Universidad Autónoma Metropolitana

UAM Iztapalapa

JUEGOS ESTOCÁSTICOS TRANSITORIOS CON RECOMPENSA TOTAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE DOCTOR

EN CIENCIAS MATEMÁTICAS

PRESENTA:

VÍCTOR MANUEL MARTÍNEZ CORTÉS

ASESORES DE TESIS:

DR. RAÚL MONTES DE OCA

DR. KAREL SLADKÝ

Índice general

Agradecimientos

Hay personas que nos acompañan en la vida que nos inspiran que nos dan su confianza que nos dan su cariño y sin embargo, no pueden estar con nosotros en el momento donde se obtienen los éxitos. Quiero darle un especial agradecimiento a mi maestro espiritual Guru Dev Singh Khalsa que este año partió pero que me dejo enseñanzas atemporales; a tu memoria.

Paso a paso se logran muchas consas.

En esta hoja muchas personas deberían estar aquí, sin embargo me quedaré con lo mas allegados, primero gracias a Dios por permitirme llegar a este momento, después a mis padres Angel y Rosario, a mi hermana, a mi pareja Masha y a mi familia tíos, primos, sobrinos etc, que siempre han sabido estar ahí para ayudarme y darme el apoyo necesario.

Pero que quiero hacer una mención especial a mis asesores el Dr. Karel Sladký y al Dr. Raúl Montes de Oca por su apoyo y guía durante todo este proceso. No ha sido sólo un día sino han sido años en los que se tuvieron que hacer cambios, mejoras donde tuve que vivir muchas experiencias para llegar aquí. Es importante también agradecer a los referis anónimos por su ayuda y apoyo en la consecusión del artículo.

Un agradecimiento muy especial a mis amigos Alberto, Victoria, Carol, Marlene, Elisandro, Erika (Harliv), Karime, Lenca, Fabiola, Tito, Livtar, AmritPal, Cony, Fernanda, y muchos otros amigos que me brindaron un apoyo incondicional, porque lograr esto no sólo tiene una parte académica sino también una parte emocional.

Por último me gustaría agredecerle a la UAM-Iztapalapa y al CONACYT por los apoyos brindados como lo fueron la beca y los apoyos para ir al extranjero.

Resumen

Este trabajo está relacionado con dos variantes de juegos: los juegos estocásticos sensibles al riesgo y los juegos estocásticos transitorios con tiempos de paro. La idea principal del trabajo consiste en encontrar condiciones para obtener las soluciones o equilibrios de Nash de estos juegos. Una primera parte del trabajo se enfocó en determinar condiciones para establecer las soluciones y sus aproximaciones, así como en construir algoritmos para generar estas aproximaciones y en dar estimaciones del error que se comete al emplear éstas. La segunda parte del trabajo se centró en determinar condiciones que nos permiten asegurar que el modelo utilizado para el caso de juegos estocásticos con tiempos de paro, en el cual el primer jugador tiene una función objetivo dada por la recompensa total, tiene solución. Para esto fue necesario imponer condiciones de transitoriedad sobre los juegos estocásticos considerados. Es importante observar que el modo en el que se obtuvo la solución fue mediante el teorema del punto fijo de Banach, a través de un operador de contracción adecuado. Este modo de trabajo, nos permitió no solo dar la solución y su caracterización sino además, nos permitió proponer una selección en caso de múltiples equilibrios respecto al enfoque del primer jugador.

Introducción

Un juego es una lista que consiste de tres elementos. El primer elemento son los jugadores. El segundo elemento son las acciones o estrategias que podrán usar los jugadores, dependiendo del juego y el tercer elemento es una función de recompensa para los participantes [7, 31] y [72]. En resumen, un juego es una situación en la cual los jugadores pueden elegir una estrategia para poder participar en el juegos; la idea, para cada uno de los participantes, es encontrar la acción que les da una recompensa máxima. Cuando se puede hablar de una repetición del juego, los jugadores pueden usar una combinación entre sus estrategias para alcanzar su máxima recompensa. Esta combinación de estrategias son conocidas como "estrategias mixtas", mientras las otras se les conoce como "estrategias puras". El objetivo principal de los juegos es encontrar la estrategia (si ésta existe) que le da a cada jugador la mejor recompensa. Para un tratado formal sobre la definición de un juego se pueden consultar las siguientes referencias [7, 30, 31, 65] y [72].

El objetivo principal de un juego, como se mencionó antes, es encontrar la estrategia que le da a cada jugador la máxima recompensa. Este concepto es conocido como "Equilibrio de Nash" [40], que en un sentido heurístico es una medida del comportamiento de los participantes. Este equilibrio depende del tipo de estrategias que los jugadores usarán durante el juego. Se puede mostrar que hay algunos juegos para los cuales no existe el Equilibrio de Nash en estrategias puras, pero existe el equilibrio en estrategias mixtas. (Uno de los juegos más famosos para ilustrar este caso es el juego de tirar un volado [72]). Sin embargo, es bien conocido en la literatura que para el caso de los juegos finitos en estrategias mixtas el Equilibrio de Nash existe [7, 18, 31, 30] y [72].

En la siguiente sección, un tipo especial de juegos será mencionado y estos juegos serán nuestros antecedentes para el trabajo a realizar [31, 66]. Más aún, para nuestro trabajo nosotros nos enfocaremos en juegos bi-personales (solamente dos jugadores) con estrategias finitas los cuales serán realizados a través de tiempos discretos, en una serie de distintos estados de manera secuencial, i.e. un jugador toma una decisión y entonces el siguiente jugador tiene la opción de seleccionar su

estrategia y así sucesivamente. Siguiendo esta regla también se le puede asociar una función de recompensa, así mismo el objetivo será encontrar una estrategia que le permita a cada uno de los jugadores obtener la mayor recompensa posible o en lenguaje de teoría de juegos encontrar el Equilibrio de Nash.

En nuestro caso, el primer jugador trabajará con la recompensa total y el segundo jugador buscará detener el juego en el mejor momento, esto nos conducirá a trabajar con los Juegos Estocásticos Transitorios con Tiempos de Paro. El nombre anterior resumirá el objetivo del trabajo y la clase de juegos con la que trabajaremos en esta tesis. Más aún, la técnica que se utilizó fue buscar un operador contractivo en un espacio de Banach, para así asegurar la existencia de un punto fijo [16], lo cual nos permitió garantizar la solución de los juegos propuestos.

Ahora nos enfocaremos en describir el orden de este trabajo. En la primera sección se hablará de los Juegos Markovianos [41], para después saltar a los Juegos Estocásticos y así poder trabajar con los Procesos de Decisión de Markov [7, 22]. Esto nos conllevará a una clase más especial que serán los Juegos Estocásticos Transitorios y mostrará la necesidad de la convergencia de la serie asociada a la recompensa total; para el final lograremos ir a la clase de los Juegos Estocásticos con Tiempos de Paro. En esta clase de juegos se mostrarán las condiciones para la existencia del Equilibrio de Nash. También se incluyeron tres apéndices los cuales redondean el trabajo. El primer apéndice es la formalización del Proceso Estocástico adjunto a los Juegos Estocásticos. El segundo apéndice es la base de la técnica que se utiliza para la generación del algoritmo en los Juegos Estocásticos Sensibles al Riesgo y el último apéndice resume las ideas básicas respecto al Tiempo de Paro.

Notación y Conceptos Básicos

Un juego simultáneo y estático es una tripleta [7, 31, 72] (A_1, A_2, R) donde A_1 denota el espacio de estrategias para el jugador 1, A_2 representa el espacio de estrategias para el jugador 2 y R es una función real que depende de A_1 y A_2. Como representación de lo anterior el jugador 1 elige una estrategia a_1 de A_1 y el jugador 2 elige una estrategia a_2 de A_2. Para el par (a_1, a_2) el pago para el jugador 1 es $R(a_1, a_2)$ y el pago para el jugador 2 es -$R(a_1, a_2)$. Supóngase que el jugador 2 anuncia al jugador 1 que estrategia va a utilizar, digamos $a_2(0)$ entonces el jugador 1 buscará maximizar su pago entonces de manera natural elegirá su estrategia $a_1(0)$ tal que $R(a_1(0), a_2(0)) = \text{máx}_{a_1} R(a_1, a_2(0))$. La mejor respuesta para el jugador 2 respecto a estas circunstancias será anunciar la estrategia $a_2(0)$ tal que

$$\max_{a_1} R(a_1(0), a_2(0)) = \min_{a_2} \max_{a_1} R(a_1, a_2) = \overline{v},$$

donde $\overline{v}$ (el valor máximo) como lo máximo que el jugador 1 puede obtener si el jugador 2 usa la estrategia $a_2(0)$. Supóngase que los roles son invertidos y el jugador 1 tiene que anunciar su estrategia $a_1(0)$. Como el jugador 2 esta seguro de usar $a_2(0)$ tal que

$$R(a_1(0), a_2(0)) = \min_{a_2} R(a_1(0), a_2),$$

entonces para el jugador 1 la mejor manera de protegerse es seleccionar $a_1(0)$ tal que

$$\min_{a_2} R(a_1(0), a_2) = \max_{a_1} \min_{a_2} R(a_1, a_2) = \underline{v},$$

donde $\underline{v}$ (el valor mínimo) que puede ser interpretado como el valor mínimo que puede obtener el jugador 1 independientemente de las selecciones del jugador 2. Siguiendo esta línea de argumentos.

Entonces es razonable llamar a esta valor común v el valor del juego para el jugador 1 y análogamente $-v$ es el valor del juego para el jugador 2.

Todas las estrategias $a_1(0)$ y $a_2(0)$ tal que $R(a_1(0), a_2) \geq v$ para todo a_2 y $R(a_1, a_2(0)) \leq v$ para todo a_1 serán conocidas como estrategias óptimas para el jugador 1 y 2 respectivamente.

Si existe $a_1(0) \in A_1$, $a_2(0) \in A_2$ y un número real v tal que $R(a_1(0), a_2(0)) \geq v$ para todo $a_2 \in A_2$ y $R(a_1, a_2(0)) \leq v$ para todo $a_1 \in A_1$,

$$\overline{v} = \min_{a^2} \max_{a^1} R(a^1, a^2) = v = \max_{a^1} \min_{a^2} R(a^1, a^2) = \underline{v}$$

y a la inversa.

Capítulo 1

Juegos Markovianos

Definición 1.0.1. *Un **Juego Estocástico o un Juego Markoviano** es una sexteta* $(X, N, A, \{A_i(x) : x \in X\}_{j=1,\ldots,n}, P, R)$ *la cual consiste de:*

- X *un conjunto no vacío y finito mejor conocido como el **espacio de estados**; donde "i" es un elemento arbitrario de* X.
- N *un conjunto finito y no vacío que representa el número de jugadores.* $n = \#(N)$.
- A *es un conjunto finito y no vacío conocido como el **espacio de acciones** o el **espacio de controles**. Sea* $A = A_1 \times \ldots \times A_n$.
- *Sea* $A_j(i)$ *un subconjunto* A_j *que representa las acciones disponibles para el jugador* j *respecto al estado* i. *Nótese que* $\{(i, a) : i \in X, a \in A_j(i)\}$ *es un subconjunto de* $X \times A_j$. *Sea* $A(i) = A_1(i) \times A_2(i) \times \ldots \times A_m(i), i \in X$. *Un **perfil*** $\sigma(i) = (a_1(i), a_2(i), \ldots, a_n(i))$ *consiste de las elecciones de cada jugador con respecto al estado* i.
- $P : X \times A \times X \to [0, 1]$ *representa la probabilidad de transición;* $P(\{\hat{j}\}|\sigma(i), i)$ *se puede interpretar como la probabilidad de que el siguiente estado sea* $\hat{j}$ *dado que se encuentra en el estado* i *y se utiliza el perfil* $\sigma(i)$.
- $R = (r_1, \ldots, r_n)$, *en la cual* $r_j : X \times A \to \mathbb{R}$ *es una función real valuada que se considera la **función de recompensa** con respecto al jugador* j.

Un juego estocástico está dado por un conjunto de estados X en donde el juego se va a jugar para cada estado i, en donde el jugador j puede elegir una acción de sus espacio de acciones $A_j(i)$ con respecto al estado i. Cada uno de los juegos se realizan a tiempo discreto $t = 0, 1, 2, \ldots$. La elección tomada por el conjunto de jugadores en cada estado determina inmediatamente una recompensa $r_j(i, \sigma(i))$ y una probabilidad $P(\{j\}|i, \sigma(i))$, que da la ley de movimiento del juego. Para que

esta explicación quede más clara se pondrá como ejemplo un juego con dos jugadores esto es: Dado el estado x y que los jugadores eligieron las acciones $a_1 \in A_1(i)$ y $a_2 \in A_2(i)$, respectivamente entonces los jugadores reciben la recompensa $r_1(i, a_1, a_2)$ y $r_2(i, a_1, a_2)$ y la probabilidad de que el próximo estado sea j en el siguiente paso es $p(\{j\}|i, a_1, a_2)$.

Es importante en este momento mencionar que las ideas y definiciones de estrategias puras y mixtas así como la de equilibrio de Nash serán introducidas; esto debido a que se repetirán en muchas de las secciones precedentes. Sin embargo daremos un ejemplo para ver en acción las definiciones anteriores.

Ejemplo 1 ([69])

El espacio de estados es: $X = \{1, 2\}$, el de acciones es: $A_1(1) = A_2(1) = \{a, b\}$, $A = \{(a, a), (a, b), (b, a), (b, b)\}$ y el número de jugadores es: $N = \{1, 2\}$. En el estado 1, los dos jugadores pueden elegir una acción de su espacio de acciones $A_1(1) = A_2(1) = \{a, b\}$ respectivamente. La recompensa inmediata para el jugador 1 dado las posibles acciones son $r_1(1, a, a) = 4$, $r_1(1, a, b) = 5, r_1(1, b, a) = 3$ y $r_1(1, b, b) = 2$. La suposición para este juego es que es de suma cero, esto significa que $r_2(1, a_1, a_2) = -r_1(1, a_1, a_2)$ para todas las parejas de acciones. Si los jugadores eligieron el par de acción (a, b) en el estado 1, entonces ellos se mueven al estado 2 con probabilidad (1/2,1/2), es decir con probabilidad $1/2$ se quedan en el mismo estado o se mueven al siguiente estado; la probabilidad de quedarse en el mismo estado dado que cualquier otro par es seleccionado es 1.

Si los jugadores se encuentran en el estado 2 entonces solo tienen una elección en su espacio $A_1(2) = A_2(2) = \{b\}$ y las recompensas inmediatas son $r_1(2, b, b) = r_2(2, b, b) = 0$. Una vez que los jugadores alcanzaron el estado 2 entonces se quedan ahí con probabilidad 1 (esto significa que el estado 2 es un estado de recompensa 0 y además absorbente).

Esta descripción anterior se puede presentar mediante un diagrama; en el cual primero se presentan los estados y después en cada estado se dan las posibles alternativas de acciones en la parte superior de la diagonal se muestran los pagos mientras en la parte inferior se menciona con la probabilidad de quedarse o moverse al siguiente estado esto gráficamente se muestra a continuación:

	a	b
a	4, -4 / (1 , 0)	5,-5 / (1/2,1/2)
b	3, -3 / (1 , 0)	2,-2 / (1 , 0)

Estado 1

0, 0 / (0 , 1)

Estado 2

El equilibrio de Nash hablando de una forma intuitiva representa la estrategia del sistema en el cual ninguno de los jugadores tiene la intención de modificar su decisión. Ahora supóngase que existe un factor de descuento δ, esto significa que las ecuaciones que caracterizan las correspondientes estrategia con respecto al equilibrio de Nash son:

$$\pi_1^*(i) = \max_{a_1 \in A_1(i)} (r_1(i, a_1, a_2^*) + \delta \sum_{j \in X} p(j|i, a_1, a_2^*)\pi_1^*(j)),$$

$$\pi_2^*(i) = \max_{a_2 \in A_2(i)} (r_2(i, a_1^*, a_2) + \delta \sum_{j \in X} p(j|i, a_1^*, a_2)\pi_2^*(j)).$$

La idea intuitivas detrás de estas ecuaciones significa que hay que analizar el juego del final al inicio generando las mejores recompensan usando la información que proporcionan las transiciones. Si se supone que el factor de descuento δ es $2/3$, por la forma del juego la recompensa en el segundo estado es 0 entonces en la siguiente tabla muestra el juego al que nos estamos enfrentando.

	a	b
a	$4 + 2/3v,\ -4 - 2/3v$	$5 + 1/3v,\ -5 - 1/3v$
b	$3 + 2/3v,\ -3 - 2/3v$	$2 + 2/3v,\ -2 - 2/3v$

Se puede notar que el para (b, a) no es un equilibrio de Nash para cualquier valor de v. Si se ignoran los casos marginales entonces se puede observar que el equilibrio de Nash para el estado 1 son:

- (a, a) si $v < 3$;
- (b, b) si $v > 9$;
- (a, b) si $3 < v < 9$.

Sólo para ejemplificar, si el par (a, a) fuera el equilibrio de Nash entonces se deben cumplir las siguiente desigualdades $4 + 2/3v > 3 + 2/3v$ y $-4 - 2/3v > -5 - 1/3v$ lo que nos lleva a observar que entonces v tiene que ser menor que 3. Los otros pares que resultan ser equilibrios de Nash para calcular el valor de la v se hacen de manera similar.

Supóngase que los jugadores eligen (a, a) entonces $v = 4 + 2/3v$ por lo que $v = 12$ lo cual es inconsistente con la condición de que $v < 3$. Ahora supóngase que se eligieron (b, b) entonces se tiene

que $v = 2 + 2/3v$ y entonces $v = 6$ lo cual es inconsistente con la condición de que $v > 9$. Finalmente, supóngase que los jugadores selección (a, b) entonces $v = 5 + 1/3v$ y $v = 15/2$ lo cual cumple con la condición de que $3 < v < 9$. Por lo que la única que es una estrategia que es un equilibrio de Nash es par (a, b) en el estado 1. Cabe mencionar que se puede caer en recursiones como lo muestran en [19].

Capítulo 2

Juegos Estocásticos

En este trabajo nosotros consideramos la dinámica de un sistema $S = \{S_n, n = 0, 1, \ldots\}$ el cual va a tiempo discreto $t = 0, 1, \ldots$ con espacio de estados finitos $X = \{1, 2, \ldots, N\}$. Se considera que el comportamiento del sistema esta influenciado S por dos jugadores $P^{(1)}$ and $P^{(2)}$ los cuales tienen objetivos contrarios. Supóngase que al tiempo t el sistema se encuentra en el estado $i \in \mathcal{I}$ entonces el jugador $J^{(1)}$ y el jugador $J^{(2)}$ respectivamente seleccionan una acción $a(1)$, $a(2)$ respectivamente de su conjunto de acciones disponibles $A_1(i)$, $A_2(i)$, entonces el estado j es alcanzado en la siguiente transición con probabilidad $p_{ij}(a(1), a(2))$ y se obtiene la recompensa inmediata $r_i(a(1), a(2))$, algunas veces por simplicidad la recompensa $r_i(a(1), a(2))$ será remplazada por r_i. A lo anterior, lo conoceremos como un juego bipersonal de suma cero como un juego de Markov.
Una función de utilidad se dice separable si existe una función $u_i : X_i \to \mathbb{R}$ tal que $U(x_1, \ldots, x_n) = u_1(x_1) + \ldots + u_n(x_n)$ donde X_i representa el conjunto de preferencias del jugador i.

2.1. Juegos Estocásticos con Flujo de Recompensas mediante una Función de Utilidad Exponencial

Para esto, consideremos una función de utilidad exponencial, digamos $\bar{u}^\gamma(\cdot)$, i.e. una función de utilidad separable con coeficiente de sensibilidad al riesgo $\gamma \in \mathbb{R}$. La utilidad asociada a la recompensa ξ (aleatoria) esta dado por:

$$\bar{u}^\gamma(\xi) := \begin{cases} (\text{sign } \gamma)\, \exp(\gamma\xi), & \text{if } \gamma \neq 0, \quad \text{caso sensible al riesgo,} \\ \xi \qquad \text{for } \gamma = 0 & \text{caso neutral al riesgo.} \end{cases} \tag{2.1}$$

Nótese que $\bar{u}^\gamma(\cdot)$ es continua y estrictamente creciente; para $\gamma > 0$ $\bar{u}^\gamma(\cdot)$ es convexa, si $\gamma < 0$ $\bar{u}^\gamma(\cdot)$ es cóncava. Finalmente si $\gamma = 0$ (caso neutral al riesgo) $\bar{u}^\gamma(\cdot)$ es lineal. Obsérvese que la función de

utilidad $\bar{u}^\gamma(\cdot)$ es separable y multiplicativa si el factor sensible al riesgo $\gamma \neq 0$ y es aditiva si $\gamma = 0$. En particular, para $u^\gamma(\cdot) := \exp(\gamma\xi)$ obtenemos $u^\gamma(\xi_1+\xi_2) = u^\gamma(\xi_1)\cdot u^\gamma(\xi_2)$ si $\gamma \neq 0$ y $u^\gamma(\xi_1+\xi_2) \equiv \xi_1+\xi_2$ para $\gamma = 0$.

