\
&= \max_{\mathbf{u}\in U} E\left[\int_t^{t+\mathrm{d}t} F(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}s + \int_{t+\mathrm{d}t}^T F(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}s + \Phi(X_T^{\mathbf{u}}) \mid \mathcal{F}_t\right] \\
&= \max_{\mathbf{u}\in U} E\left[\int_t^{t+\mathrm{d}t} F(s, X_s^{\mathbf{u}}, \mathbf{u}_s)\mathrm{d}s + \widehat{J}(t+\mathrm{d}t, X_t^{\mathbf{u}} + \mathrm{d}X_t^{\mathbf{u}}) \mid \mathcal{F}_t\right].
\end{aligned}$$

En esta expresión se aplica al primer sumando el teorema del valor medio de cálculo integral y al segundo una expansión en serie de Taylor, de lo cual resulta

$$\begin{aligned}
\widehat{J}(t, X_t^{\mathbf{u}}) = \max_{\mathbf{u}\in U} E\Big[&F(t, X_t^{\mathbf{u}}, \mathbf{u}_t)\mathrm{d}t + o(\mathrm{d}t) + \widehat{J}(t, X_t^{\mathbf{u}}) \\
&+ \mathrm{d}\widehat{J}(t, X_t^{\mathbf{u}}) + o(\mathrm{d}t) \mid \mathcal{F}_t\Big].
\end{aligned}$$

[10]La ecuación HJB es el resultado central en la teoría de control óptimo. La ecuación correspondiente en tiempo discreto se conoce como la ecuación de Bellman.

Después de simplificar, se tiene

$$0 = \max_{\mathbf{u}\in U} E\left[F(t, X_t^{\mathbf{u}}, \mathbf{u}_t)\mathrm{d}t + o(\mathrm{d}t) + \mathrm{d}\widehat{J}(t, X_t^{\mathbf{u}}) \mid \mathcal{F}_t\right].$$

En la expresión anterior se aplica el lema de Itô para obtener la diferencial estocástica de $\widehat{J}$, así

$$\begin{aligned}
0 = \max_{\mathbf{u}\in U} E\Bigg\{ & F\left(t, X_t^{\mathbf{u}}, \mathbf{u}_t\right)\mathrm{d}t + o\left(\mathrm{d}t\right) \\
& + \left[\frac{\partial \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial t} + \sum_{i=1}^{n} \frac{\partial \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial x_i}\mu_i^{\mathbf{u}}\left(x_i, t\right)\right. \\
& \left. + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial x_j \partial x_i}\sigma_i^{\mathbf{u}}\left(x_i, t\right)\sigma_j^{\mathbf{u}}\left(x_i, t\right)\rho_{ij}\right]\mathrm{d}t \\
& + \sum_{i=1}^{n} \frac{\partial \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial x_i}\sigma_i^{\mathbf{u}}\left(x_i, t\right)\mathrm{d}W_{it} \,|\mathcal{F}_t \Bigg\},
\end{aligned}$$

donde, como antes, ρ_{ij} satisface $\rho_{ij}\mathrm{d}t = \mathrm{Cov}(\mathrm{d}W_{it}, \mathrm{d}W_{jt})$. Después de tomar valores esperados a los términos aleatorios de la ecuación anterior y dado que $\mathrm{d}W_{it} \sim N(0, \mathrm{d}t)$, se obtiene

$$\begin{aligned}
0 = \max_{\mathbf{u}\in U} \Bigg[& F\left(t, X_t^{\mathbf{u}}, \mathbf{u}\right) + \frac{\partial \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial t} + \sum_{i=1}^{n} \frac{\partial \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial x_i}\mu_i^{\mathbf{u}}\left(x_i, t\right) \\
& + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 \widehat{J}\left(t, X_t^{\mathbf{u}}\right)}{\partial x_j \partial x_i}\sigma_i^{\mathbf{u}}\left(x_i, t\right)\sigma_j^{\mathbf{u}}\left(x_i, t\right)\rho_{ij}\Bigg].
\end{aligned}$$