Más aún, recuérdese que la certeza equivalente correspondiente a ξ, digamos $Z^\gamma(\xi)$, esta dado por

$$\bar{u}^\gamma(Z^\gamma(\xi)) = \mathsf{E}\,[\bar{u}^\gamma(\xi)] \qquad \text{(el símbolo } \mathsf{E} \text{ representara el operador esperanza).} \tag{2.2}$$

De (2.1), (2.2) se puede concluir inmediatamente que

$$Z^\gamma(\xi) = \begin{cases} \gamma^{-1}\,\ln\{\mathsf{E}\,u^\gamma(\xi)\}, & \text{if } \gamma \neq 0 \\ \mathsf{E}\,[\xi] & \text{for } \gamma = 0. \end{cases} \tag{2.3}$$

El desarrollo del sistema S a través del tiempo es controlada por las acciones de los dos jugares los cuales tiene información completa sobre la historia del sistema. En particular, el jugador $J^{(1)}$, resp. el jugador $J^{(2)}$, trata de maximizar, resp. minimizar la recompensa total. Supóngase que el sistema esta en el estado $i \in X$ si la decisión $a(1) \in A_1(i)$ es tomada por el primer jugador y el jugador $J^{(2)}$ selecciona la decisión$a(2) \in A_2(i)$ tal que "minimiza"la posible recompensa (entonces las decisiones $a(1), a(2)$ no son simultáneas, el jugador 1 $J^{(1)}$ es el líder y el jugador $J^{(2)}$ es el que sigue en el modelo del duopolio de Stackelberg). Las cadenas de decisión de Markov sensible al riesgo pueden ser consideradas como un caso especial de los juegos de Markov con un solo jugador.

Una política (Markoviana) que controla las decisiones del proceso, $\pi = (f^0, f^1, \ldots)$, es identificada como una sucesión de vectores de decisión $\{f^n, n = 0, 1, \ldots\}$ en donde $f^n = (f^{(1)n}, f^{(2)n}) \in \mathcal{F} \equiv \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$. En particular, el jugador $J^{(1)}$, resp. el jugador $J^{(2)}$, genera un secuencia de decisiones $f^{(1),n}$ donde $f^{(1),n} \in \mathcal{F}^{(1)} \equiv \mathcal{A}_1^{(1)} \times \ldots \times \mathcal{A}_N^{(1)}$, resp. $f^{(2),n}$ donde $f^{(2),n} \in \mathcal{F}^{(2)} \equiv \mathcal{A}_1^{(2)} \times \ldots \times \mathcal{A}_N^{(2)}$.

Sea $\pi^m = (f^m, f^{m+1}, \ldots)$, por lo tanto $\pi = (f^0, f^1, \ldots, f^{m-1}, \pi^m)$, en particular $\pi = (f^0, \pi^1)$. El símbolo E_i^π denotará la esperanza matemática si $X_0 = i$ y la política $\pi = (f^n)$ es seguida, en particular, $\mathsf{E}_i^\pi(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f_i^0) \ldots p_{i_{m-1},j}(f_{m-1}^{m-1})$; $\mathrm{P}(X_m = j)$ es la probabilidad de que se elija el estado j al tiempo m.

La política π en la cual se selecciona la misma regla de decisión para todos los tiempos, i.e. $\pi \sim (f)$, se llama estacionaria, por lo tanto si se sigue está política $\pi \sim (f)$, S es una cadena de Markov homogénea con matriz de transición de probabilidad $P(f)$ en la cual ij-ésimo elemento es $p_{ij}(f_i) = p_{ij}(f_i^{(1)}, f_i^{(2)})$. Así mismo $r_i(f_i) := r_i(f_i^{(1)}, f_i^{(2)})$ es la recompensa obtenida para el estado i en un paso. Similarmente, $r(f)$ es el N-ésimo vector columna de las recompensas en un paso en la cual el i-ésimo elemento es igual a $r_i(f_i)$. Una política estacionara $\tilde{\pi}$ se dice aleatoria si existe un vector de decisión $f^{[1]}$, $f^{[2]}, \ldots, f^{[m]} \in \mathcal{F}$ (obsérvese que $f^{[1]} = (f^{[1](1)}, f^{[1](2)}) \in \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$). En la siguiente política $\tilde{\pi}$ se selecciona en el estado i la acción $f_i^{[j]}$ con probabilidad $\kappa_i^{[j]}$ (claro, $\kappa_i^{[j]} \geq 0$ con $\sum_{j=1}^N \kappa_i^{(j)} = 1$ para todo $i \in \mathcal{I}$). Obsérvese que $\mathsf{E}_i^{\tilde{\pi}}(X_m = j) = [P^m(f)]_{ij}$ (por lo tanto $[A]_{ij}$

denota que ij-ésimo elemento de la matriz A, $A \geq B$, resp. $A > B$ sii para cada i, j $[A]_{ij} \geq [B]_{ij}$ resp. $[A]_{ij} \geq [B]_{ij}$ y $[A]_{ij} > [B]_{ij}$ para algún i, j). El símbolo I denota la matriz identidad y el símbolo e esta reservado para el vector columna unitario. Nótese también que $\mathsf{E}_i^{\pi^n}(X_m = j)$ representa que el esto inicial de la política π es X_n.

2.2. Optimalidad en los Procesos de Markov Sensibles y Neutrales al Riesgo

2.2.1. Modelos No Controlados

Sea ξ_n la recompensa acumulada obtenida en las primeras n transiciones consideradas de la cadena de Markov S. Como el proceso comienza en el estado S_0, $\xi_n = \sum_{k=0}^{n-1} r_{S_k}$. Similarmente sea $\xi_{(m,n)}$ la recompensa (aleatoria) acumulada que se obtiene de la m-ésima hasta n-ésima transición (obviamente, $\xi_n = r_{S_0} + \xi_{(1,n)}$, estamos asumiendo tácitamente que $\xi_{(1,n)}$ comienza en el estado X_1).

Introduciendo para arbitrarios $g, w_j \in \mathbb{R}$ $(i, j \in X)$ la función de discrepancia (cf. [36])

$$\tilde{\varphi}_{i,j}(w, g) := r_i - w_i + w_j - g \tag{2.4}$$

fácilmente se puede verificar la siguiente identidad:

$$\xi_n \quad = \quad ng + w_{S_0} - w_{S_n} + \sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w, g). \tag{2.5}$$

Si el proceso comienza en el estado i y se utiliza la política $\pi = (f^n)$ entonces la recompensa total esperada es $V_i^{\pi}(n) := \mathsf{E}_i^{\pi} \xi_n$ la obtenemos inmediatamente de (2.5)

$$V_i^{\pi}(n) \quad = \quad ng + w_i + \mathsf{E}_i^{\pi}\{\sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w, g) - w_{S_n}\}, \qquad \text{en donde} \tag{2.6}$$

$$\mathsf{E}_i^{\pi} \sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w, g) \quad = \quad \sum_{j \in X} p_{ij}(f_i)\{\tilde{\varphi}_{i,j}(w, g) + \mathsf{E}_j^{\pi^1} \sum_{k=1}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w, g)\} \tag{2.7}$$

donde π^1 denota la política π que comienza en (f^1). Los siguientes casos son bien conocidos en el ambiente de la programación dinámica estocástica (véase e.g. [28, 45, 50, 56]).

Si el proceso de Markov S considerado es irreducible (o al menos de una sola cadena [1]) para cualquier política estacionaria $\pi \sim (f)$ g y w_i's (que dependen de $f \in \mathcal{F}$) pueden ser elegidas de manera que:

$$w_i + g = r_i(f_i) + \sum_{j \in X} p_{ij}(f_i) + w_j \Longleftrightarrow \sum_{j \in X} p_{ij}(f_i)\tilde{\varphi}_{i,j}(w, g) = 0 \quad \text{for all} \quad i \in X \tag{2.8}$$

Por lo tanto

$$V_i^{\pi}(n) = ng + w_i - \mathsf{E}_i^{\pi} w_{S_n}.$$

[1] i.e. S contiene una sola clase de estados recurrentes y (posiblemente) algunos estados transitorios

Considérese un modelo sensible al riesgo en virtud de (2.1), (2.4), (2.5) para la esperanza de ξ_n el caso sensible al riesgo $U_i^{\pi}(\gamma, n) := \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \xi_n}$ podemos concluir que

$$U_i^{\pi}(\gamma, n) = \mathrm{e}^{\gamma[ng+w_i]} \times \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g) - w_{S_n}]}. \tag{2.9}$$

Obsérvese que;

$$\mathsf{E}_i^{\pi}\, \mathrm{e}^{\gamma \sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g)} = \sum_{j \in X} p_{ij}(f_i^0)\, \mathrm{e}^{\gamma[r_i - w_i + w_j - g]} \times \mathsf{E}_j^{\pi^1}\, \mathrm{e}^{\gamma \sum_{k=1}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g)} \tag{2.10}$$

$$\mathsf{E}_j^{\pi} \{\mathrm{e}^{\gamma \sum_{k=m}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g)} | S_m = j\} = \sum_{\ell \in X} p_{j,\ell}(f_j^m)\, \mathrm{e}^{\gamma[r_j - w_j + w_\ell - g]} \times \mathsf{E}_\ell^{\pi^{m+1}}\, \mathrm{e}^{\gamma \sum_{k=m+1}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g)} \tag{2.11}$$

En analogía con el caso neutral al riesgo, si la política estacionaria $\pi \sim (f)$ es tomada (2.9) puede ser drásticamente simplificada si los números g, w_j's son seleccionados de tal manera que $\sum_{j \in X} p_{ij}(f_i)\, \mathrm{e}^{\gamma \tilde{\varphi}_{ij}(g,w)} = 1$ for all $i \in X$ [2]. Recordando (2.4) esta condición es equivalente al siguiente conjunto de ecuaciones lineares

$$\mathrm{e}^{\gamma[g(f) + w_i(f)]} = \sum_{j \in X} p_{ij}(f_i)\ \mathrm{e}^{\gamma[r_i + w_j(f)]} \quad (i \in \mathcal{I}) \tag{2.12}$$

para los valores de $g(f)$, $w_i(f) (i = 1, \ldots, N)$; nótese que estos valores dependen del coeficiente de sensibilidad al riesgo seleccionado γ; las Eqs. (2.9) pueden ser llamadas γ-*recompensa promedio/ecuación optimalidad de costo*. En particular, si $\gamma \downarrow 0$ usando la expansión de Taylor (3.5) se obtiene

$$g(f) + w_i(f) = \sum_{j \in X} p_{ij}(f_i)\, [r_i(f_i) + w_j(f)]$$

que corresponde a (2.8).

Introduciendo las nuevas variables $v_i(f) := \mathrm{e}^{\gamma w_i(f)}$, $\rho(f) := \mathrm{e}^{\gamma g(f)}$, y reemplazando las probabilidades de transición $p_{ij}(f_i)$'s por números no negativos definidos por $q_{ij}(f_i) := p_{ij}(f_i) \cdot \mathrm{e}^{\gamma r_i}$ entonces (2.12) puede escribirse alternativamente como el siguiente conjunto de ecuaciones

$$\rho(f) v_i(f) = \sum_{j \in X} q_{ij}(f_i)\, v_j(f) \quad (i \in X). \tag{2.13}$$

De aquí en adelante es conveniente considerar (2.13) en su forma matricial. Para este hecho introduzcamos (cf. [25]) $N \times N$ la matriz no negativa $Q(f) = [q_{ij}(f_i)]$ con radio espectral (eigenvalor de Perron) $\rho(f)$ y con su respectivo eigenvector derecho de Perron $v(f) = [v_i(f_i)]$. Entonces (2.13) puede ser reescrito en su forma matricial como:

$$\rho(f) v(f) = Q(f) v(f). \tag{2.14}$$

[2] Para verificar esta aseveración es suficiente aplicar recursivamente (11) hacia atrás comenzando al tiempo $n - 1$ (cf. [29])

Más aún, si la matriz de transición de probabilidad $P(f)$ es irreducible también $Q(f)$ es *irreducible* y el eigenvector derecho de Perron $v(f)$ puede ser seleccionado *estrictamente positivo.*

De (2.3),(2.9),(2.10),(2.12) inmediatamente se puede obtener que para una política estacionaria $\pi \sim (f)$ que

$$U_i^\pi(\gamma, n) = \mathrm{e}^{\gamma[ng(f)+w_i(f)]} \times \mathsf{E}_i^\pi \mathrm{e}^{\gamma w_{S_n}(f)}, \quad Z_i^\pi(\gamma, n) = \frac{1}{\gamma} \ln U_i^\pi(\gamma, n). \tag{2.15}$$

Como $g(f) = \gamma^{-1} \ln \rho(f)$, $w_i(f) = \gamma^{-1} \ln v_i(f)$ de (2.15) se obtiene que

$$n^{-1} Z_i^\pi(\gamma, n) = g(f) + o(n). \tag{2.16}$$

2.2.2. Procesos de Control de Markov

Para los modelos neutrales al riesgo los siguientes resultados son bien conocidos en la literatura de la programación dinámica (cf. e.g. [11, 45, 50]).

Si el proceso de Markov es irreducible (o con una sola clase) entonces existe una política de decisión $\hat{f} \in \mathcal{F}$ (resp. $f^* \in \mathcal{F}$) además de números $\hat{g}$ (resp. g^*), $\hat{w}_i, i \in X$ (resp. $w_i^*, i \in X$) (son únicos salvo constantes) tal que

$$\hat{w}_i + \hat{g} = \min_{a \in \mathcal{A}_i} \sum_{j \in X} p_{ij}(a)[r_i(a) + \hat{w}_j] = \sum_{j \in \mathfrak{X}} p_{ij}(\hat{f}_i)[r_i(\hat{f}_i) + \hat{w}_j], \tag{2.17}$$

$$\varphi_i(f, \hat{f}) := \sum_{j \in X} p_{ij}(f_i)[r_i(f_i) + \hat{w}_j] - \hat{w}_i - \hat{g} \geq 0 \quad \text{with} \quad \varphi_i(\hat{f}, \hat{f}) = 0, \tag{2.18}$$

resp.

$$w_i^* + g^* = \max_{a \in \mathcal{A}_i} \sum_{j \in X} p_{ij}(a)[r_i(a) + w_j^*] = \sum_{j \in X} p_{ij}(f_i^*)[r(f_i^*) + w_j^*], \tag{2.19}$$

$$\varphi_i(f, f^*) := \sum_{j \in X} p_{ij}(f_i)[r(f_i) + w_j^*] - w_i^* - g^* \leq 0 \quad \text{with} \quad \varphi_i(f^*, f^*) = 0. \tag{2.20}$$

De (2.6),(2.9),(2.17),(2.19) se sigue que $\hat{g} \leq g(f) \leq g^*$, y para políticas estacionarios $\hat{\pi} \sim (\hat{f}), \pi^* \sim (f^*)$ se puede conseguir

$$V_i^{\hat{\pi}}(n) = n\hat{g} + \hat{w}_i - \mathsf{E}_i^{\hat{\pi}} \hat{w}_n, \qquad V_i^{\pi^*}(n) = ng^* + w_i^* - \mathsf{E}_i^{\pi^*} w_n^*. \tag{2.21}$$

En el caso que la ecuación de optimalidad se cumpla para varias decisiones es razonable preferir la decisión que garantice la varianza mínima.

El procedimiento algorítmico aplicado al conjunto de decisiones que cumplen (2.17)–(2.20) permite seleccionar la clase de las políticas óptimas con varianza mínima.

Similarmente, para el caso sensible al riesgo se cumple que:

Si la cadena de Markov es irreducible existe $\hat{f}, f^* \in \mathcal{F}$ y números $\hat{\rho} = \rho(\hat{f})$, $\rho^* = \rho(f^*)$ y vectores estrictamente positivos $\hat{v} = v(\hat{f})$, con elementos $v_i(\hat{f})$ y $v^* = v(f^*)$ con elementos $v_i(f^*)$ tales que para cualquier $f \in \mathcal{F}$ (el max y el min vectoriales deben ser considerados componente a componente)

$$Q(f) \cdot \hat{v} \geq \min_{f \in \mathcal{F}} \{Q(f) \cdot \hat{v}\} = Q(\hat{f}) \cdot \hat{v} = \hat{\rho} \cdot \hat{v} \tag{2.22}$$

$$Q(f) \cdot v^* \leq \max_{f \in \mathcal{F}} \{Q(f) \cdot v^*\} = Q(f^*) \cdot v^* = \rho^* \cdot v^* \tag{2.23}$$

$$\rho(\hat{f}) \equiv \hat{\rho} \leq \rho(f) \leq \rho(f^*) \equiv \rho^* \qquad \text{for all} f \in \mathcal{F}. \tag{2.24}$$

Esto se traduce a palabras como que:
$\hat{\rho} \equiv \rho(\hat{f})$ (resp. $\rho^* = \rho(f^*)$) es mínimo (resp. máximo) eigenvalor posible de $Q(f)$ sobre todo $f \in \mathcal{F}$ (cf. [6, 5, 29]).

2.3. Procesos de Decisión de Markov a los Juegos Estocásticos

Como los procesos de decisión de Markov pueden se considerados como un caso muy especial de los juegos de Markov, es interesante mencionar que los juego estocásticos fueron formulados por Shapley [52] en 1953, muchos años antes de la explosión del interés en los procesos de decisión de Markov. En su artículo fundamental, Shapley estableció la existencia del valor y la estrategia estacionaria para los juegos estocásticos transitorios, i.e. bajo la suposición de que $\sum_j p_{ij} < 1$ lo cual también extiende los juegos estocásticos descontados. Más aún Gillette [26] amplio los resultados dados por Shapley para los modelos "no descontadosz para funciones de pago promediados. (Para observar mejoras a los resultados de Gillette véase los artículos por Ligget y Lippman [32]). Para los primeros resultados en procesos de decisión de Markov se pueden encontrar en los artículos de Bellman y su monografía [4, 5, 6], así como en los artículos de Blackwell [8, 9] y especialmente en el libro de Howard [28].

En contraste con los procesos de decisión de Markov nosotros tenemos que tener en consideración las decisiones tomadas por los dos jugadores. Para esto las ecuaciones de optimalidad (2.17),(2.19) deben ser reemplazadas por las condiciones dadas por el equilibrio de Nash, [40]). Dado la condición del equilibrio de Nash existe $f^* = (f^{(1)*}, f^{(2)*}) \in \mathcal{F} = \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ tal que para cualquier $f_i^{(1)} \in \mathcal{F}_i^{(1)}$ y cualquier $f_i^{(2)} \in \mathcal{F}_i^{(2)}$ para las decisiones resultantes $f_i^d = (f_i^{(1)}, f_i^{(2)*})$, $f_i^u = (f_i^{(1)*}, f_i^{(2)})$, se cumple que:

$$\sum_{j \in \mathcal{J}} p_{ij}(f_i^d)[r(f_i^*) + w_j^*] \leq \sum_{j \in \mathcal{J}} p_{ij}(f_i^*)[r(f_i^*) + w_j^*] \leq \sum_{j \in \mathcal{J}} p_{ij}(f_i^u)[r(f_i^*) + w_j^*], \tag{2.25}$$

Considérese un juego bipersonal de Markov, si las decisiones son seleccionadas por un sólo jugador entonces se puede emplear los métodos estándares que se usan en los procesos de decisión de Markov (véase por ejemplo [10, 14, 22, 23, 46, 67]).

Si se quiere observar extensiones de los métodos mencionados al caso general sobre los juegos bipersonales de Markov véase e.g. [27, 42, 44, 61, 62, 63], en el artículo revisado [47] y en la monografía [24], estos resultados que se presentarán aquí son el desarrollo de [39].

Encontrar la recompensa optima promedio puede ser resuelta a través de un problema de programación lineal, esta forma es la usada en la mayoría de los artículos mencionados previamente. Otra forma de atacar este problema es a través del método de iteración de valores (aproximaciones sucesivas) los cuales son usados en [20, 61, 63]. Iteración de políticas para encontrar la política óptima se pueden encontrar en [62].

2.4. Optimalidad Sensible al Riesgo en los Juegos Estocásticos

En contraste con los modelos no controlados considerados en la sección 3.1 y su extensión a los modelos controlados en la sección 3.2 nosotros suponemos que la utilidad esperada $U_i^{\pi}(\gamma, n)$ depende de las decisiones $f^{(1),n}$, $f^{(2),n}$ tomadas por los dos, en el caso de un sólo jugador este tipo de políticas son abordadas en [12, 13, 59]. Recuérdese que:

$$U_i^{\pi}(\gamma, n) = \mathrm{e}^{\gamma[ng+w_i]} \times \mathsf{E}_i^{\pi} \mathrm{e}^{\gamma[\sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(w,g) - w_{S_n}]}. \tag{2.26}$$

pueden ser drásticamente simplificada si los números g, w_j son seleccionados de tal manera que $\sum_{j\in X} p_{ij}(f_i)\, \mathrm{e}^{\gamma\tilde{\varphi}_{ij}(g,w)} = 1$ y además

$$\mathrm{e}^{\gamma[g(f)+w_i(f)]} = \sum_{j\in\mathcal{J}} p_{ij}(f_i)\ \mathrm{e}^{\gamma[r_i+w_j(f)]} \quad (i \in X)$$

En contraste con los procesos de decisión de Markov nosotros debemos tener en consideración las decisiones tomadas por lo dos jugadores. De hecho las ecuaciones de optimalidad (2.22) y (2.23) deben ser reemplazadas por la condición del equilibrio de Nash, [40]. De acuerdo a la condición del equilibrio de Nash existe $f^* = (f^{(1)*}, f^{(2)*}) \in \mathcal{F} = \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ tal que para cualquier $f_i^{(1)} \in \mathcal{F}_i^{(1)}$ y cualquier $f_i^{(2)} \in \mathcal{F}_i^{(2)}$ para las decisiones obtenidas $f_i^d = (f_i^{(1)}, f_i^{(2)*})$, $f_i^u = (f_i^{(1)*}, f_i^{(2)})$, se cumple que:

$$\sum_{j\in X} q_{ij}(f_i^d) v_j^* \le \sum_{j\in X} q_{ij}(f_i^*) v_j^* = \rho(f^*) v_i^* \le \sum_{j\in X} q_{ij}(f_i^u) v_j^*, \tag{2.27}$$

o en notación matricial $\rho(f)v(f) = Q(f)v(f)$ por lo que buscamos $f^* = (f^{(1)*}, f^{(2)*}) \in \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ tal que:

$$Q(f^d)v(f^*) \le \rho(f^*)v(f^*) = Q(f^*)v(f^*) \le Q(f^u)v(f^*) \tag{2.28}$$

en donde $\rho(f^*) = \mathrm{e}^{\gamma g(f^*)}$, $v_i(f^*) = \mathrm{e}^{\gamma w_i(f^*)}$.

De (2.3),(2.9),(3.3),(3.5) de manera inmediata obtenemos un política estacionaria $\pi \sim (f)$ tal que

$$U_i^{\pi^*}(\gamma, n) = \mathrm{e}^{\gamma[ng(f^*)+w_i(f^*)]} \times \mathsf{E}_i^{\pi^*} \mathrm{e}^{\gamma w_{S_n}(f^*)}, \quad Z_i^{\pi^*}(\gamma, n) = \frac{1}{\gamma} \ln U_i^{\pi^*}(\gamma, n), \tag{2.29}$$

$$n^{-1} Z_i^{\pi^*}(\gamma, n) = g(f^*) + o(n). \tag{2.30}$$

2.4.1. Cotas en la Optimalidad Sensible al Riesgo para el Caso Promedio

Como la recompensa sensible al riesgo promediada $g(f) = \gamma^{-1} \ln[\rho(f)]$ y $\rho(f)$ es el eigenvalor de Perron para una matriz no negativa $Q(f)$, es bien sabido que (e.g. [25, 55, 57, 67]) que para cualquier $f', f'' \in \mathcal{F}$ $Q(f') \leq Q(f'') \Rightarrow \rho(f') \leq \rho(f'')$. Para generar cotas superiores e inferiores respecto al mínimo y máximo eigenvalor de Perron $\rho(f^*)$ nosotros reemplazamos los elementos $q_{ij}(f_i^{(1)}, f_i^{(2)})$ por sus mínimos y máximos valores posibles $q_{ij}^{'}$ y $q_{ij}^{''}$. Entonces el problema es aproximar mediante una cadena de Markov sensible al riesgo (no controlada) y como consecuencia es posible generar cotas superiores e inferiores para $\rho(f^*) = \mathrm{e}^{\gamma g(f^*)}$ calculando los eigenevalores de Perron (i.e. el radio espectral) de matrices no negativas. Desafortunadamente, usando este método podemos esperar cotas no muy eficientes respecto al valor óptimo de la recompensa sensible al riesgo en el caso promedio.

Mejores cotas pueden ser obtenidas si se hace un análisis más detallado sobre el conjunto de matrices admisibles. Claro, es razonable sugerir procesos algorítmicos que no necesiten evaluar todas la matrices admisibles. El algoritmo 1 contiene una pequeña modificación al método de iteración de políticas citado en [29] solamente para encontrar el máximo eigenvalor de Perron en un conjunto de matrices no negativas e irreducibles.

Algoritmo 1. Iteración de políticas para encontrar máximos, resp. mínimos, eigenvalores de Perron.

Paso 0. Encuéntrese la matriz $Q^{(0)} := Q(f^{(1),0}, f^{(2),0})$ con $f^{(1),0} \in \mathcal{F}^{(1)}$, $f^{(2),0} \in \mathcal{F}^{(2)}$ tal que la suma de las filas sea máxima (resp. mínima).

Paso 1. Para las matrices $Q^{(k)}$ $(k = 0, 1, \ldots)$ calcúlese el radio espectral $\rho^{(k)}$ con su respectivo eigenvector de Perron derecho $v^{(k)}$.

Paso 2. Construyase (si es posible) la matriz $Q^{(k+1)} := Q(f^{(1),k+1}, f^{(2),k+1})$ con $f^{k+1} := (f^{(1),k+1}, f^{(2),k+1})$ en donde $f^{(1),k+1} \in \mathcal{F}^{(1)}$, $f^{(2),k+1} \in \mathcal{F}^{(2)}$, tal que

$$Q^{(k+1)} \cdot v^{(k)} \geq \rho^{(k)} v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{resp.} \quad Q^{(k+1)} \cdot v^{(k)} \leq \rho^{(k)} v^{(k)} = Q^{(k)} \cdot v^{(k)} \tag{2.31}$$

Paso 3. Si $Q^{(k+1)} = Q^{(k)}$ entonces váyase al paso 4, de otra manera sea $k := k + 1$ y repítase el paso 1.

Paso 4. Sea $\hat{Q} := Q^{(k+1)}$, $\hat{\rho} := \rho^{(k+1)}$, $\hat{v} := v^{(k+1)}$, $\hat{f} := f^{(k+1)}$ y deténgase. $\hat{\rho}$ es el máximo (resp. mínimo) eigenvalor de Perron .

El corazón del algoritmo anterior es lo siguiente:

Rutina de mejoramiento de política:

Como el eigenvector derecho (resp. izquierdo) de Perron $v^{(k)}$ (resp. $z^{(k)}$) de una matriz irreducible $Q^{(k)}$ cumple que $Q^{(k)} \cdot v^{(k)} = \rho^{(k)} v^{(k)}$ (resp. $z^{(k)} Q^{(k)} = \rho^{(k)} z^{(k)}$) if $\varphi^{(k+1)} := Q^{(k+1)} \cdot v^{(k)} - Q^{(k)} \cdot v^{(k)} > 0$ (resp. < 0) entonces

$$Q^{(k+1)} \cdot v^{(k+1)} - Q^{(k)} \cdot v^{(k)} = \rho^{(k+1)}[v^{(k+1)} - v^{(k)}] + [\rho^{(k+1)} - \rho^{(k)}]v^{(k)}$$

Por lo cual premultiplicando la desigualdad anterior $z^{(k+1)}$ (un vector fila estrictamente positivo) entonces llegamos a que:

$$\rho^{(k+1)} \cdot z^{(k+1)}[v^{(k+1)} - v^{(k)}] + [\rho^{(k+1)} - \rho^{(k)}] \cdot z^{(k+1)} v^{(k)} = z^{(k+1)} Q^{(k+1)}[v^{(k+1)} - v^{(k)}] + z^{(k+1)} \varphi^{(k+1)}$$

lo cual implica que: $z^{(k+1)} \varphi^{(k+1)} = [\rho^{(k+1)} - \rho^{(k)}] z^{(k+1)} v^{(k)}$.
Como $z^{(k+1)} v^{(k)} > 0$ si $z^{(k+1)} \varphi^{(k+1)} > 0$ (resp. $z^{(k+1)} \varphi^{(k+1)} < 0$) entonces $\rho^{(k+1)} > \rho^{(k)}$ (resp. $\rho^{(k+1)} < \rho^{(k)}$).

Un ejemplo ilustrativo.

Sea $X = \{1, 2\}$, $\mathcal{A}_1^{(1)} = \mathcal{A}_1^{(2)} = \mathcal{A}_2^{(1)} = \mathcal{A}_2^{(2)} = \{1, 2\}$ y las correspondientes transiciones de probabilidad dadas por los vectores fila $p_i(f_i^{(1)}, f_i^{(2)}) = [p_{i1}(f_i^{(1)}, f_i^{(2)}), p_{i2}(f_i^{(1)}, f_i^{(2)})]$ for $f_i^{(1)}, f_i^{(2)} = 1, 2$. La recompensa asociada al estado i es igual a $r_i(f_i^{(1)}, f_i^{(2)})$.

El siguiente ejemplo es tomado de [24], Ejemplo 3.2.2, de la página 96. Sea la transición dada y las recompensas en un paso las siguientes:

$p_1(1,1) = [0{,}5; 0{,}5]$	$r_1(1,1) = 10$	$p_1(1,2) = [0{,}5; 0{,}5]$	$r_1(1,2) = 6$
$p_1(2,1) = [0{,}8; 0{,}2]$	$r_1(2,1) = 4$	$p_1(2,2) = [0{,}8; 0{,}2]$	$r_2(2,2) = 8$
$p_2(1,1) = [0{,}3; 0{,}7]$	$r_2(1,1) = 2$	$p_1(1,2) = [0{,}3; 0{,}7]$	$r_2(1,2) = 5$
$p_2(2,1) = [0{,}9; 0{,}1]$	$r_2(2,1) = 4$	$p_2(2,2) = [0{,}9; 0{,}1]$	$r_2(2,2) = 10$

Considerando el modelo sensible al riesgo entonces se reemplazan las recompensas en un paso $r_i(f_i^{(1)}, f_i^{(2)})$ por
$\bar{r}_i(f_i^{(1)}, f_i^{(2)}) := \ln[r_i(f_i^{(1)}, f_i^{(2)})]$ si $r_i(f_i^{(1)}, f_i^{(2)}) > 0$ o por
$\bar{r}_i(f_i^{(1)}, f_i^{(2)}) := \ln[-r_i(f_i^{(1)}, f_i^{(2)})]$ if $r_i(f_i^{(1)}, f_i^{(2)}) < 0$.
Obsérvese que $\mathrm{e}^{\gamma \bar{r}_i(f_i^{(1)}, f_i^{(2)})} = |r_i(f_i^{(1)}, f_i^{(2)})|^\gamma$. Recuérdese que
$q_{ij}(f_i^{(1)}, f_i^{(2)}) =: p_{ij}(f_i^{(1)}, f_i^{(2)}) \times r_i(f_i^{(1)}, f_i^{(2)})$, sean los vectores fila
$q_i(f_i^{(1)}, f_i^{(2)}) =: [q_{i1}(f_i^{(1)}, f_i^{(2)}), q_{i2}(f_i^{(1)}, f_i^{(2)})]$. Entonces $Q(f^{(1)}, f^{(2)})$ es una matriz cuadrada (no negativa) en la cual la i-ésima fila es igual a $q_i(f_i^{(1)}, f_i^{(2)})$.

En particular, si $\gamma = 1$, resp. $\gamma = 0{,}5$,

$\gamma = 1$	$\gamma = 1$		$\gamma = 0{,}5$	$\gamma = 0{,}5$
$q_1(1,1) = [5;5]$	$q_1(1,2) = [3;3]$		$q_1(1,1) = [1{,}581;1{,}581]$	$q_1(1,2) = [1{,}225;1{,}225]$
$q_1(2,1) = [3{,}2;0{,}8]$	$q_1(2,2) = [6{,}4;1{,}6]$		$q_1(2,1) = [1{,}6;0{,}4]$	$q_1(2,2) = [2{,}2628;0{,}5656]$
$q_2(1,1) = [0{,}6;1{,}4]$	$q_2(1,2) = [1{,}5;3{,}5]$		$q_2(1,1) = [0{,}4242;0{,}9899]$	$q_2(1,2) = [0{,}671;1{,}5652]$
$q_2(2,1) = [3{,}6;0{,}4]$	$q_2(2,2) = [9;1]$		$q_2(2,1) = [1{,}8;0{,}2]$	$q_2(2,2) = [2{,}846;0{,}3162]$

Como podemos ver, si $\gamma = 1$, si se selecciona en el estado 1 la acción (1,1) y en el estado 2 la decisión (2,2) entonces el radio espectral de la matriz resultante es igual a 10 – el valor máximo posible. Similarmente, seleccionando en el estado 1 la decisión (2,1) y en el estado 2 la decisión (1,1) el radio espectral de la matriz resultante es igual a 3.4358 – el mínimo eigenvalor posible.
Sin embargo, si $\gamma = 0{,}5$, seleccionando en el estado 1 la decisión (1,1) y en el estado 2 la decisión (2,2) el radio espectral respecto a la matriz resultante es 3.1621 – el máximo valor posible. El mínimo valor posible del radio espectral otra vez se obtiene al utilizar en el estado 1 la decisión (2,1) y en el estado 2 la decisión (1,1) el radio espectral de la matriz es igual a 1.8075, el cual es muy cercano al radio espectral 1.8378 que se obtiene de utilizar en el estado 1 la decisión (1,2) y en el estado 2 la decisión (1,1) que es incambiable.
Obviamente, si $\gamma = 0$ el radio espectral es igual a 1 para todas las decisiones.

Más aún, si el coeficiente de sensibilidad al riesgo $\gamma = 1$ calculando de manera directa se puede observar que si el jugador selecciona la decisión 2 (resp 1) en los dos estados, entonces el jugador 1 maximiza su beneficio si selecciona la acción 2 en los dos estados (resp. la acción 1 en el estado 1 y la acción 2 en el estado 2. Si el jugador 2 elige la decisión 2 en el estado 1 y la decisión 1 en el estado 2 la política optima del jugador 1 es seleccionar la acción 2 en los dos estados. Finalmente, si el el segundo jugador selecciona la decisión 1 en el estado 1 y la decisión 2 en el estado 2 la política óptima del jugador 1 es seleccionar en el estado 1 la acción 1 y la acción 2 en el estado 2. Obsérvese que en este caso es necesario resolver 4 problemas que conciernen a encontrar la política óptima de la cadena de decisión de Markov sensible al riesgo.

Para finalizar nosotros sugerimos el siguiente procedimiento algorítmico.

2.4.2. Código Computacional

```
clear all
clc
%Probabilidades de los jugadores para los estados
p111=[.5 .5];
p121=[.8 .2];
p112=[.5 .5];
p122=[.8 .2];

p211=[.3 .7];
p221=[.9 .1];
p212=[.3 .7];
p222=[.9 .1];

%Factor de sensibilidad al riesgo
ga=1

%Recompensa ya incluyendo el factor sensible al riesgo
r111=10^ga;
r121=4^ga;
r112=6^ga;
r122=8^ga;



r211=2^ga;
r221=4^ga;
r212=5^ga;
r222=10^ga;

%Calculo de las pol ticas

comb=combinator(2,2,'p','r')
[m1,m2]=size(comb)
D=[];
F=[];
for i=1:m1
    D=comb(i,:);
    DO=num2str(D);
    DOO=regexprep(DO,'[^\w'']','');
    F=[F;DOO]

end
%C lculo de la q
```

```
for i=1:m1
    for l=1:2
        clear p
        w=strcat('p',num2str(l),num2str(F(i,:)));
        p=eval(w);
        k=strcat('r',num2str(l),num2str(F(i,:)));
        r=eval(k);
        nu=num2str(F(i,:));
        lc=sprintf('q%d%s=p*r',l,nu);
        eval(lc);
    end
end

%Cálculo de la matriz A, en la variable W se guardan los eigenvalores
W=[];
for i=1:m1
        nu=num2str(F(i,:));
        z= strcat('q',num2str(1),nu);
        z1=eval(z);

    for l=1:m1;
        nu1=num2str(F(l,:));
        q= strcat('q',num2str(2),nu1) ;
        w=eval(q);

        lc=sprintf('A%d%s = [z1;w]',l, nu);
        eval(lc);
        e= strcat('A',num2str(l),num2str(nu));
        w1=eval(e);
       W=[W,eig(w1)];

    end
end



W;
%Aquí encontramos el máximo valor de W y encontramos las coordenadas
[M,I] = max(W(:));
[R,C1]=ind2sub(size(W),I);


%Usando la información previa se obtiene el index para la actualización

b=mod(C1,m1);
```

```
if b==0;
    b=m1;
end
a=[1:4*m1];
v=reshape(a,m1,[]);
[r,c]=find(v==C1);
C2=c;
CO=comb(C2,:);
CO2=num2str(CO);
CO1=regexprep(CO2,'[^\w'']','');

%Obtenci n de la matriz y de los eigenvalores para el segundo jugador pero ahora
    hicimos una minimizaci n para lo eigenvalores



P1=[];
lc=strcat('A',num2str(b),num2str(CO1));
eval(lc)
lc1=eval(lc);
[eiv, eig]=eig(lc1);
to=eiv(:,1);

        for j=1:m1;
        nu=num2str(F(C2,:));
        z1=strcat('A',num2str(j),num2str(nu));
        w1=eval(z1)*to;
        P1=[P1,w1];
        end

[M1,I1] = min(P1(:));
[R1,C3]=ind2sub(size(P1),I1);
COO=comb(C3,:);
COO2=num2str(COO);
COO1=regexprep(COO2,'[^\w'']','');
lc2=strcat('A',num2str(C2),num2str(COO1));
eval(lc2)
lc3=eval(lc2);
clear eig;
[eiv1,eig1]=eig(lc3);
to1=eiv1(:,1);

%Una vez obtenida la informaci n del segundo jugador entonces necesitamos otra vez
    maximizar
%respecto al primer jugador y eso es lo que sigue a continuaci n.
```

```
P2=[];
for i=1:m1;
        nu1=num2str(F(i,:));
        z2=strcat('A',num2str(C3),num2str(nu1));
        w2=eval(z2)*to1;
        D=sum(w2);
        P2=[P2,D];
end
[M11,I11]=max(P2);

COO=comb(I11,:);
COO2=num2str(COO);
COO1=regexprep(COO2,'[^\w'']','');
lc=strcat('A',num2str(C2),num2str(COO1));
eval(lc)

%Una vez obtenida la informaci n del segundo jugador entonces necesitamos otra vez
    minimizar
%respecto al primer jugador y eso es lo que sigue a continuaci n.


P3=[];
clear eig
lc1=eval(lc);
[eiv, eig]=eig(lc1);
to=eiv(:,1);
        for j=1:m1;
        nu=num2str(F(C2,:));
        z1=strcat('A',num2str(j),num2str(nu));
        w1=eval(z1)*to;
        P3=[P3,w1];
        end
[M1,I1] = min(P3(:));
[R1,C3]=ind2sub(size(P3),I1);
COOO=comb(C3,:);
COOO2=num2str(COOO);
COOO1=regexprep(COOO2,'[^\w'']','');
lc2=strcat('A',num2str(C3),num2str(COO1));
eval(lc2)
tf=strcmp(lc,lc2);

%Necesitamos repetir este procesos hasta encontrar la estrategia  ptima  para los dos
    jugadores
```

```
% la estrategia max min.
while tf~=1;
lc3=eval(lc2);
clear eig;
[eiv1,eig1]=eig(lc3);
to1=eiv1(:,1);
P2=[];
for i=1:m1;
        nu1=num2str(F(i,:));
        z2=strcat('A',num2str(C3),num2str(nu1));
        w2=eval(z2)*to1;
        D=sum(w2);
        P2=[P2,D];
end
[M11,I11]=max(P2);

COO=comb(I11,:);
COO2=num2str(COO);
COO1=regexprep(COO2,'[^\w'']','');
lc=strcat('A',num2str(C2),num2str(COO1));
eval(lc)



P3=[];
clear eig
lc1=eval(lc);
[eiv, eig]=eig(lc1);
to=eiv(:,1);
        for j=1:m1;
        nu=num2str(F(C2,:));
        z1=strcat('A',num2str(j),num2str(nu));
        w1=eval(z1)*to;
        P3=[P3,w1];
        end
[M1,I1] = min(P3(:));
[R1,C3]=ind2sub(size(P3),I1);
COOO=comb(C3,:);
COOO2=num2str(COOO);
COOO1=regexprep(COOO2,'[^\w'']','');
lc2=strcat('A',num2str(C3),num2str(COO1));
eval(lc2)

end
```

```
%Aqu  encontramos la soluci n
'La_soluci n_es:'
eval(lc2)
```

Nota 2.4.1. *Es importante notar que el siguiente código funciona para recompensas estrictamente positivas porque con tan solo cambiar* $r_1(1,1) = -10$ *y* $r_2(2,2) = -10$ *en el ejemplo anterior entonces se puede comprobar que el código se cicla debido a que no hay un equilibrio de Nash.*

Algoritmo 2. (Iteración de políticas para aproximar la recompensa óptima para el caso promediado.)

Paso 0. Encuéntrese la matriz $Q^{(0)} := Q(f^{(1),0}, f^{(2),0})$ con $f^{(1),0} \in \mathcal{F}^{(1)}$, $f^{(2),0} \in \mathcal{F}^{(2)}$ tal que el radio espectral sea máximo (resp. mínimo).

Paso 1. Para la matriz $Q^{(k)}$ $(k = 0, 1, \ldots)$ cálcula el radio espectral $\rho^{(k)}$ con su respectivo eigenvector derecho de Perron $v^{(k)}$.

Paso 2. Construyase (si es posible) la matriz $Q^{(k+1)} := Q(f^{(1),k+1}, f^{(2),k+1})$ con

$f^{k+1} := (f^{(1),k+1}, f^{(2),k+1})$ donde $f^{(1),k+1} \in \mathcal{F}^{(1)}, f^{(2),k+1} \in \mathcal{F}^{(2)}$, tal que

$f^{(1),k+1} = f^{(1),k}$ para k impar, resp. $f^{(2),k+1} = f^{(2),k}$ para k par, y

$$Q^{(k+1)} \cdot v^{(k)} \leq \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{si } k \text{ es impar} \qquad \text{resp.} \tag{2.32}$$

$$Q^{(k+1)} \cdot v^{(k)} \geq \rho^{(k)}\, v^{(k)} = Q^{(k)} \cdot v^{(k)} \quad \text{si } k \text{ es par} \tag{2.33}$$

Paso 3. Si para algún $\ell = 0, 1, \ldots k$ sucede que $Q^{(k+1)} = Q^{(\ell)}$ entonces vaya al paso 4, de otro modo sea $k := k + 1$ y repítase el paso 1.

Paso 4. Sea $\bar{Q} := Q^{(\ell)} Q^{(\ell+1)} \cdot \ldots Q^{(k)}$. Calcúlese $\bar{\rho}$, el radio espectral de $\bar{Q}$ y pare.

Entonces $\rho^* = (\bar{\rho})^{\frac{1}{k-\ell}}$ es igual al límite de la recompensa promediada sensible al riesgo generada por las decisiones por el primer y segundo jugador en la clase de las políticas no aleatorias.

Evaluación de políticas estacionarias no aleatorias

con coeficiente sensible al riesgo $\gamma = 1$.