Toda vez que el análisis ha sido realizado sobre un punto fijo pero arbitrario, entonces la ecuación es válida para todo $(t, x) \in (0, T) \times \mathbb{R}^n$, de tal forma que:

1) $\widehat{J}$ satisface la ecuación de HJB:

$$\begin{aligned}
0 = \max_{\mathbf{u}\in U} \Bigg[& F(t, X_t^u, u) + \frac{\partial \widehat{J}(x, X_t^u)}{\partial t} + \sum_{j=1}^{n} \frac{\partial \widehat{J}(t, X_t^u)}{\partial x_i}\mu_i^u(x_i, t) \\
& + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 \widehat{J}(t, X_t^u)}{\partial x_j \partial x_i}\sigma_i^u(x_i, t)\sigma_j^u(x_i, t)\rho_{ij}\Bigg]
\end{aligned}$$

$\forall (t, x) \in (0, T) \times \mathbb{R}^n$, $\widehat{J}(T, x) = \Phi(x)\ \forall\ x \in \mathbb{R}^n$.

2) Para cada $(t,x) \in (0,T) \times \mathbb{R}^n$, el máximo en la ecuación HJB es alcanzado por $u = \mathbf{u}(t,x)$.

A partir de la ecuación HJB se sigue que $\mathbf{u}$ es única ya que x y t son fijos y las funciones $F, \widehat{J}, \mu_i^u, \sigma_i^u$ y σ_j^u son consideradas como dadas. Si se supone que $u \in U$ es óptimo, entonces se obtiene la siguiente ecuación diferencial parcial de segundo orden en $\widehat{J}$,

$$0 = F(t, X_t^u, u) + \frac{\partial \widehat{J}(t, X_t^u)}{\partial t} + \sum_{i=1}^{n} \frac{\partial \widehat{J}(t, X_t^u)}{\partial x_i} \mu_i^u(x_i, t) + \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 \widehat{J}(t, X_t^u)}{\partial x_j \partial x_i} \sigma_i^u(x_i, t) \sigma_j^u(x_i, t) \rho_{ij}.$$

Al derivar esta ecuación con respecto de la variable de control, u, se tiene la siguiente condición de primer orden (condición necesaria):

$$0 = \frac{\partial F(t, X_t^u, u)}{\partial u} + \frac{\partial^2 \widehat{J}(t, X_t^u)}{\partial u \partial t} + \frac{\partial}{\partial u} \left(\sum_{i=1}^{n} \frac{\partial \widehat{J}(t, X_t^u)}{\partial x_i} \mu_i^u(x_i, t) \right) + \frac{\partial}{\partial u} \left(\frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 \widehat{J}(t, X_t^u)}{\partial x_j \partial x_i} \sigma_i^u(x_i, t) \sigma_j^u(x_i, t) \rho_{ij} \right).$$

La ecuación anterior caracteriza al control óptimo u en función de x y t y $\widehat{J}$; es decir $\widehat{u} = \widehat{u}(t, x, \widehat{J})$. Para resolver la ecuación anterior y encontrar la trayectoria óptima del control, se procede a utilizar el método de funciones en variables separables, aunque es necesario recordar que, en general, es difícil obtener una solución explicita de la ecuación HBJ. Sin embargo, en diversas aplicaciones en las ciencias naturales y sociales la ecuación de HJB tiene una solución analítica; véanse, por ejemplo, Merton (1990) y Hakansson (1970).

7 Conclusiones

La forma en que los agentes definen su actuar requiere de un proceso de abstracción en el que el individuo escoge y organiza sus acciones, en su propio beneficio, de acuerdo con un criterio preestablecido, elaborando con ello un plan para anticipar posibles efectos no deseados. En esta investigación se ha realizado una revisión de las contribuciones de Onésimo Hernández-Lerma a la teoría y práctica de los procesos markovianos. Se resaltan los avances recientes de Onésimo Hernández-Lerma

que han impulsado el potencial y las bondades técnicas de los procesos markovianos en el modelado de los procesos de toma de decisiones de agentes racionales incorporando dinámicas más realistas a diversas variables (económicas y financieras) de interés. Particularmente, se destacan las extensiones y reformulaciones de Onésimo Hernández-Lerma en los procesos markovianos de decisión, los juegos estocásticos, la optimalidad de Blackwell para procesos de difusión controlados y el control óptimo estocástico donde las restricciones son procesos markovianos de difusión.