Matriz No.	1	2	3	4
estado/acción	1/(1,1)	1/(1,2)	1/(2,1)	1/(2,2)
estado/acción	2/(1,1)	2/(1,1)	2/(1,1)	2/(1,1)
Matriz resultante	$\begin{bmatrix} 5 & 5 \\ 0{,}6 & 1{,}4 \end{bmatrix}$	$\begin{bmatrix} 3 & 3 \\ 0{,}6 & 1{,}4 \end{bmatrix}$	$\begin{bmatrix} 3{,}2 & 0{,}8 \\ 0{,}6 & 1{,}4 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 0{,}6 & 1{,}4 \end{bmatrix}$
Eigenvalor	5.698	3.762	3.4357	6.5851
Eigenvector	[0,9904;0,1383]	[0.9692;0.2462]	[0.9592;0.2826]	[0.9931;0.1149]
Matriz No.	5	6	7	8
estado/acción	1/(1,1)	1/(1,2)	1/(2,1)	1/(2,2)
estado/acción	2/(2,1)	2/(2,1)	2/(2,1)	2/(2,1)
Matriz resultante	$\begin{bmatrix} 5 & 5 \\ 3{,}6 & 0{,}4 \end{bmatrix}$	$\begin{bmatrix} 3 & 3 \\ 3{,}6 & 0{,}4 \end{bmatrix}$	$\begin{bmatrix} 3{,}2 & 0{,}8 \\ 3{,}6 & 0{,}4 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 3{,}6 & 0{,}4 \end{bmatrix}$
Eigenvalor	7.5260	5.2341	4.0000	7.2419
Eigenvector	[0.8926;0.4509]	[0.8020;0.59732]	[0.7071;0.7071]	[0.885;0.4656]
Matriz No.	9	10	11	12
estado/acción	1/(1,1)	1/(1,2)	1/(2,1)	1/(2,2)
estado/acción	2/(1,2)	2/(1,2)	2/(1,2)	2/(1,2)
Matriz resultante	$\begin{bmatrix} 5 & 5 \\ 1{,}5 & 3{,}5 \end{bmatrix}$	$\begin{bmatrix} 3 & 3 \\ 1{,}5 & 3{,}5 \end{bmatrix}$	$\begin{bmatrix} 3{,}2 & 0{,}8 \\ 1{,}5 & 3{,}5 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 1{,}5 & 3{,}5 \end{bmatrix}$
Eigenvalor	7.0895	5.360	4.4557	7.0719
Eigenvector	[0.9227;0.3856]	[0.7826;0.6225]	[0.5373;0.8434]	[0.9220;0.3872]
Matriz No.	13	14	15	16
estado/acción	1/(1,1)	1/(1,2)	1/(2,1)	1/(2,2)
estado/acción	2/(2,2)	2/(2,2)	2/(2,2)	2/(2,2)
Matriz resultante	$\begin{bmatrix} 5 & 5 \\ 9 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 3 \\ 9 & 1 \end{bmatrix}$	$\begin{bmatrix} 3{,}2 & 0{,}8 \\ 9 & 1 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 9 & 1 \end{bmatrix}$
Eigenvalor	10.0000	7.2915	5.0000	8.3573
Eigenvector	[0.7071;0.7071]	[0.5729;0.8196]	[0.4061;0.9139]	[0.6329;0.7742]

Ejemplo ilustrativo – continuación. Buscar la tasa de crecimiento usando el algoritmo 2.

Comenzando con el máximo eigenvalor de Perron; se obtiene la siguiente sucesión de matrices :

$Q13 \longrightarrow Q6 \longrightarrow Q8 \longrightarrow Q7 \longrightarrow Q5 \longrightarrow Q6$

Comenzando con el mínimo eigenvalor de Perron; se obtiene la siguiente sucesión de matrices :

$Q3 \longrightarrow Q5 \longrightarrow Q6 \longrightarrow Q8 \longrightarrow Q7 \longrightarrow Q5 \longrightarrow Q6$

La matriz resultante $Q = Q6 \cdot Q8 \cdot Q7 \cdot Q5$

El eigenvalor de Perron del producto de matrices es igual a .....

El eigenvalor de Perron de una transición es igual a .....

Coeficiente sensible al riesgo $\gamma = 1$. Comenzando con el eigenvalor máximo de Perron.

Paso	0	1	2
estado/acción	1/(1,1)	1/(1,1)	1/(2,2)
estado/acción	2/(2,2)	2/(1,1)	2/(1,1)
Matriz resultante	$\begin{bmatrix} 5 & 5 \\ 9 & 1 \end{bmatrix}$	$\begin{bmatrix} 5 & 5 \\ 0{,}6 & 1{,}4 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 0{,}6 & 1{,}4 \end{bmatrix}$
Eigenvalor	10	5.579	6.5851
Eigenvector	[1;1]	[0.934;0.1151]	[0.9934;0.1149]
estado/mejoramiento	1/(1,1)	1/(2,2)	1/(2,2)
estado/mejoramiento	2/(1,1)	2/(1,1)	2/(1,1)

Coeficiente sensible al riesgo $\gamma = 1$. Comenzando con el mínimo eigenvalor de Perron.

Paso	0	1
estado/acción	1/(2,1)	1/(2,2)
estado/acción	2/(1,1)	2/(1,1)
Matriz Resultante	$\begin{bmatrix} 3{,}2 & 0{,}8 \\ 0{,}6 & 1{,}4 \end{bmatrix}$	$\begin{bmatrix} 6{,}4 & 1{,}6 \\ 0{,}6 & 1{,}4 \end{bmatrix}$
Eigenvalor	3.435	6.5851
Eigenvector	[0.9592;0.2887]	[0.9934;0.1149]
estado/mejoramiento	1/(2,2)	1/(2,2)
estado/mejoramiento	2/(1,1)	2/(1,1)

Coeficiente sensible al riesgo $\gamma = 0{,}5$. Comenzando con el eigenvalor máximo de Perron.

Paso	0	1	2
estado/acción	1/(1,1)	1/(1,2)	1/(2,2)
estado/acción	2/(2,2)	2/(1,2)	2/(1,2)
Matriz Resultante	$\begin{bmatrix} 1{,}581 & 1{,}581 \\ 2{,}846 & 0{,}3162 \end{bmatrix}$	$\begin{bmatrix} 1{,}225 & 1{,}225 \\ 0{,}671 & 1{,}5652 \end{bmatrix}$	$\begin{bmatrix} 2{,}263 & 0{,}5657 \\ 0{,}671 & 1{,}5652 \end{bmatrix}$
Eigenvalor	3.1621	2.137	2.6221
Eigenvector	[1;1]	[0.746;0.6656]	[0.8442;0.53601]
estado/mejoramiento	1/(1,1)	1/(2,2)	1/(2,2)
estado/mejoramiento	2/(1,2)	2/(1,2)	2/(1,2)

Coeficiente sensible al riesgo $\gamma = 0{,}5$. Comenzando con el mínimo eigenvalor de Perron.

Paso	0	1
estado/acción	1/(2,1)	1/(2,2)
estado/acción	2/(1,2)	2/(1,2)
Matriz Resultante	$\begin{bmatrix} 1{,}6 & 0{,}4 \\ 0{,}671 & 1{,}5652 \end{bmatrix}$	$\begin{bmatrix} 2{,}263 & 0{,}5657 \\ 0{,}671 & 1{,}565 \end{bmatrix}$
Eigenvalor	2.101	2.6221
Eigenvector	[0.624;0.7815]	[0.8442;0.5360]
estado/mejoramiento	1/(1,2)	1/(2,2)
estado/mejoramiento	2/(1,2)	2/(1,2)

Capítulo 3

Procesos de Decisión de Markov Transitorios

Considérese la matriz de probabilidad estándar $P(f)$ para la cual el radio espectral $P(f)$ es igual a uno. Recuérdese que $P^*(f) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k(f)$ (el límite de Cesaro $P(f)$) con elementos $p^*_{ij}(f)$) existe); si $P(f)$ es apériodica entonces $P^*(f) = \lim_{k\to\infty} P^k(f)$ y la convergencia es geométrica.

En lo que sigue pondremos atención en lo que llamamos los *modelos transitorios*, en donde el radio espectral de cualquier *matriz de transición de probabilidad admisibles es menor que uno.* Las matrices de transición de probabilidad $\tilde{P}(f)$ son llamadas *transitorias* si el radio espectral de la matriz $\tilde{P}(f)$ es menor que la unidad, i.e. al menos una de las sumas de las filas de $\tilde{P}(f)$ es menor que uno. Entonces $\lim_{n\to\infty}[\tilde{P}(f)]^n = 0$, $\tilde{P}^*(f) = 0$. Obsérvese que si $P(f)$ es estocástica y $\alpha \in (0,1)$ entonces $\tilde{P}(f) := \alpha P(f)$ es transitoria, sin embargo, si $\tilde{P}(f)$ es transitoria entonces puede suceder que la suma de algunas filas puede ser más grande que la unidad.

Estamos ahora en la posición de presentar la caracterización de las políticas de control de los modelos transitorios mediante funciones de discrepancia.

[1] Para este hecho, introduciendo para arbitrario $v_j \in \mathbb{R}$ $(i, j \in X)$ la función de discrepancia $\tilde{\varphi}_{i,j}(v) := r_{ij} - v_i + v_j$ para la recompensa aleatoria obtenida hasta la n-ésima transición se tiene que:

$$\xi_n = v_{S_0} - v_{S_n} + \sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(v). \tag{3.1}$$

De hecho por (3.1) para la esperanza de ξ_n se obtiene que:

$$V_i^{\pi}(n) = v_i + \mathsf{E}_i^{\pi}\{\sum_{k=0}^{n-1} \tilde{\varphi}_{S_k, S_{k+1}}(v) - v_{S_n}\}, \tag{3.2}$$

Los siguiente resultados son bien conocidos en la literatura de la programación dinámica (cf. e.g. [28, 45, 50]).

[1] Las funciones de discrepancia fueron originalmente introducidas por [36], aunque desafortunadamente su trabajo fue reconocido mucho tiempo después ([35] o el artículo revisado [1], page 319).

(i) Para cada $f \in \mathcal{F}$ existen números $v_i(f), i \in X$ tales que:

$$v_i(f) = \sum_{j \in X} p_{ij}(f_i)[r_{ij} + v_j(f)], \qquad (i \in X), \quad \text{i.e.} \tag{3.3}$$
$$\sum_{j \in X} p_{ij}(f_i)\, \tilde{\varphi}_{i,j}(v) = 0 \ \text{ donde } \quad \tilde{\varphi}_{i,j}(v) := r_{ij} - v_i(f) + v_j(f).$$

(ii) Existen decisiones $\hat{f} \in \mathcal{F}$ (resp. $f^* \in \mathcal{F}$) con números $\hat{v}_i, i \in X$ (resp. $v_i^*, i \in X$) tales que

$$\hat{v}_i = \min_{a \in \mathcal{A}_i} \sum_{j \in X} p_{ij}(a)[r_{ij} + \hat{v}_j] = \sum_{j \in \mathcal{J}} p_{ij}(\hat{f}_i)[r_{ij} + \hat{v}_j], \tag{3.4}$$
$$\varphi_i(f, \hat{f}) := \sum_{j \in X} p_{ij}(f)[r_{ij} + \hat{v}_j] - \hat{v}_i \geq 0 \ \text{ with } \ \varphi_i(\hat{f}, \hat{f}) = 0, \tag{3.5}$$

resp.

$$v_i^* = \max_{a \in \mathcal{A}_i} \sum_{j \in X} p_{ij}(a)[r_{ij} + v_j^*] = \sum_{j \in X} p_{ij}(f_i^*)[r_{ij} + v_j^*], \tag{3.6}$$
$$\varphi_i(f, f^*) := \sum_{j \in X} p_{ij}(f)[r_{ij} + v_j^*] - v_i^* \leq 0 \ \text{ with } \ \varphi_i(f^*, f^*) = 0. \tag{3.7}$$

De (3.1),(3.2),(3.4),(3.6) inmediatamente se tiene que $\hat{v}_i \leq v_i(f) \leq v_i^*$, y
para políticas estacionarias $\hat{\pi} \sim (\hat{f}), \pi^* \sim (f^*)$

$$V_i^{\hat{\pi}}(n) = \hat{v}_i - \mathsf{E}_i^{\hat{\pi}} \hat{v}_n, \qquad V_i^{\pi^*}(n) = v_i^* - \mathsf{E}_i^{\pi^*} v_n^*. \tag{3.8}$$

Capítulo 4

Juegos Estocásticos Transitorios

Definición 4.0.1. *Un **Juego Estocástico Transitorio** es considerado un juego en el cual en el estado N el juego finaliza. Esto significa:*

- *El estado N es absorbente, i.e. para todo par (a_1, a_2) de los jugadores se tiene que $P^N_{a_1,a_2}(N) = 1$.; y*
- *La recompensa en este estado es igual a cero. No importando que acciones usen los jugadores. $r^N_1(a_1, a_2) = r^N_2(a_1, a_2) = 0$.*

Es importante recalcar que una vez que el estado N es alcanzado cualquier acción que tomen los jugadores no tendrá ningún efecto en la dinámica del sistema.

Una convención es que el juego se restringe al estado N pero esto no significa ninguna pérdida de generalidad, esto es porque el estado en cuestión puede representar cualquier estado del sistema y entonces solo por notación se considerará el estado N.
De aquí en adelante $r^{X_t}_j(a_1(t), a_2(t))$ denotará la recompensa ganada por el jugador "j" ($j = 1, 2$) al tiempo $t = 0, 1, 2, \ldots$ dado que el juego se encuentra en el estado X_t y las decisiones tomadas por el jugador 1 (resp. 2) son $A_{1,t}$ (resp. $A_{2,t}$).

Nota 4.0.2. *Dado un Modelo de Control de Markov Transitorio en el sentido de los juegos (véase [24], página 161), es un modelo para el cual $\forall i \in X$, $\forall\, \pi \in \mathcal{P}$, la siguiente condición de transitoriedad se cumple:*

$$\sum_{t=0}^{\infty} P^{\pi}_i(X_t \neq N) < \infty,$$

que es equivalente a:

$$\forall i \in X, \forall\, \pi \in \mathcal{P}, \sum_{t=0}^{\infty} \sum_{j=1}^{N-1} P^{\pi}_i(X_t = j) < \infty.$$

*esta fórmula será conocida como la **Condición de Transitoriedad**.*

Teorema 4.0.3. *Esta Condición de Transitoriedad garantiza que la suma total con horizonte infinito será finita i.e., para todo estado inicial $i \in X$ and $j = 1, 2$ se cumple que:*

$$v^{\pi,\phi}(j)_i := \sum_{t=0}^{\infty} E_{i,\pi,\phi} r_j^{X_t}(A_{1,t}, A_{2,t}) < \infty.$$

en donde $v^{\pi,\phi}(j)_i$ es la recompensa total esperada por el jugador "j" si el juego inicia en el estado "i" y la política π (resp. ϕ) son seguidas por el jugador 1 (resp. 2).

Demostración. Es suficiente con comprobar que para cualquier estado inicial i,

$$\begin{aligned} \mid E_{i,\pi,\phi} r_j^{X_t}(A_{1,t}, A_{2,t}) \mid &= \mid \sum_{k \in X} \sum_{(a_1,a_2) \in A_1^k \times A_2^k} r_j^k(a_1, a_2) P_{i,\pi,\phi}(X_t = k, A_{1,t} = a_1, A_{2,t} = a_2) \mid \\ &= \mid \sum_{k=1}^{N-1} \sum_{(a_1,a_2) \in A_1^k \times A_2^k} r_j^k(a_1, a_2) P_{i,\pi,\phi}(X_t = k, A_{1,t} = a_1, A_{2,t} = a_2) \mid \\ &\leq K P_{i,\pi,\phi}(X_t \neq N), \end{aligned}$$

donde $K = \text{máx}_{k \in X}, a_1 \in A_1^k, a_2 \in A_2^k \mid r_j^k(a_1, a_2) \mid$. entonces

$$\sum_{t=0}^{\infty} \mid E_{i,\pi,\phi} r_j^{X_t}(A_{1,t}, A_{2,t}) \mid \leq K \sum_{t=0}^{\infty} P_{i,\pi,\phi}(X_t \neq N) < \infty.$$

La desigualdad se cumple por la condición de transitoriedad y entonces la serie converge por que es convergente

□

Definición 4.0.4. *Un **Juego Estocástico Sumable** es un Juego Estocástico para el cual el criterio de la suma con horizonte infinito esta bien definido*

Observación 4.0.5. *Un Juego Estocástico Transitorio es Sumable.*

Teorema 4.0.6. *Dado un Juego Estocástico Transitorio, si denotamos por τ el tiempo hasta que X_t alcanza el estado N, esto significa que:*

$$\tau = \text{ínf}\{t \mid X_t = N\},$$

o $\tau = \infty$ if $\text{ínf}\{t \mid X_t = N\}$ es un conjunto vacío, se cumple que $\mathbb{P}_{i,\pi,\phi}(\tau < \infty) = 1$, para cualquier par de estrategias π, ϕ del jugador 1 y del jugador para cualquier estado inicial $i \in X$.

Teorema 4.0.7. *Dado un Juego Estocástico Transitorio se cumple que para cada para de estrategias generales π, ϕ y para cada estado inicial $i \in X$ que:*

$$\mathbb{P}_{i,\pi,\phi}(X_N = N) > 0.$$

Teorema 4.0.8. *(Existencia del valor en estrategias estacionarias.) Para cada juego estocástico transitorio de suma cero existe v^* el valor del juego con respecto a las estrategias estacionarias. Esto significa que si*

consideramos un juego estocástico transitorio de suma cero y estrategias estacionarias Π, Φ *del conjunto del jugador 1 y 2 entonce se cumple que para todo estado inicial* $i \in S$ *que:*

$$\sup_{\pi\in\Pi}\inf_{\phi\in\Phi} v_i^{\pi,\phi} = \inf_{\phi\in\Phi}\sup_{\pi\in Pi} v_i^{\pi,\phi} = v_i^*;$$

también se tiene que v^* *es el único vector en* $\mathbb{R}^N$ *que cumple que* $v_N^* = 0$ *y*

$$v_i^* = val\left[r^i(a_1,a_2) + \sum_{k\in X} P_{i,a_1,a_2}(k)v_k^*\right] a_1 \in A_1^i, a_2 \in A_2^i,$$

donde $val\left[(a_1,a_2)\right]_{a_1\in A_1^i,\ a_2\in A_2^i}$ *representa el valor de la matriz con filas* $a_1 \in A_1^i$*, columnas* $a_2 \in A_2^i$ *y entradas* $f(a_1,a_2)$ *el cual nosotros ya conocíamos que existía por el teorema min-max de Von Neumann. Más aún los dos jugadores tienen óptimas estrategias (estacionarias).*

Antes de la demostración se necesitan los siguientes lemas y después se procederá a la realización de la misma.

Lema 4.0.9. *Si nosotros fijamos un Juego Estocástico Transitorio de suma cero con estrategias estacionarias* f *y* g *para el jugador 1 y 2 respectivamente entonces la siguiente desigualdad se cumple:*

$$v^{f,g} = \sum_{t=0}^{\infty} P(f,g)^t r(f,g),$$

donde $v^{f,g}$ *es el vector del valor del juego para las estrategias* f *y* g*.*

Lema 4.0.10. *Si* v *es un vector de* $\mathbb{R}_0^N$ *y* f*,* g *son estrategias estacionarias para un juego estocástico de suma cero tal que*

$$v \leq r(f,g) + P(f,g)v, \tag{4.1}$$

entonces considerando la desigualdad coordenada a coordenada sucede que:

$$v \leq v^{f,g},$$

donde $v^{f,g}$ *es el valor del juego.*

Demostración. Si en la parte izquierda de la ecuación (4.1) sustituimos v usando otra vez la misma ecuación entonces obtenemos:

$$v \leq r(f,g) + P(f,g)r(f,g) + P(f,g)^2 v,$$

entonces si repetimos este proceso k veces se tiene que:

$$v \leq \sum_{t=0}^{k} P(f,g)^t r(f,g) + P(f,g)^{k+1} v.$$

Tenemos que observar que cuando k tiende a infinito la parte de la derecha de la desigualdad tiende a $v^{f,g}$ lo cual nos da el resultado que estamos buscando. Sabemos por el lema anterior que: $\sum_{t=0}^{k} P(f,g)^t r(f,g) + P(f,g)^{k+1} v$

lo cual va a $v^{f,g}$, por lo cual solo nos queda probar que $P^(f,g)^t r(f,g)$ tiende a cero. Así que vamos a analizar que paso con el valor absoluto con respecto al i-ésimo elemento $i = 1, 2, \ldots, N-1, deP(f,g)^{k+1}v$.

$$\begin{aligned} \mid (P(f,g)^{k+1}v)_i \mid &\leq \sum_{j=1}^{N} \mathbb{P}_{i,f,g}(X_{k+1} = j) \mid v_j \mid \\ &= \sum_{j=1}^{N-1} \mathbb{P}_{i,f,g}(X_{k+1} = j) \mid v_j \mid \\ &\leq \sum_{j=1}^{N-1} \mathbb{P}_{i,f,g}(X_{k+1} = j) \parallel v \parallel_\infty \\ &= \mathbb{P}_{i,f,g}(X_{k+1} \neq N) \parallel v \parallel_\infty, \end{aligned}$$

para la cual la primera desigualdad se tienen debido a que $v_N = 0$. Por la definición de juego estocástico transitorio sabemos que $P_{i,f,g}(X_{k+1} \neq N)$ tiendo a cero por que es un término general de una serie convergente, por lo cual la demostración queda completada. □

En este momentos nos encontramos con las condiciones para probar el Teorema 4.0.8.

Demostración. La demostración se dividirá en dos partes. La primera parte se trata de demostrar que existe un único vector $v^* \in \mathbb{R}_0^N$ que cumple que:

$$v_i^* = val\left[r^i(a_1,a_2) + \sum_{j\in X} P_{i,a_1,a_2}(j)v_j\right]_{a_1\in A_1^i, a_2\in A_2^i}$$

para todo $i = 1, \ldots, N$; después veremos que las estrategias f^* y g^* son óptimas y que el valor para estas estrategias es precisamente v^*.

Primer parte. Definimos el la siguiente función $U : \mathbb{R}_0^N \to \mathbb{R}_0^N$ como sigue:

$$(Uv)_i := val\left[r^i(a_1,a_2) + \sum_{j=1}^{N} P_{i,a_1,a_2}(j)v_j\right]_{a_1\in A_1^i, a_2\in A_2^i},$$

Nótese que U es creado de manera que v^* será un punto fijo. Como se conoce la diferencia entre dos valores de dos matrices de la misma dimensiones puede ser acotado por la máxima diferencia entre las entradas de las matrices. Por lo que tenemos que:

$$\begin{aligned} \mid Uv - Uw \mid_i &\leq \max_{a_1,a_2} \mid \sum_{j=1}^{N} P_{i,a_1,a_2}(j)(v-w)_j \mid \\ &\leq \max_{a_1,a_2} \sum_{j=1}^{N} P_{i,a_1,a_2}(j) \mid v - w \mid_j . \end{aligned}$$

en donde $\mid v \mid$ denota el vector $(\mid v_1 \mid, \ldots, \mid v_n \mid)$. Usando la última desigualdad en $U^{n-1}v$ y $U^{n-1}w$ entonces se tienen las siguientes desigualdades:

$$\mid U^n v - U^n w \mid_i \leq \max_{a_1,a_2} \sum_{j=1}^{n} P_{i,a_1,a_2}(j) \mid U^{n-1}v - U^{n-1}w \mid_j .$$

Si nosotros llamamos a f_n y g_n las estrategias puras en las que en el estado i se seleccionan las acciones $a_1(i)$ y $a_2(i)$ para las cuales se toma el máximo en la última ecuación entonces se pueden reescribir las ecuaciones como sigue:

$$\begin{aligned} \mid U^n v - U^n w \mid &\leq P(f_n, g_n) \mid U^{n-1} v - U^{n-1} w \mid \\ &\leq P(f_n, g_n) \dots P(f_1, g_1) \mid v - w \mid . \end{aligned}$$

la segunda desigualdad proviene de aplicar la primera varias veces. Ahora es tiempo de considerar la $\parallel v \parallel$, max-norma y nosotros llamamos a i la coordenada del vector $\mid U^N v - U^N w \mid$ en la cual se alcanza el máximo, esto significa que $\mid U^N v - U^N w \mid_i = \parallel U^N v - U^N w \parallel$; sea π una estrategia para la cual en los primeros N pasos y sigue la regla de decisión $f_1, \dots, f_N$ y sea ϕ una estrategia para el segundo jugador en la cual en los primeros N pasos se sigue la regla de decisión $g_1, \dots, g_N$. Usando la última desigualdad se obtiene que:

$$\begin{aligned} \parallel U^N v - U^N w \parallel &= \mid U^N v - U^N w \mid_i \\ &\leq \sum_{j=1}^{N} P_{i,\pi,\phi}(X_N = j) \mid v - w \mid_j \\ &= \sum_{j=1}^{N-1} P_{i,\pi,\phi}(X_N = j) \mid v - w \mid_j \\ &\leq \sum_{j=1}^{N-1} P_{i,\pi,\phi}(X_N = j) \parallel v - w \parallel \\ &= P_{i,\pi,\phi}(X_N \neq N) \parallel v - w \parallel . \end{aligned}$$

Sabemos por el Teorema de caracterización de los Juegos Estocásticos Transitorios que: $\mathbb{P}_{i,\pi,\phi}(X_N \neq N) < 1$; más aún $\mathbb{P}_{i,\pi,\phi}(X_N \neq N)$ depende de i y en estrategias puras $f_1, \dots, f_N$ y $g_1, \dots, g_N$ las cuales son un número finito. Por esto existe $k < 1$ tal que:

$$\forall v, w \in \mathbb{R}_0^N, \qquad \parallel U^N v - U^N w \parallel \leq k \parallel v - w \parallel,$$

por lo cual U resulta ser una contracción en N pasos; por lo cual existe un único punto fijo $v^* \in \mathbb{R}_0^N$.

Segunda parte.

Considérese las estrategias estacionarias f^* y g^* como fueron definidas al principio. Vamos a probar que para todas las estrategias estacionarias f del jugador 1 y g para el segundo jugador se tiene que:

$$v^{f,g^*} \leq v^* \leq v^{f^*,g},$$

lo cual prueba que $v^{f^*,g^*} = v^*$ y que el par de estrategias (f^*, g^*) so un equilibrio de Nash lo que implica por el teorema que v^* es el valor del juego y que las estrategias son óptimas. Por la definición de f^* tenemos que f_i^* es una estrategia óptima para el jugador 1 en el juego matricial:

$$\left[r^i(a_1, a_2) + \sum_{j \in X} P_{i,a_1,a_2}(j) v_j^* \right]_{a_1 \in A_1^i, a_2 \in A_2^i},$$

el cual tiene valor v_i^*, para esto, para todas las estrategias g del jugador 2 se cumple:

$$r(f^*, g)_i + \sum_{j=1}^{N} P(f^*, g)_{i,j} v_j^* \geq v_i^*,$$

esto es porque el lado izquierdo es el valor del juego matricial con respecto a las estrategias f^* y g. En notación matricial la desigualdad se observa así:

$$r(f^*, g) + P(f^*, g)v^* \geq v^*,$$

Aplicando el último lema se obtiene que $v^* \leq v^{f^*,g}$. Usando el mismo análisis obtenemos que $v^* \geq v^{f,g^*}$ lo cual termina la demostración. □

Recordando la definición de juego estocástico en ([30] p.238) y el criterio de optimalidad para los jugadores $j,\ j = 1, 2.$ dado por:

$$v(f(1), f(2))(j)_i = \sum_{t=0}^{\infty} E_i^{(f(1),f(2))} r_{X_t}^j (f_{X_t}^{(1)}, f_{X_t}^{(2)}).$$

Para ilustrar lo anterior, tomaremos un juego genérico para dos jugadores de una sola etapa, esto es: Dado el estado x y que los jugadores eligen las acciones $a_1 \in A_1(i)$ y $a_2 \in A_2(i)$, lo jugadores reciben la recompensa $r_1(a_1(i), a_2(i))$ y $r_2(a_1(i), a_2(i))$ y la probabilidad de que el siguiente estado sea j es $P(\{j\}|i, a_1, a_2)$ o equivalentemente $p_{ij}(a_1, a_2)$.

Nota 4.0.11. *Un Juego Estocástico es* **Transitorio** *si la probabilidad de transición*

$$\sum_{t=0}^{\infty} P_i(f^{(1)}, f^{(2)})(X_t \neq N) < \infty, \text{ for all } i \in X.$$

o

$$\sum_{t=0}^{\infty} \sum_{j=1}^{N-1} P_i(f^{(1)}, f^{(2)})(X_t = j) < \infty, \text{ for all } i \in X.$$

o equivalentemente,

$$\sum_{t=0}^{\infty} \sum_{x \in X} P(j|\delta(i), i) < \infty.$$

$P(\{j\}|\delta(i), i)$ puede ser interpretada con la probabilidad de que el siguiente estado sea j, dado el estado actual es i y que se uso el perfil $\delta(i)$, en donde el perfil $\delta(i) = (a_1(i), a_2(i))$ depende el estado i.

En las partes siguiente se presentarán condiciones suficiente que garantizan que un juego estocástico sea transitorio. Para este fin, sea $f \in \mathcal{F}$ una política estacionaria tal que $i, j \in X$, $q_{i,j}(f)$ es el i, j's elemento de la matriz $Q(f)$ considerada en el sección 5. En particular, $Q(f^*)$ es la matriz que procede de $\mathcal{F}$ con máximo radio espectral posible, en donde: $v^* = v(f^*)$ es el correspondiente eigenvector derecho.

$$\rho^* v^* = Q(f^*)v^* = \max_{f \in \mathcal{F}} Q(f) \cdot v^* \geq Q(f) \cdot v^*, \tag{4.2}$$

$$\hat{\rho}\hat{v} = Q(\hat{f})\hat{v} = \min_{f \in \mathcal{F}} Q(f) \cdot \hat{v} \leq Q(f) \cdot \hat{v}. \tag{4.3}$$

Bajo estas condiciones estamos en posibilidades de aseverar que:

Teorema 4.0.12. *Si $\rho^* < 1$ en (4.2) el Juego Estocástico es Transitorio.*

Para probar el Teorema 4.0.12 se necesita el siguiente lema.

Lema 4.0.13. *Sea $\pi(n) = (f^{(0)}, \ldots, f^{(n-1)})$ una política (Markoviana arbitraria), entonces para cualquier $v(0) > 0$ existe un eigenvector derecho de Perron $v^* \geq v(0)$, resp. $\hat{v} \leq v(0)$ tal que*

$$(\hat{\rho})^n v(0) \leq v(n, \pi) \leq (\rho^*)^n v^*. \tag{4.4}$$

La prueba se sigue directamente de iterar (4.2) y (4.3) y así se obtiene la demostración de 4.0.12 de manera inmediata.

Capítulo 5

Juegos Estocásticos con Tiempos de Paro

Este tipo de juegos se pueden usar para modelar algunos problemas en los que no se pueden acotar a priori el número de pasos en el que el juego terminará pero cumplen que el número de paso es finito con probabilidad uno, i.e. existe un Tiempo de Paro para estos juegos.

5.1. Preliminares

5.1.1. Modelo de Control de Markov

Definición 5.1.1. *Un **Modelo de Control de Markov** (MCM), [45] es una quintupla*

$$M := \{X, A, \{A(i)|i \in X\}, Q, R\} \text{ que consiste en:}$$

1. *X es un conjunto finito que será conocido como el **espacios de estados**; para el cual "i" se refiere a un elemento arbitrario de X.*
2. *A es un conjunto finito conocido como el **espacio de acciones** o **espacio de controles**.*
3. *$\{A(i)|i \in X\}$ es una familia de subconjuntos no vacíos $A(i)$ de A; $A(i)$ representa el subconjunto de **controles permitidos** para el estado $i \in X$. $\mathbb{K} := \{(i,a)|i \in X, a \in A(i)\}$ es el **espacio de pares estado-acción permitidos**.*
4. *$Q(B|i,a)) := P(X_{t+1} \in B|X_t = i, A_t = a)$, $B \in X$ para $t = 0, 1, 2, \ldots$.*
5. *$R : \mathbb{K} \to \mathbb{R}$ es la **función de recompensa** en el sentido que es el resultado obtenido por aplicar el control "a" dado que el estado fue "i". Nótese que vamos a considerar que: $0 \leq R(i,a) < \infty$, $\forall\, i, \forall\, a$; $\mathbb{R}$ denotará el conjunto de los números reales..*

Consideremos un Sistema de Control Estocástico y supóngamos que podemos observar el sistema en cada época, i.e. el MCM representa las consideraciones previas con espacio de estados X y espacio de controles A, este sistema puede ser observado a cada tiempo $t = 0, 1, \ldots$. Así que si denotamos por X_t al *estado* del sistema y por A_t a la *acción* al tiempo t, entonces podemos describir el sistema como sigue a continuación: Si el sistema se encuentra en el estado $X_t = i \in X$ al tiempo t y se utiliza el control $A_t \in A(i)$ entonces se obtiene una respuesta al sistema dada por $R(i, a)$ que es la consecuencia de elegir la acción $A_t = a$ dado que se esta en el estado i, así mismo el sistema se mueve al siguiente estado digamos X_{t+1} el cual es una función X-valuada con distribución $Q(\cdot|i, a)$. Una vez el proceso se encuentra en el siguiente estado, estamos en condiciones para proceder con otro control y así continuar. Un MCM tiene la característica de que el para cualquier estado la respuesta al sistema y la ley de transición solo dependen del estado actual.

Políticas o Estrategias

Considérese un MCM para cada $t = 0, 1, 2, \ldots$, definimos el espacio H_t de historias admisibles hasta el tiempo t como sigue: $H_t := \mathbb{K}^t \times X = \mathbb{K}^t \times H_{t-1}$ para $t = 1, 2, \ldots$ y $H_0 := X$. Un elemento genérico de H_t se denotará por $h_t = (i_0, a_0, \ldots, i_j, a_j, \ldots, i_t)$, en donde $a_j \in A(i_j)$.

Definición 5.1.2. *Una **Política o Estrategia** [45] es una sucesión $\pi = \{\pi_t, t = 0, 1, \ldots\}$ de medidas de probabilidad π_t sobre el conjunto de controles A dado H_t que satisface que $\pi_t(A(i_t)|h_t) = 1$ para todo $h_t \in H_t$, $t = 0, 1, \ldots$.*

Definición 5.1.3. *Sea F es espacio de todas las funciones $f : X \to A$ tal que $f(i) \in A(i)$ para todo $i \in X$. Llamaremos **Políticas Estacionarias** para las cuales existe una función $f \in F$ tal que $\pi_t(\cdot|h_t)$ está concentrado en $f(i_t) \in A(i_t)$ para todo $h_t \in H_t$ and $t = 0, 1 \ldots$.*

Obsérvese que el conjunto de todas las ***Políticas*** se denotará por $\mathcal{P}$ y el conjunto de las ***Políticas Estacionarias Deterministicas*** por $\mathbb{F}$. Es claro que $\mathbb{F} \subset \mathcal{P}$.

Dada la política π y un estado inicial $X_0 = i$, existe una única medida de probabilidad P_i^π determinada en el espacio producto $\mathbb{H} := \prod_{t=0}^{\infty} \mathbb{K}$ de todos los pares posibles estado-acción del proceso $\{(X_t, A_t)\}$ ([3], [45]); Así mismo al respectivo operador esperanza se le denotará por: E_i^π.

Una ***Función Objetivo*** es una función que cumple lo siguiente:

$$V : \mathcal{P} \times X \to \mathbb{R},$$

la cual es una manera de medir el resultado de todo el proceso.

Definición 5.1.4. *Dado un MCM $\{X, A, \{A(i)|i \in X\}, Q, R\}$, el conjunto de políticas $\mathcal{P}$ y la función objetivo V . El **Problema de Control Óptimo** consiste en determinar una política $\pi \in \mathcal{P}$ si esta existe, tal que:*

$$V(\pi^*, i) = \sup_{\pi \in \mathcal{P}} V(\pi, i), i \in X.$$

Se define la ***Función Óptima de la Recompensa Total Esperada*** como:

$$\mathcal{V}(i) = \sup_{\pi \in \mathcal{P}} E_i^{\pi} \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \quad \text{for all } i \in X, \text{ for all } \pi \in \mathcal{P},$$

y en el mismo sentido podemos decir que π^* es ***Óptima*** si

$$\mathcal{V}(i) = V(\pi^*, i), \text{ for all } i \in X,$$

en donde la ***Función Objetivo de la Recompensa Total Esperada*** esta definida por:

$$V(\pi, i) = E_i^{\pi} \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \quad \text{for all } i \in X.$$

5.1.2. Modelo de Markov Transitorio

Definición 5.1.5. *Un **Modelo de Control de Markov Transitorio** (TMCM) [24] es un Modelo de Control de Markov con una condición adicional la cual es: existe un estado $N \in X$ tal que $Q(\{N\}|N, a) = 1$ y $R(N, a) = 0$ para todo $a \in A(N)$.*

La existencia del estado absorbente N en el TMCM implica que la Recompensa Total sea acotada si el estado N puede ser alcanzado desde cualquier estado de X. Nótese que para cualquier $\beta \in (0, 1)$ un MCM que utiliza el criterio descontado $V_\beta(\pi, i) = E_i^{\pi} \left[\sum_{t=0}^{\infty} \beta R(X_t, A_t) \right]$, para todo $i \in X, \pi \in \mathcal{P}$, puede ser visto como un caso espacial de los TMCM.

5.2. Definición de un Juego Estocástico con Tiempo de Paro

Dada lo anterior, nosotros podemos proponer dos jugadores: El primer jugador con Modelo de Control de Markov Transitorio (TMCM) y para el segundo jugador se propone el concepto de Tiempo de Paro [17, 53].

Como un breviario, este tema concierne a la clase de los juegos bipersonales de suma cero (esto significa que lo que un jugador gana es lo que el otro pierde) a tiempo discreto con transiciones de Markov en un espacio finito. La idea detrás del juego es que para cada tiempo de elección: el Jugador 2 puede parar el sistema pagando un Recompensa Terminal al Jugador 1, y si el juego no es detenido entonces el Jugador 1 puede seleccionar una acción y mover el sistema a otro estado recibiendo una recompensa por el Jugador 2. Para este trabajo el sistema será medido a través del criterio de la Recompensa Total.

Sea $\mathcal{G} = (X, A, \{A(i)\}_{i \in X}, P, R, G)$ un juego secuencial de suma cero con tiempos de paro para dos jugadores 1 y 2, en donde el espacio de estados X es finito dotado con la topología discreta y el espacio de acciones A es un conjunto finito. El jugador 1 esta jugando un TMCM $\{X, A, \{A(i) \mid i \in X\}, P, R\}$ en el ambiente de los juegos y el jugador 2 esta decidiendo si detener el juego al costo de pagar la Recompensa Terminal G al jugador 1. Se introduce la *Recompensa Terminal* G del jugador 2 como $G : X \to \mathbb{R}$ que es el

resultado de detener el juego al estado "i" podemos asumir que: $G(i) \geq R(i) \geq 0$.

Nota: Dado un espacio topológico $\mathbb{K}$, el ***Espacio de Banach*** $B(\mathbb{K})$ consiste en el conjunto de todas las funciones continuas $\hat{R} : \mathbb{K} \to \mathbb{R}$ para las cuales la norma del supremo $\| \hat{R} \|$ es finita, análogamente se define $B(X)$. Obsérvese que para nuestro trabajo nos concentraremos en $R \geq 0$ y $G \geq 0$, y se tiene que $R \in B(\mathbb{K})$ y $G \in B(X)$.

El modelo $\mathcal{G}$ se interpreta como sigue: para cada tiempo $t = 0, 1, 2, \ldots$, los jugadores 1 y 2 observan el estado del sistema digamos, $X_t = i \in X$, y entonces el Jugador 2 puede decidir si detener el juego a expensas de dar un pago terminal $G(i)$ al Jugador 1 o dejar que el sistema continue, en dado caso el Jugador 1 usas la historia observada en los estados anteriores así como sus acciones tomadas y en el estado actual $X_t = i$, para seleccionar una acción (control) $A_t = a \in A(x)$ para mover el sistema. Como resultado de esto el Jugador 1 obtiene una recompensas $R(i, a)$ de parte del Jugador 2 y como resultado al tiempo $t + 1$ el sistema se encontrará en el estado $X_{t+1} = j \in X$ con probabilidad $p_{ij}(a)$.

5.2.1. Las Estrategias para el Juego con Tiempos de Paro

En el caso del Jugador 1 las estrategias (políticas) son las heredadas por los TMCM. Por notación sea $\mathcal{P}$ el conjunto de todas las estrategias.

Sea $\mathcal{F}_t := \bar{\sigma}(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$, donde todos los elementos usados están basados en las estrategias del Jugador 1. ($\bar{\sigma}(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$ representa la σ-álgebra que es generada por todos los elementos $(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$).

El conjunto de estrategias para el Jugador 2 es el espacio $\mathcal{T}$ que consiste en todos los tiempos de paro $\tau : \mathbb{H} \to \mathbb{N}$, ($\mathbb{H} := \prod_{t=0}^{\infty} \mathbb{K}$ y $\mathbb{N} := \{1, 2, 3, \ldots, \} \bigcup \{\infty\}$) con respecto a la filtración $\{\mathcal{F}_t\}$, i.e. para cada entero no negativo t, el evento $[\tau = t]$ pertenece a $\mathcal{F}_t$. ($\mathbb{K} := \{(i, a) \mid i \in X, a \in A(i)\}$ conocido como los pares disponibles estado-acción.

Obsérvese que el juego $\mathcal{G}$ es jugado en el espacio producto $\mathbb{H}$ el cual puede ser construido en la forma canónica [3]; dada una política π y un estado $i \in X$ existe una única medida de probabilidad en $\mathbb{H}$ que puede ser denotada por P_i^{π} (la probabilidad única inducida en el espacio producto de forma natural) en la cual también se puede definir el operador esperanza como E_i^{π}. Nótese también que el los tiempos de paro τ también pertenecen a $\mathbb{H}$. Definimos un *Par de Estrategias* para el juego $\mathcal{G}$ como: (π, τ) donde $\pi \in \mathcal{P}$ y $\tau \in \mathcal{T}$.

Definición 5.2.1. *Dado un estado inicial $i \in X$, la **Recompensa Total Esperada** para el Jugador 1 correspondiente al par $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$ esta dado por:*

$$\begin{aligned} V(i;\pi,\tau): &= E_i^{\pi}\left[\sum_{t=0}^{\tau-1} R(X_t,A_t) + G(X_\tau)I\left[\tau<+\infty\right]\right] \\ &= V(\pi,i) - E_i^{\pi}\left[\sum_{t=\tau}^{\infty} V(\pi,i)\right] + E_i^{\pi}[G(X_\tau)I\left[\tau<+\infty\right]]. \end{aligned}$$

Nota 5.2.2. *Nótese que:*

$$\mid V(i;\pi,\tau)\mid \leq E_i^{\pi}\left[\sum_{t=0}^{\tau-1}\mid R(X_t,A_t)\mid + \parallel G(X_\tau)I\left[\tau<+\infty\right]\parallel\right] \leq \parallel R\parallel E_i^{\pi}[\tau] + \parallel G\parallel < \infty,$$

como consecuencia se obtiene que $V(i;\pi,\tau)$ esta bien definida para cada $\pi\in\mathcal{P}$ y para cada $\tau\in\mathcal{T}$.