Varios temas de gran potencial para la investigación en sistemas controlados con procesos markovianos has sido ampliamente estudiados por Hernández-Lerma, entre ellos destacan los refinamientos de los criterios de utilidad promedio tales como optimalidad rebasante, optimalidad en sesgo y los llamados criterios sensibles al descuento, los cuales incluyen la optimalidad de Blackwell. En este sentido, se destacan los trabajos de Hernández-Lerma y varios coautores sobre las condiciones para la existencia y caracterización de equilibrios óptimos en sesgo y rebase, sobre todo en lo que se refiere a la caracterización de estrategias óptimas de ganancia promedio.

Francisco Venegas-Martínez
Escuela Superior de Economía
Instituto Politécnico Nacional
Plan de Agua Prieta, No. 66
Col. Plutarco Elías Calles
Del. Miguel Hidalgo
México, D. F. 11340
México
fvenegas1111@yahoo.com.mx

Referencias

[1] Álvarez-Mena J.; Hernández-Lerma O., *Existence of Nash equilibria for constrained stochastic games*, Math. Methods Oper. Res. **63** (2006), 261–285.

[2] Atsumi H. *Neoclassical growth and the efficient program of capital accumulation*, Rev. Econ. Stud. **32** (1965), 127–136.

[3] Björk T.; Myhrman J.; Persson M., *Optimal consumption with stochastic prices in continuous time*, J. Appl. Probab. **24** No. 1 (1987), 35–47.

[4] Cerra V.; Saxena S. C., *Did output recover from the Asian crisis?*, IMF Staff Papers **52** (2005), 1–23.

[5] Escobedo-Trujillo B. A.; López-Barrientos J. D.; Hernández-Lerma O., *Bias and overtaking equilibria for zero-sum stochastic differential games*, J. Optim. Theory Appl. **153** No. 3 (2012), 662–687.

[6] Escobedo-Trujillo B. A.; Hernández-Lerma O., *Overtaking optimality for controlled Markov-modulated diffusion*, Optimization (2011), 1–22.

[7] Feinberg E. A., *Controlled Markov processes with arbitrary numerical criteria*, Theor. Probability Appl. **27** (1982), 486–503.

[8] Federgruen A.; Schweitzer P. J., *Nonstationary Markov decision problems with converging parameters*, J. Optim. Theory Appl. **34** (1981), 207–241.

[9] Goldfeld S. M.; Quandt R. E., *A Markov model for switching regressions*, J. Econom. **1** (1973), 3–16.

[10] Guo H.; Hsu W., *A survey of algorithms for real-time bayesian network inference*, Joint Workshop on Real-Time Decision Support and Diagnosis Systems, Edmonton, Canada (2002).

[11] Guo X. P.; Hernández-Lerma O., *Continuous-time controlled Markov chains*, Ann. Appl. Probab., **13** (2003a), 363–388.

[12] Guo X. P.; Hernández-Lerma O., *Drift and monotonicity conditions for continuous-time Markov control processes with an average criterion*, IEEE Trans. Automat. Control, **48** (2003b), 236–245.

[13] Guo X. P.; Hernández-Lerma O., *Continuous-time controlled Markov chains with discounted rewards*, Acta Appl. Math., **79** (2003c), 195–216.

[14] Guo X. P.; Hernández-Lerma O., *Zero-sum games for continuous-time Markov chains with unbounded transition and average payoff rates*, J. Appl. Probab., **40** No. 2 (2003d), 327–345.

[15] Guo X. P.; Hernández-Lerma O., *Zero-sum continuous-time Markov games with unbounded transition and discounted payoff rates*, Bernoulli **11** No. 6 (2005a), 1009–1029.