Cuando el Jugador 2 utiliza la estrategia τ, la ***Mejor Recompensa Total Esperada*** para el Jugador 1 es $\sup_{\pi\in\Pi} V(i;\pi,\tau)$, y la ***Función del Valor*** del juego es:

$$V^*(i) := \inf_{\tau\in\mathcal{T}}\left[\sup_{\pi\in\mathcal{P}} V(i;\pi,\tau)\right],\ i\in X. \tag{5.1}$$

Intercambiando el orden, el menor valor de la función del juego tiene la siguiente forma:

$$V_*(i) := \sup_{\pi\in\mathcal{P}}\left[\inf_{\tau\in\mathcal{T}} V(i;\pi,\tau)\right],\ i\in X. \tag{5.2}$$

Definición 5.2.3. *Un par $(\pi^*,\tau^*)\in\mathcal{P}\times\mathcal{T}$ es un **Equilibrio de Nash** si*

$$V(i;\pi,\tau^*)\leq V(i;\pi^*,\tau^*),\ i\in X,\ \pi\in\mathcal{P},$$

y

$$V(i;\pi^*,\tau)\geq V(i;\pi^*,\tau^*),\ i\in X,\ \tau\in\mathcal{T}.$$

5.2.2. Teorema Principal

En esta parte vamos a dar una caracterización sobre el Equilibrio de Nash para los Juegos Estocásticos Transitorios Bipersonales con Tiempos de Paro.

Si recordamos la Nota 4.0.2, la Condición de Transitoriedad puede ser reescrita como sigue [33]:

$$\sup_{i\in X,\pi\in\mathcal{P}}\sum_{t=0}^{\infty} P_i^{\pi}[X_t\neq N] < \infty. \tag{5.3}$$

Esta condición es equivalente a la Condición Simultánea de Doeblin [33] , [34]:

$$\sup_{i\in X,\pi\in\mathcal{P}} E_i^{\pi}[\tau] < \infty. \tag{5.4}$$

Proposición 5.2.4. *Existe un entero positivo K tal que:*

$$\sum_{t=K}^{\infty} P_i^{\pi}[X_t\neq N] < 1/2,\ \ i\in X,\pi\in\mathcal{P}.$$

Demostración. El lado izquierdo de la desigualdad es $P_i^{\pi}[\tau>K]$ así que usando la desigualdad de Markov, (5.3) implica que la se cumple la condición deseada si $K>2M$. □

Proposición 5.2.5. *Sea $B'(X)$ la clase de todas las funciones $W : X \to \mathbb{R}$ que satisfacen $W(N) = 0$ definimos el operador $\hat{C}$ en $B'(X)$ como sigue: Para cada $W \in B'(X)$ la función $\hat{C}[W]$ esta dada por:*

$$\hat{C}\,[W]\,(i) = \text{mín}\left\{G(i),\ \sup_{a\in A(i)}\left[R(i,a) + \sum_{j=1}^{N} P_{ij}(a)W(j)\right]\right\}.$$

El operador $\hat{C}^K$ es contractivo, donde K cumple las condiciones de la Proposición 5.2.4.

Demostración. Para $W, V \in B'(X)$ la siguiente desigualdad se cumple:

$$|\hat{C}[W] - \hat{C}[V]|(i) \leq \sup_{a\in A(i)}\left[\sum_{j=1}^{N} P_{ij}(a)|W(j) - V(j)|\right].$$

Como X y A son finitos esta relación implica que existe una política $f \in \mathbb{F}$ tal que:

$$\begin{aligned} |\hat{C}[W](i) - \hat{C}[V](i)| &\leq E_i^f[|W(X_1) - V(X_1)|] \\ &= E_i^f[|W(X_1) - V(X_1)|I[X_1 \neq N]] \\ &= E_i^f[|W(X_1) - V(X_1)|I[\tau > 1]], \end{aligned}$$

en donde la segunda igualdad se da por la inclusión de $W, V \in B'(X)$, obsérvese que la definición de τ se uso en el último paso. Vía inducción se sigue que existe una política Markoviana π tal que para que cada $i \in X$,

$$\begin{aligned} |\hat{C}^K[W](i) - \hat{C}^K[V](i)| &\leq E_i^\pi[|W(X_K) - V(X_K)|I[\tau > K]] \\ &\leq \parallel W - V \parallel P_i^\pi[\tau > K] \leq \parallel W - V \parallel /2, \end{aligned}$$

donde la última desigualdad se cumple por la Proposición 5.2.4 □

Así que con esto probamos en Proposición 5.2.5 que $\hat{C}^K$ es contractivo, por lo que se cumple las condiciones del Teorema del Punto Fijo de Banach [16], esto significa $\hat{C}$ tiene un punto fijo $W^* \in B'(X)$. Más aún, este punto fijo es el único que satisface que:

$$W^*(i) = \text{mín}\left\{G(i),\ \sup_{a\in A(i)}\left[R(i,a) + \sum_{j=1}^{N} P_{ij}(a)W^*(j)\right]\right\},\ i \in X.$$

Para nuestro modelo, W^* es exactamente V^*. Para cada $i \in X$, sea $f^*(i) \in A(i)$ el maximizador de la parte en corchetes de la ecuación mostrada arriba definimos el Tiempo de Paro τ^* como sigue: $\tau^*(i) = 1$ (detener) si $V^*(i) = G(i)$, $\tau(i) = 0$ (continuar) cuando $G(i) > V^*(i)$.

Teorema 5.2.6. *Para el Juego Estocástico con Tiempo de Paro introducido previamente, el par (π^*, τ^*) dado por la Proposición 5.2.5 es un Equilibrio de Nash.*

Demostración. Sea $S^* = \{i \in X | V^*(i) = G(i)\}$ y sea $\tau^*(h) = \text{mín}\{t \geq 0 | X_t \in S^*\}$, $h = (X_0, A_0, X_1, A_1, \ldots) \in \mathbb{H}$. Necesitamos probar dos desigualdades para mostrar que es un Equilibrio de Nash.

Consideremos la primera desigualdad $V(i, \pi, \tau^*) \leq V(i, \pi^*, \tau^*)$, $i \in X$, $\pi \in \mathcal{P}$, donde estamos usando que π^* es la estrategia que es la que desarrolla la sucesión de $f^*(i)$ como en la proposición anterior.

Caso 1. Si $i \in S^*$ entonces $[\tau^* = 0]$ tiene probabilidad 1 con respecto a P_i^π y $P_i^{\pi^*}$ entonces para este caso las recompensa son $G(i) = V^*(i)$.

Caso 2. Si $i \notin S^*$entonces $V^*(i) \geq R(i,a) + \sum_{j\in X} P_{ij}(a)V^*(j)$ y para cada $\pi \in \mathcal{P}$ y como $i \in X \setminus S^*$ se tiene:

$$\begin{aligned} V^*(i) &\geq E_i^\pi[R(X_0, A_0) + V^*(X_1)] \\ &= E_i^\pi\left[R(X_0, A_0)I[\tau^* > 0] + G(X_{\tau^*})I[\tau^* < 1] + I[\tau^* \geq 1]V^*(X_1)\right], \end{aligned}$$

donde la segunda desigualdad se cumple debido a la relación $P_i^\pi[\tau^* \geq 1] = 1$, y por un argumento inductivo para cada entero n y $\pi \in \mathcal{P}$ se obtiene que:

$$\begin{aligned} V^*(i) &\geq E_i^\pi\left[\sum_{t=0}^{n-1} R(X_t, A_t)I[\tau^* > t]\right] \\ &\quad + E_i^\pi\left[G(X_{\tau^*})I[\tau^* < n]\right] + E_i^\pi\left[I[\tau^* \geq n]V^*(X_n)\right],\ i \in X \setminus S^*, \end{aligned}$$

tomando el límite $n \to \infty$, vía el Teorema de la Convergencia Acotada, se sigue de la Definición 5.2.1 y el Teorema 4.0.3 que:

$$\begin{aligned} V^*(i) &\geq E_i^\pi\left[\sum_{t=0}^{\infty} R(X_t, A_t)I[\tau^* > t]\right] + E_i^\pi\left[G(X_{\tau^*})I[\tau^* < +\infty]\right] \\ &= E_i^\pi\left[\sum_{t=0}^{\tau^*-1} R(X_t, A_t)\right] + E_i^\pi\left[G(X_{\tau^*})I[\tau^* < +\infty]\right] \\ &= V(i; \pi, \tau^*); \end{aligned}$$

El Caso 1 junto con el Caso 2 nos llevan a la desigualdad requerida.

Para completar la prueba, hace falta mostrar que: $V(i, \pi^*, \tau) \geq V(i, \pi^*, \tau^*), i \in X, \tau \in \mathcal{T}$. Para obtener esto, consideremos el siguiente juego $\hat{G} := (X, A, \{\hat{A}(i)\}_{i\in X}, P, R, G)$ que es el resultado de reducir $A(i)$ a $\hat{A}(i) = \{\pi^*(i)\} = \{f^*(i)\},\ i \in X$, y restringir el dominio de $R(\cdot)$ y cada $P_{xy}(\cdot)$ a$\hat{A}(i)$. Para este nuevo modelo, la correspondiente clase $\hat{\mathcal{P}}$ de estrategias del Jugador 1 es el singulete $\{f^*\}$ así que la función de valor (superior) asociada a $\hat{G}$ esta dada por:

$$\hat{V}^*(i) = \inf_{\tilde{\tau}\in\mathcal{T}} V(i; \pi^*, \tilde{\tau}),\ i \in X. \tag{5.5}$$

Esta ecuación puede ser obtenida de aplicar (5.1) reemplazando $\mathcal{P}$ por $\hat{\mathcal{P}}$. Aplicando la Proposición 5.2.5 a este juego reducido $\hat{G}$, la función $\hat{V}^*$ se caracteriza por ser la única solución en $B'(X)$ de la ecuación de equilibrio:

$$\hat{V}^*(i) = \min\left\{G(i), \left[R(i, f^*(i)) + \sum_{j\in X} P_{ij}(f^*(i))\hat{V}^*(j)\right]\right\},\ i \in X.$$

así que podemos reemplazar $\hat{V}^*(i)$ por $V^*(i)$ usando la unicidad del resultado dado por la Proposición 5.2.5. Combinando la última observación con (5.5), se sigue que para cada $\tau \in \mathcal{T}$ e $i \in X$,

$$\begin{aligned} V(i; \pi^*, \tau) &\geq \inf_{\tilde{\tau}} V(i; \pi^*, \tilde{\tau}) \\ &= \hat{V}^*(i) \\ &= V^*(i). \end{aligned}$$

Así que obtuvimos que: $V(i;\pi,\tau^*) \leq V^*(i) \leq V(i;\pi^*,\tau)$ y también tenemos que: $V(i;\pi^*,\tau^*) = V^*(i)$, $\forall i \in X$ y que el par (π^*,τ^*) es un *Equilibrio de Nash* y es óptimo para cada jugador. □

Corolario 5.2.7. *Para los Juegos Estocásticos Transitorios con Tiempos de Paro el análisis del problema puede ser hecho por componentes una parte sobre los estados transitorios y en la otra el estado absorbente.*

Demostración. Para este tipo de juegos se puede observar que:

$$E_i^{\pi}\left[\sum_{t=0}^{\tau}\sum_{k\in X} r_1^k(a)p_{ik}^t(a)\right] = G(X_\tau),$$

para algún τ, esto se puede reescribir usando la información del estado absorbente y del tiempo de paro τ^* como:

$$E_i^{\pi}\left[\sum_{t=0}^{\tau^*}\sum_{k\in X} r_1^k(a)p_{ik}^t(a)\right] = G(X_{\tau^*}),$$

estas ecuaciones también se pueden observar como:

$$E_i^{\pi}\left[\sum_{t=0}^{\tau^*-1}\sum_{k\in X\setminus N} r_1^k(a)p_{ik}^t(a) + r_1^N(a)p_{iN}^t(a)\right] = G(X_{\tau^*}),$$

en la que la primera parte puede ser observada como la parte transitoria y la segundo como el estado absorbente.

□

5.3. Ejemplos

Para ilustrar lo anterior mostraremos algunos modelos con sus respectivos detalles que revelarán las aplicaciones de esta teoría.

5.3.1. Ejemplo con un Único Equilibrio de Nash

Ejemplo 1

Sea N un número no negativo fijo y $p \in [0,1]$ entonces los cinco elementos para el TMCM son los siguiente:

I. $X := \{0,1,2,\ldots,N\}$.

II. $A := \{0,1,2,\ldots,\lfloor N/2\rfloor\}$ en donde $\lfloor z\rfloor$ representa el máximo entero de z.

III. Para cada $i \in X$, $A(i) = \{1,2,\ldots,\text{mín}\{i,N-i\}\}$.

IV. La ley de transición de probabilidad definida para $P = [p_{ij}(a)]$ para $i \in X$ y $a \in A(x)$ como : $p_{ii+a}(a) = p$, $p_{ii-a}(a) = q$ donde $q = 1 - p$, $p_{N0}(a) = 1$, $p_{00}(a) = q$, $p_{01}(a) = p$.

V. La función de recompensa R para cada época como:

$$R(i, a) = 1,\ i \neq N;\ R(N, a) = 0.$$

Sea $v(0)_i$ la recompensa inicial para el estado i y sea $v(n)_i$ la recompensa esperada máxima para el problema en el n-ésimo periodo usando que se comenzó en el estado i.

Usando el principio de optimalidad de Bellman [45] (Sección 4.2 Finite-Horizon Policy Evaluation), se tiene que la siguiente recursión se cumple para $v(n)_i$

$$v(n)_i = \max_{a \in A} \left\{ R_i^a + \sum_{j=1}^{N} p_{ij}^a v(n-1)_j \right\}. \tag{5.6}$$

La última desigualdad se puede reescribir en notación matricial usando la terminología de las políticas como sigue:

$$v(n) = \max_{\pi \in \Pi} \{R(\pi) + P(\pi)v(n-1)\}, \qquad n \in \mathbb{N}, \tag{5.7}$$

donde $v(n)$ representa el vector con componentes $v(n)_i$, $i = 1, 2, \ldots N$, introduciendo una simple constante como en [29] podemos obtener que:

$$\begin{pmatrix} v(n) \\ 1 \end{pmatrix} = \underset{\pi \in \Pi}{max} \begin{pmatrix} P(\pi) & R(\pi) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v(n-1) \\ 1 \end{pmatrix}, \qquad n \in \mathbb{N}. \tag{5.8}$$

Sea $\mathcal{G} = (X, A, \{A(i)\}_{i \in X}, R, G, P)$ los elementos para este juegos son:

$X := \{0, 1, 2, 3, 4\}$; esto implica $N = 4$, $A(i) := \{A(0) = \{1\}, A(1) = \{1\}, A(2) = \{2\}, A(3) = \{1\}, A(4) = \{0\}\}$
Sea P y R como se definieron previamente. La función de recompensa G para el segundo jugador como sigue:

$$G(X_t) = \begin{cases} 356 & \text{if } t \leq 560, \\ 356 + \sum_{k=560}^{t} (1/2)^{(k-560)} & \text{if } t > 560, \end{cases}$$

con $p =$,2 y $q =$,8.

Para este ejemplo en particular se tiene dos políticas que puede ser representadas como sigue:

$$\pi_1 = (1, 1, 1) \text{ and } \pi_2 = (1, 2, 1).$$

En donde (1,1,1) representa que $A(1) = \{1\}$, $A(2) = \{1\}$ y $A(3) = \{1\}$ (respectivamente la misma notación para π_2). Las siguientes matrices representan las políticas π_1 y π_2 respectivamente, añadiendo una simple variable y la recompensa (para esta parte se están utilizando algunas ideas que fueron desarrolladas en [73] sobre las matrices no-negativas y la fórmula 5.8).

$$P_{\pi_1} := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ q & 0 & p & 0 & 0 & 1 \\ 0 & q & 0 & p & 0 & 1 \\ 0 & 0 & q & 0 & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P_{\pi_2} := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ q & 0 & p & 0 & 0 & 1 \\ q & 0 & 0 & 0 & p & 1 \\ 0 & 0 & q & 0 & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Podemos observar que estas dos matrices tienen un solo estado absorbente si observamos la sub-matriz P_{π_i} quitando la columna de la recompensa entonces nosotros podemos analizar el tiempo de absorción. Tomemos P_{π_1} y P_{π_2} restringidas (sub-matrices sin el estado absorbente), calculando los tiempos de absorción se obtiene que son 560 y 155 respectivamente. (Los tiempos de absorción se pueden calcular usando la matriz fundamental [49]).

Nota 5.3.1. *En esta tesis estamos usando la fórmula (5.8) para obtener el resultado mediante algoritmos computaciones. Sin embargo, el proceso de la Recompensas Total Esperada puede ser hecho a través de (5.6) siguiendo las instrucciones desarrolladas en [73].*

Lema 5.3.2. *Para el ejemplo 1 el par ($\pi^* = \pi_1$, $\tau^* = 560$) es un equilibrio de Nash para el juego $\mathcal{G}$.*

Demostración. Tomemos el par $(\pi_1, 560)$. La primera desigualdad para ser equilibrio de Nash se cumple $V(\pi_1, 560) \leq V(\pi_1, \tau)$ para todo $\tau \in \mathbb{N}$ esto debido a como $G(X_t)$ fue definido, obsérvese que π_1 es óptima para este MDP. $V(x, \pi_2, \tau^*) \leq V(x, \pi_1, \tau^*)$ con valores (152, 356) de la recompensa esperada respectivamente, supóngase que el estado inicial x fue el primer estado. Esto puede ser calculado para cualquier estado inicial y puede ser hecho mediante algunos cálculos, se puede observar que la desigualdad se cumple para el par (π_1, τ^*) por lo cual es un equilibrio de Nash para el juego estocástico transitorio de suma cero para dos jugadores. □

Nota 5.3.3. *El par $(\pi_2, 155)$ no es un equilibrio de Nash.*

Demostración. Es fácil checar que $V(\pi_2, 155) \leq V(\pi_2, \tau)$ esto debido a como $G(X_t)$ fue definida pero $V(x, \pi_1, 155) \leq V(x, \pi_2, 155)$ no se cumple porque la recompensa esperada para el TMCM sabiendo que el Jugador 2 va a detener el juego al tiempo $\tau^* = 155$ esto tiene como recompensa esperada (136, 98) respectivamente para π_1 y π_2 tomando en cuenta que el estado inicial fue el primero. Entonces el par $(\pi_2, 155)$ no es un Equilibrio de Nash. □

5.3.2. Ejemplos con Múltiples Equilibrios de Nash

Uno de los problemas más interesantes es que la mayoría de las veces el equilibrio de Nash no es único. Normalmente se presenta la pregunta de como elegir uno de ellos, con que criterio decidir. Para este problema nosotros proponemos el uso de la técnica de mínima varianza [54]; está técnica se justifica debido a que refleja el punto de vista del primer jugador. Con esta información nosotros podemos utilizar las siguientes

fórmulas que fueron introducidas en [58].

$$v_i^{(1)}(\pi,n) := E_i^{\pi}(\xi_n(\pi)) = E_i^{\pi}\left(\sum_{k=0}^{n-1} R_{X_k,X_{k+1}}\right),$$
$$v_i^{(2)}(\pi,n) := E_i^{\pi}(\xi_n(\pi))^2 = E_i^{\pi}\left(\sum_{k=0}^{n-1} R_{X_k,X_{k+1}}\right)^2,$$

esto nos lleva como resultado a las fórmulas de varianza promedio:

$$\begin{aligned}\sigma(\pi,n+1) = r^{(2)}(\pi) + P(\pi)\sigma(\pi,n) + 2P(\pi)\circ R\, v^1(\pi,n)\\ - \left[v^1(\pi,n+1)\right]^2 + P(\pi)\left[v^1(\pi,n)\right],\end{aligned} \tag{5.9}$$

(en donde el símbolo $\circ$ significa el producto de matrices de Hadamard)
en el cual para el caso de modelos transitorios, como nuestro modelos se transforma en:

$$\sigma(\pi) = r^2(f) + \tilde{P}(\pi)\sigma(\pi) + 2\tilde{P}(\pi)\circ R\, v^1(\pi) - \left[v^1(\pi)\right]^2 + \tilde{P}(\pi)\left[v^1(\pi)\right]^2. \tag{5.10}$$

en donde $\tilde{P}(\pi)$ denota una matriz transitoria en donde el radio espectral de P es menor que la unidad. Para modelos transitorios la fórmula previa puede se reescrita haciendo algunos cálculos algebraicos como:

$$\sigma(\pi) = \left[I - \tilde{P}(\pi)\right]^{-1}\left\{r^{(2)}(\pi) + 2\tilde{P}(\pi)\circ R\, v^1(\pi) - \left[v^1(\pi)\right]^2\right\}, \tag{5.11}$$

donde $r^2(\pi) = [P(\pi)\circ(R\circ R)] * e$.

Ejemplo 2

Específicamente, siguiendo las ideas del Ejemplo 1 usando la matriz extendida. Sea $N = 4$, la función de recompensa dada por: $R(i,a) = 1$, $i \neq N$ y $R(N,a) = 0$; pero con ley de transición como sigue: $p_{i,i+a}(a) = p$, $p_{i,i-a+1}(a) = q$, $p_{N0}(a) = 1$.