[16] Guo X. P.; Hernández-Lerma O., *Nonzero-sum games for continuous-time Markov chains with unbounded discounted payoffs*, J. Appl. Probab. **42** No. 2 (2005b), 303–320.

[17] Guo X. P.; Hernández-Lerma O., *Zero-sum games for continuous-time jump Markov processes in polish spaces: discounted payoffs*, Adv. in Appl. Probab. **39** No. 3 (2007), 645–668.

[18] Guo X.P.; Hernández-Lerma O., Continuous-Time Markov Decision Processes: Theory and Applications, Springer-Verlag, New York, 2009.

[19] Hakansson N., *Optimal investment and consumption strategies under risk for a class of utility functions*, Econometrica, **38** No. 5 (1970), 587–607.

[20] Hernández-Lerma O., Control óptimo y juegos estocásticos, Escuela de Matemática de América Latina y del Caribe, CIMAT, Guanajuato, México (2005).

[21] Hernández-Lerma O., Lectures on Continuous-Time Markov Control Processes, Aportaciones Matemáticas 3, Sociedad Matemática Mexicana, Mexico City (1994).

[22] Hernández-Lerma O., Lecture Notes on Discrete-Time Markov Control Processes, Departamento de Matemáticas, CINVESTAV-IPN, 1990.

[23] Hernández-Lerma O., Adaptive Markov Control Processes, Springer-Verlag, New York, 1989.

[24] Hernández-Lerma O., *Finite-state approximations for denumerable multidimensional state discounted Markov decision processes*, J. Math. Anal. Appl., **113** No. 2 (1986), 382–389.

[25] Hernández-Lerma O., *Nonstationary value-iteration and adaptive control of discounted semi-Markov processes*, J. Math. Anal. Appl., **112** (1985), 435–445.

[26] Hernández-Lerma O.; Marcus S. I., *Optimal adaptive control of priority assignment in queueing systems*, Systems Control Lett., **4** (1984), 65–72.

[27] Hernández-Lerma O.; Marcus S. I., *Adaptive control of discounted Markov decision chains*, J. Optim. Theory Appl., **46** (1985), 227–235.

[28] Hernández-Lerma O.; Marcus S. I., *Adaptive policies for discrete-time stochastic systems with unknown disturbance distribution*, Systems Control Lett., **9** (1987), 307–315.

[29] Hernández-Lerma O.; Marcus S. I., *Nonparametric adaptive control of discrete-time partially observable stochastic systems*, J. Math. Anal. Appl., **137** No. 2 (1989), 312–334.

[30] Hernández-Lerma O.; Govindan T. E., *Nonstationary continuous-time Markov control processes with discounted costs on infinite horizon*, Acta Appl. Math., **67** (2001), 277–293.

[31] Hernández-Lerma O.; Lasserre J. B., Discrete-Time Markov Control Processes, Springer-Verlag, New York, 1996.

[32] Hernández-Lerma O.; Lasserre J. B., Further Topics on Discrete-Time Markov Control Processes, Springer-Verlag, New York, 1999.

[33] Hernández-Lerma O.; Lasserre J. B., *Zero-sum stochastic games in Borel spaces: average payoff criterion*, SIAM J. Control Optimization, **39** (2001a), 1520–1539.

[34] Hernández-Lerma O.; Lasserre J. B., *Further criteria for positive Harris recurrence of Markov chains*, Proc. Amer. Math. Soc., **129** No. 5 (2001b), 1521–1524.

[35] Hernández-Lerma O.; Lasserre J. B., Markov Chains and Invariant Probabilities, Birkhäuser, Basel, 2003.

[36] Jasso-Fuentes H.; Hernández-Lerma O., *Ergodic control, bias and sensitive discount optimality for Markov diffusion processes*, Stoch. Anal. Appl., **27** (2007), 363–385.

[37] Jasso-Fuentes H.; Hernández-Lerma O., *Characterizations of overtaking optimality for controlled diffusion processes*, Appl. Math. Optim., **57** (2008), 349–369.

[38] Jasso-Fuentes H.; Hernández-Lerma O., *Blackwell optimality for controlled diffusion processes*, J. Appl. Probab., **46** No. 2 (2009), 372–391.

[39] Merton R. C., *Continuous-time finance*, Rev. Econom. Statist., **51** No. 2 (1992), 247–257.

[40] Merton R. C., Continuous-Time Finance, Basil Blackwell, Cambridge, Massachusetts, 1990.

[41] Neck R., *A differential fame model of fiscal and monetary policies: conflict and cooperation*, Optimal control theory and economic analysis, Second Viennese Workshop on Economic Applications of Control Theory, Vienna (1984), **2** 1985, 607–632.

[42] Neck R., *Non-cooperative equilibrium solution for a stochastic dynamic game of economic stabilization policies*, Dynamic Games in Economic Analysis, Lecture Notes in Control and Information Sciences, **157** (1991).

[43] Nowak A. S. (2003a). *Zero-sum stochastic games with Borel state spaces*, Proceedings of the NATO Advanced Study Institute, Stony Brook, New York (1999), **570**, 77–91.

[44] Nowak A. S. (2003b). *On a new class of nonzero-sum discounted stochastic games having stationary Nash equilibrium points*, Int. J. Game Theory, **32**, 121–132.

[45] Nowak A. S.; Szajowski P.(2003). *On Nash equilibria in stochastic games of capital accumulation*, Stochastic Games and Applications, **9**, 118–129.

[46] Nowak A. S.; Szajowski K., Advances in Dynamic Games. Annals of the International Society of Dynamic Games Vol. 7 , Birkhauser, Boston, 2005.

[47] Prieto-Rumeau T.; Hernández-Lerma O. *Bias and overtaking equilibria for zero-sum continuous time Markov games*, Math. Meth. Oper. Res., **61** (2005), 437–454.

[48] Prieto-Rumeau T.; Hernández-Lerma O. *Bias optimality for continuous-time controlled Markov chains*, SIAM J. Control Optim., **45** (2006), 51–73.

[49] Prieto-Rumeau T.; Hernández-Lerma O. *Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains*, Math. Meth. Oper. Res., **70** (2009), 527–240.

[50] Prieto-Rumeau T.; Hernández-Lerma O. *Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games*, ICP Advanced Texts in Mathematics, Vol. 5, World Scientific, 2012.

[51] Ramsey F. P. *A mathematical theory of savings*, Economic Journal, **38** (1928), 543–559.

[52] Rincón-Zapatero J. P., *Characterization of Markovian equilibria in a class of differential games*, J. Econ. Dyn. Control, **28** (2004), 1243–1266.

[53] Rincón-Zapatero J. P.; Martínez J.; Martín-Herrán G. *New method to characterize subgame perfect Nash equilibria in differential games*, J. Optim. Theory Appl., **96** (1998), 377–395.

[54] Rincón-Zapatero J. P.; Martínez J.; Martín-Herrán G. *Identification of efficient subgame-perfect Nash equilibria in a class of differential games*, J. Optim. Theory Appl., **104** (2000), 235–242.

[55] Shapley L. S. (1953). *A Value for n-person Games*, In: *Contributions to the Theory of Games*, volume II, H. W. Kuhn and A.W. Tucker (eds.).

[56] Schäl, M. *Estimation and control in discounted stochastic dynamic programming*, Stochastics, **20** (1987), 51–71.

[57] Taylor S. J., Asset Price Dynamics, Volatility, and Prediction, Princeton University Press, Princeton, 2005.

[58] Tierney L., *Markov chains for exploring posterior distributions*, Ann. Statist., **22** No. 4 (1994), 1701–1728.

[59] von Weizsäcker C. C. *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econ. Stud., **32** (1965), 85–104.

[60] White D. J., in: *Recent Developments in Markov Decision Processes* (ed. R. Hartley, L. C. Thomas and D. J. White), Academic Press, New York (1980).