Es importante mencionar que haciendo algunos cálculos numéricos se pueden calcular las política y las respectivas recompensas totales. Estos cálculos nos hacen obtener 4 políticas $\pi_1 = (1,1,1,1)$, $\pi_2 = (1,1,2,1)$, $\pi_3 = (1,2,1,1)$, $\pi_4 = (1,2,2,1)$, pero solo 2 de ellas son óptimas con respecto al TMCM. Estas políticas son:

$$\pi_2 = (1,1,2,1),\ \pi_4 = (1,2,2,1).$$

En este caso $\pi_2 = (1,1,2,1)$, $A(0) = \{1\}$, $A(1) = \{1\}$, $A(2) = \{2\}$ y $A(3) = \{1\}$. Como en el ejemplo anterior, nosotros podemos observar que las matrices asociadas a estas políticas son:

$$P_{\pi_2} := \begin{pmatrix} q & p & 0 & 0 & 0 & 1\\ 0 & q & p & 0 & 0 & 1\\ 0 & q & 0 & 0 & p & 1\\ 0 & 0 & 0 & q & p & 1\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P_{\pi_4} := \begin{pmatrix} q & p & 0 & 0 & 0 & 1\\ q & 0 & 0 & p & 0 & 1\\ 0 & q & 0 & 0 & p & 1\\ 0 & 0 & 0 & q & p & 1\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Sea $\mathcal{G} = (X, A, \{A(x)\}_{x\in X}, R, G, P)$, usando los suposiciones anteriores y la función de recompensa G para el segundo jugador como sigue:

$$G(X_t) = \begin{cases} 35 & \text{if } t \leq 95, \\ 35 + \sum_{k=95}^{t} (1/2)^{k-95} & \text{if } t > 95, \end{cases}$$

con $p = ,2$ and $q = ,8$.

Lema 5.3.4. *Para el ejemplo 2 se tienen dos Equilibrios de Nash. Más aún π_4 tiene mínima varianza.*

Demostración. De la parte previa se tiene que $\pi_2 = (1, 1, 2, 1),\ \pi_4 = (1, 2, 2, 1)$, son óptimas para el TMCM y como resultado de los cálculos con respecto al tiempo de paro obtenemos que $\tau^* = 95$ es el valor del tiempo de paro para las dos políticas. Los pares $(\pi_2, 95)$ y $(\pi_4, 95)$ son Equilibrios de Nash.

Con esta información podemos utilizar las fórmulas de la varianza media 5.9 que para nuestro modelo es 5.10 pero para hacer los cálculos nosotros utilizamos 5.11. Haciendo algunos cálculos computacionales no muy complicados para las dos políticas optimas, obtenemos que para $\sigma(\pi_2)$ y $\sigma(\pi_4)$ los resultados son los siguientes:

	$\|\sigma(\pi_i)\|_2$
π_2	1028.5
π_4	988.7

.

en donde $\| \circ \|_2$ denota ($\|A\|_2 = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$).

Entonces, π_4 es la política con Mínima Varianza y que el par $(\pi_4, \tau^* = 95)$ es el Equilibrio de Nash con mínima varianza para el juego estocástico transitorio con tiempo de paro. □

Nota 5.3.5. *Un criterio posible respecto a la visión del jugador 1 es elegir el equilibrio de Nash con varianza mínima.*

5.3.3. Código Computacional

```
clear
clc
tic
format short
N=5
p=.8;
q=1-p;
w=[zeros(1,N-1)];
for i=1:N-1
    b=[i N-i];
```

```
    a(i)=min(b);
end

for j=1:length(a)
    x(j)={1:a(j)};
end

u=allcomb(x{:});
[u1,u2]=size(u);

n=length(a)+3;

for r=1:u1
A=zeros(n);
A(n-1,n-1)=0;
A(n-1,n-2)=1;
A(n,n-1)=1;
    for i=2:n-2
        for j=1:n
            if i-j==u(r,i-1);
            A(i,j)=q ;
            elseif i-j-1==-u(r,i-1);
            A(i,j)=p;
            end
        end
         A(i,n-1) = 1;
    end
   A=A(2:n,1:n-1);
    %A=sparse(A);
    B(r)={A};
end

%n=(N*(N+11))^2;
 m=95;
%m=(N+2)*round(N/2);
    for i=1:u1
    B1(i)={B{i}^m};
    end

    B3=zeros(N-1,r);
for j=1:N-1
    for i=1:u1
    B2(i)=max(B1{i}(j,N+1));
    end
```

```
    B3(j,:)=B2;
end

t1=[];
max=max(B3,[],2);
j=0;
for i=1:u1
    if max-B3(:,i)<.0000000001
        j=j+1;
        t1(j)=(i);
    end
end
u(t1,:)
toc




d=length(B)
for i=1:d;
    x{i}=normform(B{i});
end

for i=1:d;
    Q{i}=x{i}(1:n-3,1:n-3);
end
t=length(Q{1});

for i=1:d;
    NP{i} = (eye(t)-Q{i})^-1;
end

for i=1:d;
    RT{i} = diag(diag(NP{i}));
end

for i=1:d;
    M{i} = NP{i}.^2;
end

for i=1:d;
    Var{i} = NP{i}*(2*RT{i}-eye(t))-M{i};
end
```

```
u;
for  i=1:d;
     Var{i};
end


 norm(Var{1})
 norm(Var{3})
 norm(Var{2})
 norm(Var{4})
```

Apéndices

Nosotros introdujimos tres extras uno asociado el proceso estocástico que esta detrás de los juegos transitorios con tiempos de paro, también pusimos unas ideas que utilizamos en los ejemplos de los juegos transitorios con tiempos de paro relacionada con la estructura matricial. Para la última parte hablamos de como llegamos a la idea de la mezcla de los juegos y la idea de los tiempos de paro.

A. Definición del Proceso Estocástico asociado al Juego Estocástico con Estrategias Fijas [15]

Dado un juego estocástico con estado inicial i y estrategias π y ϕ para los jugadores , nosotros queremos definir un espacio de probabilidad y un proceso estocástico que modele el curso del juego. Nosotros entendemos que el curso del juego es una sucesión de estados y acciones a través del juego respecto a cada acción que toman los jugadores.

Consideresen los conjuntos A y B como las acciones disponibles para el jugador 1 y 2 respectivamente, como

$$\mathbb{A} = \bigcup_{i=1}^{N} A^i \quad \text{and} \quad \mathbb{B} = \bigcup_{i=1}^{N} B^i.$$

Sea $\Omega_1 = X,\ \Omega_2 = \mathbb{A} \times \mathbb{B},\ \Omega_3 = X,\ \Omega_4 = \mathbb{A} \times \mathbb{B}$, y así como sigue. El espacio muestral en el que queremos definir la probabilidad es:

$$\Omega = \prod_{k=1}^{\infty} \Omega_k,$$

como cada elemento $(i_0, (a_0, b_0), i_1, (a_1, b_1), \ldots) \in \Omega$ representa una historia de estados y acciones tomados por los jugadores, en el cual i_t representa que al tiempo t el estado era i y las acciones que fueron tomadas por los jugadores fueron (a_t, b_t). Es importante notar que para todo $t = 0, 1, 2, \ldots$

$$\mathbb{H}_t \subseteq \prod_{k=1}^{2t+1} \Omega_k.$$

Como los conjuntos Ω_k son de cardinalidad finita entonces se puede considerar la σ-álgebra sobre cada uno, $\mathcal{F}_k = \mathcal{P}(\Omega_k)$ Para cada $k = 1, 2, \ldots,$ y para cada $(\omega_1, \ldots, \omega_{k-1})$ la dinámica del juego define la siguiente probabilidad sobre $(\Omega_k, \mathcal{F}_k)$:

- ⋆ Si $k = 1$,

$$\mathbb{P}(\omega_1) = \begin{cases} 1 & \text{if } \omega_1 = i, \\ 0 & \text{othercase.} \end{cases}$$

- ⋆ Si k es par y $(\omega_1, \ldots, \omega_{k-1}) \in \mathbb{H}_{\frac{k}{2}-1}$

$$\mathbb{P}(\omega_k) = \begin{cases} \pi(a|\omega_1, \ldots, \omega_{k-1})\phi(b|\omega_1, \ldots, \omega_{k-1}) & \text{if } \omega_k = (a, b), \\ 0 & \text{othercase.} \end{cases}$$

- ⋆ Si k es par y $(\omega_1, \ldots \omega_{k-1}) \notin \mathbb{H}_{\frac{k}{2}-1}$ no importa la distribución de probabilidad porque es un conjunto de probabilidad 0.

- ⋆ Si k es impar y diferente de 1 y $\omega_{k-1} = (a, b)$ con $(a, b) \in A^{\omega_{k-2}} \times B^{\omega_{k-2}}$

$$\mathbb{P}(\omega_k) = P_{\omega_{k-2},a,b}.$$

- ⋆ Si k es impar y diferente de 1 y $\omega_{k-1} = (a, b)$ con $(a, b) \notin A^{\omega_{k-2}} \times B^{\omega_{k-2}}$ no importa otra vez como se defina por el mismo razonamiento que en el caso parecido.

Con todas las probabilidad previas, estamos en las hipótesis del teorema de Ionescu-Tulcea. Llamando $\mathbb{P}_{i,\pi,\phi}$ a la probabilidad dada por el teorema definido en $(\Omega, \mathcal{F})$ en el cual $\mathcal{F}$ es la σ-álgebra dado por los cilindros.

En el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P}_{i,\pi,\phi})$ considérese el proceso estocástico

$$\{X_t\}_{t=0,1,\ldots}, \quad \{A_t\}_{t=0,1,\ldots}, \quad \{B_t\}_{t=0,1,\ldots},$$

el cual representa el estado y las acciones usadas por los jugadores 1 y 2 respectivamente al tiempo t; definido como se espera si $\omega = (i_0, (a_0, b_0), i_1, (a_1, b_1), \ldots) \in \Omega$ entonces:

$$X_t(\omega) = i_t, \quad A_t(\omega) = a_t, \quad B_t(\omega) = b_t.$$

También considérese los elementos aleatorios $H_t : t = 0, 1, \ldots$ de las historias del proceso hasta el tiempo t, así que H_t toma valores en el $\prod_{k=1}^{2t+1} \Omega_k$ y

$$H_t(\omega) = (i_0, a_0, b_0, i_1, a_1, b_1, \ldots, i_t).$$

Por como se definió la probabilidad $\mathbb{P}_{i,\pi,\phi}$ es fácil checar que:

- ⋆ El juego comienza en el estado i esto significa que

$$\mathbb{P}_{i,\pi,\phi}(X_0 = i) = 1.$$

- ⋆ Las variables aleatorias H_t toman valores en $\mathbb{H}_t$ con probabilidad 1.

- ⋆ Las acciones tomadas por los jugadores dependen de las estrategias e historias del juego y son independientes

$$\mathbb{P}_{i,\pi,\phi}(A_t = a_t | H_t = h_t) = \pi(a_t|h_t)$$

$$\mathbb{P}_{i,\pi,\phi}(B_t = a_t | H_t = h_t) = \phi(b_t|h_t)$$

$$\mathbb{P}_{i,\pi,\phi}(A_t = a_t, B_t = a_t | H_t = h_t) = \pi(a_t|h_t)\phi(b_t|h_t).$$

⋆ La transición entre los estados solo dependen del último estado y de las últimas acciones.

$$\mathbb{P}_{i,\pi,\phi}(X_{t+1} = i_{t+1}|H_t = h_t, A_t = a_t, B_t = b_t) = \mathbb{P}_{i,\pi,\phi}(i_{t+1}),$$

en el cual i_t es el último estado de la historia h_t.

B. Propiedades de las Matrices No Negativas.

Una cosa importante es observar que para el análisis de la matriz transitoria se necesita observar las siguientes cosas:

Sea H un conjunto de matrices no negativas $k \times m$ con $(k, m \in N)$, sea P_i que denote la i-ésima fila de la matriz $P \in H$. H tiene la propiedad producto si para cada subconjunto V de $\{1, 2, \ldots, k\}$ y si para cada par de matrices $P(1), P(2) \in H$ lo siguiente se cumple: La matriz $P(3)$, definida por

$$P(3)_i := \begin{cases} P(1)_i & \text{if} \quad i \in V \\ P(2)_i & \text{if} \quad i \in \{1, 2, \ldots, k\} \setminus V, \end{cases}$$

es también un elemento de H.

Sea $\sigma(P)$ que denota el radio espectral de P $(P \in H)$ y sea
$\hat{\sigma} := \max\{\sigma(P)|P \in H\}$. Entonces existe una matriz $\hat{P} \in H$, una partición $\{D(0), D(1), \ldots, D(\nu)\}$ de X, y un conjunto de vectores semipositivos
$\{w(1), w(2), \ldots, w(v)\}$ tales que:

$$\max_{P \in H} Pw(v) = \hat{P}w(v) = \hat{\sigma}w(v),$$

$$\max_{P \in H} Pw(k) = \hat{P}w(k) = \hat{\sigma}w(k) + w(k+1), \quad k = \nu - 1, \ldots, 2, 1.$$

Si $C \subset X$, entonces P^C es la restricción de P a $C \times C$. Similarmente, v^C es la restricción del vector (columna) v a C. Si $\{D(0), D(1), \ldots, D(n)\}$ denota una partición del espacio de estados X, entonces normalmente se escribe $P^{(k,l)}$ para la restricción de P a $D(k) \times D(l)$, $k, l = 0, 1, \ldots, n$. Nótese que $P^{(k,k)} = P^{D(k)}, k = 0, 1, \ldots, n$.

Siguiendo el Rothblum ([51] y [70]) se dice que el estado i tiene acceso al estado j (o que el estado j tiene acceso desde el estado i) si existe un entero no negativo n tal que ij-ésima entrada de P^n es positiva. La definición de accesibilidad refleja la idea que la configuración de P puede se representada por una gráfica dirigida. Como consecuencia, se considera P como una función no negativa real valuado definida en $X \times X$, y no como un operador lineal de $R^n \times R^n$. P se dice que es irreducible si para cualquier dos estados tienen acceso entre ellos, para cualquier otro caso se dice que P es reducible. Más aún, si D es un subconjunto propio de X, la restricción P^D de P a $D \times D$ se llama el menor principal de P. Una clase de P es un subconjunto C de X tal que P^C es irreducible y tal que C no puede ser engrandecida sin perder la irreducibilidad. A C se le dice un clase básica si $\sigma(P^C) = \sigma(P)$, de cualquier otra forma se le dice no básica (para la cual se tiene $\sigma(P^C) < \sigma(P)$). Entonces existe una partición del espacio de estados X en clases, $C(1), C(2), \ldots C(n)$. Si $P^{(i,j)}$

denota la restricción de P a $C(i) \times C(j), i, j = 1, \ldots, n$ entonces (posiblemente después de una permutación de los estados) P puede ser escrita de la siguiente forma:

$$P := \begin{pmatrix} P^{(1,1)} & P^{(1,2)} & \ldots & P^{(1,n)} \\ & P^{(2,2)} & \ldots & P^{(2,n)} \\ & & \ddots & \vdots \\ & & & P^{(n,n)} \end{pmatrix},$$

con $P^{(i,j)} = 0$ for $i > j, i, j = 1, \ldots, n$. Aquí las clases pueden ser parcialmente ordenadas mediante las relaciones de accesibilidad. Nosotros podemos hablar de accesibilidad de (desde) la clase si existe una acceso de (desde) (o, equivalentemente,) para cualquier estado en esas clase. Un clase C se le dice final si C no tiene acceso a cualquier otra clase. Una clase C se le dice inicial si ninguna otra clase tiene acceso a C. La existencia de eigenvectores estrictamente positivos asociados con el radio espectral $\sigma(P)$ de una matriz cuadrada no negativa P depende fuertemente en la relación entre las clases básicas y no básicas.

Lema 5.3.6. *Una matriz cuadrada no negativa P posee un eigenvector estrictamente positivo derecho (izquierdo) si y solo si las clases básicas son precisamente son las clases finales (iniciales).*

Una cadena de clases de P es una colección de clases $\{C(1), \ldots, C(n)\}$ tal que $p_{i_k, j_k} > 0$ para algún par de estados (i_k, j_k) . $i_k \in C(k)$, $j_k \in C(k+1)$, $k = 1, 2, \ldots n-1$.

Se dice que la cadena comienza con $C(1)$ y termina en $C(n)$. El tamaño de la cadena es el número de clases básicas que contiene. El peso (profundidad) de una clase C de P es la cadena más larga que termina (comienza) en C. El grado $\nu(P)$ de P es el tamaño de la cadena más larga. Claramente la profundidad de la clase con respecto a P corresponde al peso con respecto a P^T. Para este trabajo la clasificación de las clases se realiza mediante la profundidad.

Lema 5.3.7. *Sea P con radio espectral σ y grado ν. Existe una partición $\{D(\nu), D(\nu-1), \ldots, D(1), D(0)\}$ del espacio de estados X tal que $D(k)$ es la unión de todas clases con profundidad k, para $k = 0, 1, \ldots$. En particular, si $P^{(k,l)}$ denota la restricción de P a $D(k) \times D(l)$, entonces $P^{(k,l)} = 0$ para $k < l, (k, l = 0, 1, \ldots, \nu)$. De esto, posiblemente después de una permutación de estados se puede reescribir como:*

$$P := \begin{pmatrix} P^{(\nu,\nu)} & P^{(\nu,\nu-1)} & \ldots & P^{(\nu,1)} & P^{(\nu,0)} \\ & P^{(\nu-1,\nu-1)} & \ldots & P^{(\nu-1,1)} & P^{(\nu-1,0)} \\ & & \ddots & \vdots & \vdots \\ & & & P^{(1,1)} & P^{(1,0)} \\ & & & & P^{(0,0)} \end{pmatrix}.$$

Se tiene que $\sigma(P^{(k,k)}) = \sigma$ para $k = 1, 2, \ldots \nu$ y $\sigma(P^{(0,0)}) < \sigma$ (si $D(0)$ no es vacía). Más aún, existe un vector $u^{(k)} > 0$ tal que

$$P^{(k,k)} u^{(k)} = \sigma u^{(k)}, \quad k = 1, 2, \ldots, \nu.$$

Lema 5.3.8. *Sea cada $P \in H$ irreducible, entonces existe $\hat{P} \in H$, con radio espectral $\hat{\sigma}$ y un vector estrictamente positivo $\hat{u}$ tal que*

$$\hat{P}\hat{u} = \max_{P \in H} P\hat{u} = \hat{\sigma}\hat{u}.$$

Teorema 5.3.9. *Sea* $\hat{\sigma} := \max\{\sigma(P)|P \in H\}$, *y sea* $\{D(0), D(1), \ldots, D(\nu)\}$ *la partición principal de* X *con respecto a* H. *Entonces existen vectores semipositivos* $\{w(\nu), \ldots, w(2), w(1)\}$, *tales que*

$$\max_{P\in H} Pw(\nu) = \hat{\sigma}w(\nu),$$

$$\max_{P\in H_{l+1}} Pw(l) = \hat{\sigma}w(l) + w(l+1), \quad l = \nu - 1, \ldots, 2, 1,$$

donde

$$H_\nu := \{P|P \in H, Pw(\nu) = \hat{\sigma}w(v)\},$$

$$H_l := \{P|P \in H_{l+1}, Pw(l) = \hat{\sigma}w(l) + w(l+1)\}, \quad l = \nu - 1, \ldots, 2, 1.$$

Para $l = \nu, \nu - 1, \ldots, 2, 1$ *se tiene*

$$w(l)_i > 0, \quad i \in \bigcup_{k=l}^{\nu} D(k),$$

$$w(l)_i = 0, \quad i \in \bigcup_{k=0}^{l-1} D(k).$$

Teorema 5.3.10. *Sea* K *un conjunto finito de matrices no negativas con la propiedad producto entonces existe un entero* r *y una partición del espacio de estado* X, *llamémosla* $\{I(1), \ldots, I(R)\}$, *que cumple con las siguientes cosas:*

- *Sea* $P^{(k,l)}$ *la restricción a* $I(k) \times I(l)$.
  entonces $P^{(k,l)} = 0$ *si* $k > l$ $(k, l = 1, \ldots, r)$ *para cada* $P \in K$.
- *Existe una matriz* $\hat{P} \in K$ *y vectores estrictamente positivos* $\hat{u}^{(k)}$ *definidos en* $I(k)$ *tales que:*

$$\hat{P}^{(k,k)}\hat{u}^{(k)} = \max_{P\in K} P^{(k,k)}\hat{u}^{(k)} = \hat{\sigma}_k\hat{u}^{(k)}, \ k = 1, 2, \ldots, r, \tag{5.12}$$

  en donde $\hat{\sigma}_k := \sigma(\hat{P}^{(k,k)}), \ k = 1, 2, \ldots, r$.
  Para $k \leq l$ *se tiene que* $\hat{\sigma}_k \geq \hat{\sigma}_l$; *la igualdad se cumple cuando cada estado de* $I(k)$ *tiene acceso a algún estado* $I(l)$ *bajo* $\hat{P}$.
- *Sea* $x(0) > \underline{0}$ *y* $x(n)$ *como anteriormente; para cada* $k \in 1, 2, \ldots, r$ *sea* t_k *un entero positivo definido como:*

$$t_k := \begin{cases} min\{j|j > 0, \hat{\sigma}_{k+j} < \sigma_k\} & \text{if } j \text{ exist,} \\ r - k + 1 & \text{another case.} \end{cases}$$

  entonces deben existir constantes positivas c_1 *y* c_2 *tales que:*

$$c_1 u_i^k \leq \left(\frac{n}{t_k - 1}\right)^{-1} \hat{\sigma}_k^{-n} x(n)_i \leq c_2 u_i^{(k)},$$

  para cada $i \in I(k), \ k = 1, 2, \ldots, r$ *y* $n \in \mathbb{N}$.

C. Tiempos de Paro

Sea $G = (G_n)_{n\geq 0}$, una sucesión de variables aleatorias definidas en un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, P)$. Se interpreta G_n como la ganancia obtenida si la observación de G es detenida al

tiempo n. Se supone que G se a adapta a la filtración $(\mathcal{F}_n)_{n\geq 0}$ en el sentido que para cada G_n es $\mathcal{F}_n$-medible. Recuérdese que cada $\mathcal{F}_n$ es una σ-álgebra de subconjuntos de Ω tal que $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \mathcal{F}$. Típicamente $(\mathcal{F}_n)_{n\geq 0}$ coincide con la filtración natural $(\mathcal{F}_n^G)_{n\geq 0}$ pero en general podría ser mas grande. Se interpreta $\mathcal{F}_n$ como la información disponible al tiempo n. Todas las decisiones en búsqueda del óptimo paro al tiempo n deben estar basadas en esta información solamente (no se permite la anticipación) [17].

Definición 5.3.11. *Una variable aleatoria $\tau : \Omega \to \{0, 1, \ldots, \infty\}$ se dice un tiempo de Markov si $\{\tau \leq n\} \in \mathcal{F}_n$ para todo $n \geq 0$. Un tiempo de Markov se le dice un tiempo de paro si $\tau < \infty$ P-a.s.*

La familia de todos los tiempos de paro será denotado por $\mathcal{M}$, y la familia de todos los tiempos de Markov será denotada por $\overline{\mathcal{M}}$. Las siguiente subfamilias de $\mathcal{M}$ serán usadas.

$$\mathcal{M}_n^N = \{\tau \in \mathcal{M} : n \leq \tau \leq N\} \tag{5.13}$$

donde $0 \leq n \leq N$. Por simplicidad, denotaremos como $\mathcal{M}^N = \mathcal{M}_0^N$ y $\mathcal{M}_n = \mathcal{M}_n^\infty$.

El *problema de paro óptimo* que será estudiado busca resolver

$$V_* = \sup_\tau \mathsf{E}\, G_\tau \tag{5.14}$$

donde el supremo se toma sobre la familia de los tiempos de paro. Nótese que (5.14) envuelve dos tareas:

- Computar *el valor de la función* V_* tan explicitamente como sea posible
- Exhibir un *tiempo de paro óptimo* τ_* para el cual el supremo es alcanzado.

Para asegurarnos de la existencia de $\mathsf{E}\, G_\tau$ en (5.14) se necesitan imponer algunas condiciones adicionales sobre G y τ. Si la siguiente condición se cumple (con $G_N \equiv 0$ entonces $N = \infty$):

$$\mathsf{E}\,(sup_{n\leq k\leq N} \mid G_k \mid) < \infty \tag{5.15}$$

entonces $\mathsf{E}\, G_\tau$ esta bien definida para todo $\tau \in \mathcal{M}_n^N$. Aunque para muchos de los resultados siguientes es posible debilitar esta condición y reemplazar $\mid G_k \mid$ por G_k^- o G_k^+ (o más aún solo considerar los casos para los τ para los cuales $\mathsf{E}\, G_\tau$ esta bien definida) por simplicidad se asumirá a largo de este trabajo que (5.15) se satisface. Una inspección más cuidadosa de las pruebas revelará que la condición (5.15) puede ser relajada.

Con la subfamilia de tiempos de paro $\mathcal{M}_n^N$ introducida en (5.13) arriba nosotras asociaremos la siguiente función de valor:

$$V_n^N = \sup_{\tau\in\mathcal{M}_n^N} \mathsf{E}\, G_\tau \tag{5.16}$$

donde $0 \leq n \leq N$. Otra vez, por simplicidad, sea $V^N = V_0^N$ y $V_n = V_n^\infty$. Igualmente, sea $V = V_0^\infty$ donde el supremo es tomado sobre todos los $\tau \in \mathcal{M}$. El objetivo principal de esta parte esta basado en los problemas con Horizonte Finito. Considérese el problema de tiempo de paro óptimo (5.16) donde $N < \infty$. Recuérdese que (5.16) se lee más explicitamente como sigue a continuación:

$$V_n^N = \sup_{n\leq\tau\leq N} \mathsf{E}\, G_\tau \tag{5.17}$$

donde τ es un tiempo de paro y $0 \leq n \leq N$. Para resolver el problema nosotros podemos dejar que el tiempo retroceda y proceder recursivamente.

Teorema 5.3.12. *(Horizonte Finito) Considérese el problema de paro óptimo (5.17) suponiendo que la condición (5.15) se cumple. Entonces para todos* $0 \leq n \leq N$ *se tiene que:*

$$S_n^N \geq \mathsf{E}\left(G_\tau | \mathcal{F}_n\right) \text{ para cada } \tau \in \mathcal{M}_n^N, \tag{5.18}$$

$$S_n^N = \mathsf{E}\left(G_{\tau_n^N} | \mathcal{F}_n\right). \tag{5.19}$$

Más aún, si $0 \leq n \leq N$ *esta dado y fijo, entonces se tiene que: El tiempo de paro* τ_n^N *es óptimo en (5.17) Si* τ_* *es un tiempo de paro óptimo en (5.17) entonces* $\tau_{N_n} \leq \tau_*$ P*-a.s La sucesión* $(S_k^N)_{n \leq k \leq N}$ *es la más pequeña supermartingala la cual domina* $(G_k)_{n \leq k \leq N}$*. La sucesión de paro* $(S^N_{k \wedge \tau_n^N})$ *es una martingala.*

La prueba de este teorema puede ser encontrada de la página 4 en [43].

Capítulo 6

Conclusiones y Problemas Abiertos

6.1. Conclusiones

Podemos resumir este trabajo a través de dos variantes, una son los Juegos Estocásticos Sensibles al Riesgo y la segunda los Juegos Estocásticos Transitorios con Tiempos de Paro en donde se buscaron las condiciones para poder obtener los equilibrios de Nash correspondientes.

- Al principio de este trabajo dimos una pequeña introducción de los antecedentes relacionados con los juegos y cómo se da el salto de los Procesos de Decisión de Markov al área de los juegos mediante los Juegos Estocásticos.
- En la primera parte obtuvimos cotas para el valor óptimo en los Juegos Estocásticos Sensibles al Riesgo, así mismo se mostró que estas cotas estaban relacionadas con los eigenvalores de Perron y su respectivo eigenvector que se pudo elegir estrictamente positivo debido a que las matrices de transición son no negativas. Para redondear esta parte se mostraron algunos ejemplos para ilustrar el proceso de la obtención de las cotas.
- Una vez que lo anterior estaba finalizado nos enfocamos en analizar el problema de los Juegos Estocásticos con Tiempos de Paro con Recompensa Total.

  En resumen, primero se encontraron condiciones que aseguran que la Recompensa Total sea finita, debido a que el primer jugador tenía este tipo de función de recompensa y después proponer una estructura para el juego, lo cual fue el hecho mediante la transitoriedad del juego. Una vez teniendo la formulación del juego se tuvo que buscar la solución, así que lo que se hizo fue buscar un operador contractivo en un espacio de Banach apropiado; cabe resaltar que para la obtención de este operador cada vez el juego tenía que ir siendo más específico debido a que por el tipo de recompensa usada no era tan fácil encontrar una contracción natural, en comparación con el caso de la Recompensa Descontada.

  Es importante resaltar que el modo en el que fue abordado el problema de encontrar el Equilibrio

de Nash en los Juegos Estocásticos con Tiempo de Paro con Recompensa Total no sólo permite observar la solución sino además nos dio pie a proponer un método de selección en el caso de que existan múltiples equilibrios. En nuestro caso usamos el método de la mínima varianza, el cual refleja el punto de vista del primer jugador para la elección de la estrategia óptima.

Una posible aplicación sería en el campo de la industria sobre el tiempo de duración de una pieza y los costos de reparación.

6.2. Problemas Abiertos

Durante el proceso de este trabajo notamos que hay distintas opciones para mejorar y ampliar éste. Las posibles áreas para una investigación a futuro las enumeraremos a continuación.

- Generalizar los Juegos Estocásticos Sensibles al Riesgo para el caso de varios jugadores.
- Crear un programa para calcular Equilibrios de Nash en el caso de varios jugadores, para los Juegos Estocásticos Sensibles al Riesgo.
- Mejorar las cotas superiores e inferiores para encontrar el valor óptimo presentadas en la Sección 2.4.1 para los Juegos Estocásticos Sensibles al Riesgo.
- Considerar recompensas negativas en el caso de los problemas Sensibles al Riesgo.
- Para los Juegos Estocásticos Transitorios con Tiempos de Paro, extenderlos al caso con varios jugadores.
- Respecto a la Recompensa Terminal del segundo jugador en los Juegos Estocásticos Transitorios con Tiempos de Paro, debilitar las condiciones dadas en la definición.
- Será interesante introducir para los Juegos Estocásticos Transitorios con Tiempos de Paro la sensibilidad al riesgo.
- Otro punto importante será desarrollar un tipo diferente de criterios para la selección de Equilibrios. No quedarnos sólo con el criterio de la Mínima Varianza, sino a lo mejor introducir un criterio promedio o generar alguno que considere el punto de vista del segundo jugador.
- Buscar una forma de abordar los Juegos Estocásticos Transitorios con Tiempos de Paro para los cuales se modifique el espacio de acciones a un espacio continuo y compacto.
- Un problema interesante será preguntarse que pasa cuando se aceptan recompensas negativas para el caso de los Juegos Estocásticos Transitorios.

Bibliografía

[1] Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, F., Ghosh, M. K., and Marcus, S. I.: Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization* (1993).

[2] Arriaga. S.: *Problemas de Control de Markov con Recompensa Total Esperada en Espacio Finitos: Casos Neutral y Sensible al Riesgo*. Tesis de Maestría en Matemáticas UAM-Iztapalapa, 2008

[3] Ash, R.: *Real Analysis and Probability*. Academic Press, 1972.

[4] Bellman, R.: A Markovian decision process, *Journal of Mathemathics and Mechanics* **6** (1957), 679-684.

[5] Bellman, R.: *Dynamic Programming*. Princeton University Press, Princeton. Princeton 1957.

[6] Bellman, R.: On a quasi-linear equation, *Canad. J. of Math.* **8** (1956), 198–202.

[7] Binmore, K.: *Teoría de Juegos*. McGraw-Hill, 1994.

[8] Blackwell, D.: Discrete dynamic programming, *Annals of Mathematical Statistics* **33** (1962), 719-726.

[9] Blackwell, D.: Discounted dynamic programming, *Annals of Mathematical Statistics* **36** (1965), 226-725.

[10] Cavazos-Cadena, R. and Hernández-Hernández, D.: Nash equilibria in a class of Markov stopping games, *Kybernetika* **48** (2012), 1027–1044.

[11] Cavazos-Cadena. R and Montes de Oca. R.: Optimal stationary policies in controlled Markov chains with expected total-reward criterion. Research Report. No. 04.0405.I.01.010.99, UAM-Iztapalapa, 1999.

[12] Cavazos-Cadena, R. and Montes de Oca, R.: Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards. *Applicationes Mathematicae* (Warsaw), **27**, 2, (2000), 167–185.

[13] Cavazos-Cadena, R. and Montes de Oca, R.: Stationary optimal policies in a class of multichain positive dynamic programs with finite state space and risk-sensitive criterion. *Applicationes Mathematicae* (Warsaw), **28**, 1, (2001), 93–109.

[14] Cinlar, E.: *Introduction to Stochastic Processes*. Dover, 1975.

[15] Crocce, F.: *Juegos estocásticos transitorios y aplicaciones*. Tesis de Maestría. Universidad de la República de Uruguay. 2009.

[16] Dugundji, J. and Granas, J.: *Fixed Point Theory*. Springer, 2003.

[17] Dynkin, E.B.: The optimum choice for the instances for stopping Markov process. *Soviet Mathematics Doklady.* **4**, (1963), 627–629.

[18] Eeckhoudt, L. Gollier, C. Schlesinger, H.: *Economic and Financial Decision under Risk.* Princeston University Press, 2005.

[19] Etessami, K. Wojtczak, D. and Yannakakis. M.: Recursive stochastic games with positive rewards. *In Proceeding 35th ICALP.* (2008), 711–723.

[20] Federgruen, A.: Successive approximation methods in undiscounted stochastic games, *Operations Research* **28** (1984), 794–809.

[21] Filar, J.A. and Raghavan, T.E.S.: A matrix game solution of the single-controller stochastic games, *Mathematics of Operations Research* **9** (1984), 356–362.

[22] Filar, J.A.: On stationary equilibria of a single-controller stochastic games, *Mathematical Programming* **30** (1984), 315–325.

[23] Filar, J.A. and Raghavan, T.E.S.: A matrix game solution of the single-controller stochastic games, *Mathematics of Operations Research.* **9** (1984), 356–362.

[24] Filar, J.A. and Vrieze, O.J.: *Competitive Markov Decision Processes.* Springer Verlag, Berlin, 1996.

[25] Gantmakher, F. R.: *The Theory of Matrices.* Chelsea, London, 1959.

[26] Gillette, D.: Stochastic games with zero stop probabilities. In: Contribution to the theory of games, Vol. III (Dresher, M., A.W. Tucker, and P.Wolfe, eds.) *Annals of Mathematics Studies.* **39** (1957), 179–188.

[27] Hoffman, A.J. and Karp, R.M.: On non-terminating stochastic games, *Management Science* **12** (1966), 359–370.

[28] Howard, R. A.: *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, Mass., 1960.

[29] Howard, R. A. and Matheson, J.: Risk-sensitive Markov decision processes, *Management Science* **23** (1972), 356–369.

[30] Kolokoltsov, V.N. and Malafayev, O.A.: *Understanding Game Theory.* World Scientific, Singapore, 2010.

[31] Leyton, K. and Shoham, Y.: *Essentials of Game Theory.* Morgan and Claypool Publishers, USA, 2008.

[32] Liggett, T.M. and Lippman, S.A.: Stochastic games with perfect information and time average payoff, *SIAM Review* **11** (1969), 604–607.

[33] Thomas, L. C.: Connectedness conditions used in finite state Markov decision processes, *Journal of Mathematical Analysis and Applications* **68** (1979), 548–556.

[34] Thomas, L. C.: Connectedness conditions for denumerable state Markov decision processes, *Recent Developments in Markov Decision Processes (ed. R. Hartley, L. C. Thomas and D. J. White).* Academic Press, New York (1980), 181–204, 1980.

[35] Mandl, P.: Estimation and control in Markov chains. *Advances in Applied Probability.* (1974).

[36] Mandl, P.: On the variance in controlled Markov chains, *Kybernetika* **7** (1971), 1–12.

[37] Martínez. V.: Bi-personal Stochastic Transient Markov Games with Stopping Times and Total Reward Criterion, *Kybernetika* **57** (2021), 1–14.

[38] Martínez. V.: *Reducibilidad en Matrices Finitas No Negativas con Aplicaciones a Economía y Control Estocástico.* Tesis de Maestría en Matemáticas UAM-Iztapalapa, 2014.

[39] Martínez. V. and Sladký, K.: Risk-sensitive optimality in Markov games. *Conference Proceeding Mathematical Methods in Economics. Hradec Králové. CZ.* (2017), 684-689.

[40] Nash, J.: Equilibrium points in n-person games, *Proceedings of the National Academy of Sciences U.S.A.* **36** (1950), 48–49.

[41] Neyman, A. and Sorin, S.: *Stochastic Games and Applications.* NATO. New York, 1999.

[42] Nowak, A.S. and Raghavan, T.E.S.: A finite step algorithm via a bimatrix to a single controller non-zero sum stochastic game, *Mathematical Programming* **59** (1993), 249–259.

[43] Peskir, G. and Shiryaev, A.: *Optimal Stopping and Free-Boundary Problems.* Birkhauser, Boston, 2010.

[44] Pollatschek, M.A. and Avi-Itzhak, B.: Algorithms for stochastic games with geometrical interpretation, *Management Science* **15** (1969), 399–415.

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

[46] Raghavan, T.E.S., Tijs, S.H., and Vrieze, O.J.: A matrix game solution of the single-controller stochastic game, *Journal of Optimization Theory and Applications* **47**(1985), 451–464.

[47] Raghavan, T.E.S. and Filar, J.A.: Algorithms for stochastic games a survey, *Methods and Models of Operations Research ZOR* **35** (1991), 437–472.

[48] Raghavan, T. E. S., Tijs, S. H., and Vrieze, O. J.: On stochastic games with additive reward and transition structure. *Journal of Optimization Theory and Applications* (1985).

[49] Ross, S.: *Introduction to Probability Models.* Ninth edition. Elsevier, USA, 2007.

[50] Ross, S. M.: *Introduction to Stochastic Dynamic Programming.* Academic Press, New York, 1983.

[51] Rothblum, V. G.: Algebraic eigenspaces of non-negative matrices, *Linear Algebra and its applications* **12** (1975), 201–292.

[52] Shapley, L.S.: Stochastic games, *Proceedings of the National Academy of Sciences U.S.A.* **39** (1953), 1095–1100.

[53] Shiryaev, A.: *Optimal Stopping Rules.* Springer, New York, 1978.

[54] Sitar, M., and Sladký, K.: Algorithmic procedures for mean variance optimality in Markov decision chains. *Operations Research Proceedings.* (2005).

[55] Sladký, K.: Bounds on discrete dynamic programming recursions I: Models with non negative matrices. *Kybernetika* **16**, (1980), 6.

[56] Sladký, K.: On the set of optimal controls for Markov chains with rewards. *Kybernetika* **10**, 4, (1974), 350–367.

[57] Sladký, K.: On the existence of stationary optimal policies in discrete dynamic programming. *Kybernetika* **17**, 6, 489–513.

[58] Sladký, K.: Second order optimality in transient and discounted Markov decision chains. *Kybernetika* **53**, 6, 1086–1099.

[59] Sladký, K. and Montes-de-Oca R.: Risk-sensitive average optimality in Markov decision chains. *Operations Research Proceedings 2007* . Eds. J. Kalcsics and S. Níkel , 69–74, Springer, 2008.

[60] Thuijsman, F.: *Optimality and Equilibria in Stochastic Games*. Mathematical Centre Tracts, Amsterdam, 1992.

[61] Van der Wal, J.: Discounted Markov games: successive approximations and stopping times, *Internat. J. Game Theory* **6** (1977), 11–22.

[62] Van der Wal, J.: Discounted Markov games: generalized policy iteration method, *Journal of Optimization Theory and Applications* **25** (1978) 125–138.

[63] Van der Wal, J.: Succesive approximations for average reward Markov games. Memorandum COSOR 77-10, Eidhoven, April 1977. *International Journal of Game Theory* **9** (1980), 213–224.

[64] Van der Wal, J.: *Stochastic Dynamic Programming*. Mathematical Centre Tracts, Amsterdam, 1981.

[65] Von Neumann, J. and Morgenstern, O.: *Theory of Games and Economic Behavior*. Springer. London, 1944.

[66] Vrieze, O. J.: *Stochastic Games with Finite State and Action Spaces*. Mathematical Centre Tracts, Amsterdam, 1987.

[67] Vrieze, O.J., Tijs, S.H., Raghavan, T.E.S. and Filar, J.A.: A finite algorithm for the switching control stochastic game, *OR Spektrum* **5** (1983), 5–24.

[68] Wald, A.: *Sequential Analysis*. Wiley, New York. Chapman and Hall, London, 1947.

[69] Webb, N.: *Game Theory Decisions. Interaction and Evolution*. Springer. London, 2006.

[70] Whittle, P.: *Optimization over Time*. Dynamic Programming and Stochastic Control, Volume II. Wiley, Chichester. 1983

[71] Zachrisson, L.: *Markov games. In advances in game theory* (M.Dresher, L.S. Shapley amd A.W. Tucker, eds). Princeston University Press, Princeton 1964, 211-253.

[72] Zapata P. : *Economía, Política y otros juegos. Una introducción a los juegos no cooperativos.* Las Prensas de Ciencias, 2007.

[73] Zijm, W. H. M.: *Nonnegative Matrices in Dynamic Programming*. Mathematisch Centrum, Amsterdam, 1983.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE DISERTACIÓN PÚBLICA

No. 00077
Matrícula: 2143805683

JUEGOS ESTOCÁSTICOS TRANSITORIOS CON RECOMPENSA TOTAL.

Con base en la Legislación de la Universidad Autónoma Metropolitana, en la Ciudad de México se presentaron a las 17:00 horas del día 14 del mes de septiembre del año 2021 POR VÍA REMOTA ELECTRÓNICA, los suscritos miembros del jurado designado por la Comisión del Posgrado:

DR. ROLANDO CAVAZOS CADENA
DR. EVGUENI ILICH GORDIENKO
DR. VICTOR HUGO VAZQUEZ GUEVARA
DR. JOSE RAUL MONTES DE OCA MACHORRO
DR. JULIO CESAR GARCIA CORTE

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron a la presentación de la Disertación Pública cuya denominación aparece al margen, para la obtención del grado de:

DOCTOR EN CIENCIAS (MATEMATICAS)

DE: VICTOR MANUEL MARTINEZ CORTES

y de acuerdo con el artículo 78 fracción IV del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

VICTOR MANUEL MARTINEZ CORTES
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTE

DR. ROLANDO CAVAZOS CADENA

VOCAL

DR. EVGUENI ILICH GORDIENKO

VOCAL

DR. VICTOR HUGO VAZQUEZ GUEVARA

VOCAL

DR. JOSE RAUL MONTES DE OCA MACHORRO

SECRETARIO

DR. JULIO CESAR GARCIA CORTE

El presente documento cuenta con la firma –autógrafa, escaneada o digital, según corresponda- del funcionario universitario competente, que certifica que las firmas que aparecen en esta acta – Temporal, digital o dictamen- son auténticas y las mismas que usan los c.c. profesores mencionados en ella