Problemas de Control de Markov con Recompensa Total Esperada en Espacios Finitos

Casos Neutral y Sensible al Riesgo

Presenta:
María Soledad Arriaga
Asesor:
Dr. Raúl Montes-de-Oca
Departamento de Matemáticas
Universidad Autónoma Metropolitana-Iztapalapa

17 de agosto de 2008

Índice general

Agradecimientos

A Rojo por todo su amor y comprensión.

A Raúl Montes de Oca por su generosidad y su infinita paciencia.

A todos mis maestros.

A mis compañeros por su solidaridad, sus risas, sus consejos,... su ayuda.

A Evgueni Gordienko y Juan González por sus sugerencias sobre este trabajo.

A mi madre por traerme a este mundo.

A mi hermano Adrián porque sin su sacrificio no estaría yo aquí.

A Deva por enseñarme un nuevo camino.

A Helena que se ha convertido en la luz de mi vida.

A Luis por enseñarme que aún hay mucho que hacer.

Al Consejo Nacional de Ciencia y Tecnología.

Y por supuesto gracias a Guru Ram Das.

Introducción

Esta tesis trata con problemas de control de Markov con recompensa total esperada ([6], [9], [15] y [16]). En particular aquí se encontrará la solución a dos ejemplos de problemas de control de Markov en espacios finitos y con un estado absorbente, ambos ejemplos tienen a su función objetivo relacionada con la recompensa total esperada. El primero de ellos será llamado ejemplo neutral al riesgo mientras que al segundo le llamaremos *sensible al riesgo*. Una forma alternativa de distinguirlos será a través de un parámetro λ conocido como *coeficiente de sensibilidad al riesgo*. Cuando $\lambda = 0$ se tiene el caso neutral al riesgo [16], y una de las aportaciones de este trabajo es proporcionar detalladamente la solución a este ejemplo. Por otro lado cuando $\lambda \neq 0$ estamos ante un problema de control de Markov sensible al riesgo con recompensa total esperada [5]. Hasta donde sabemos son escasos y poco detallados los ejemplos a problemas de control de Markov con este tipo de función objetivo, por ello resulta interesante la aportación de la tesis en este sentido. En este trabajo proponemos la extensión del ejemplo neutral al sensible a través de la función objetivo adecuada, también proponemos y verificamos la solución de este nuevo ejemplo utilizando los resultados de Cavazos-Cadena y Montes-de-Oca en [5]. Esta extensión constituye otra aportación de la tesis.

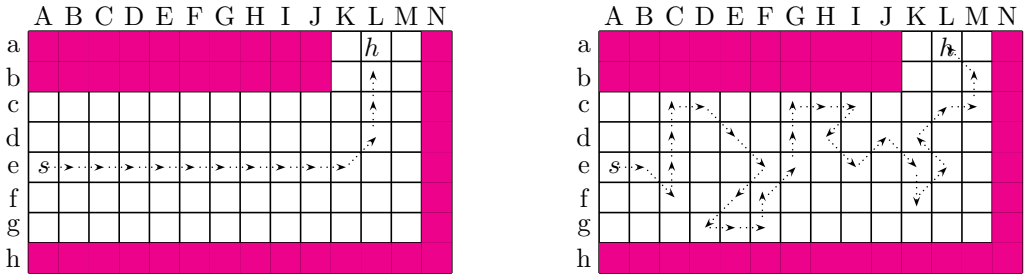
En esta introducción además de proveer un panorama general del trabajo queremos mostrar un ejemplo que nos sirva para ilustrar el concepto de *modelo de control de Markov* y también discutir brevemente el de *sensibilidad al riesgo*. Este ejemplo es interesante porque aparte de que permite describir los elementos de un modelo de control es también un ejemplo de una de las primeras aplicaciones de PCMs con sensibilidad al riesgo en el campo de la inteligencia artificial (véase [12]).

Considérese una partícula situada en el interior de una cuadrícula, de hecho

en el interior de alguno de los cuadrados, no sobre los vértices ni sobre los ejes; la partícula puede cambiar su posición a un cuadrado adyacente. Supongamos que podemos observar periódicamente estos cambios. Los movimientos de la partícula son producto de la acción o decisión de un controlador. Esta acción puede ser de dos tipos: determinista o aleatoria. Con acción determinista nos referimos al hecho de que el controlador no tiene más que una opción para la nueva posición y es ésta la que *elige*. El caso aleatorio es más interesante; aquí el controlador tiene un distribución probabilística que le ayuda a elegir la nueva posición. Puede ilustrarse esto en el siguiente cuadro.

(i) Acción determinista (ii) Acción aleatoria

Cuadro 1: Dos formas de moverse



(i) (ii)

Cuadro 2: Dos tipos de trayectorias

La figura del Cuadro 1 es sólo ilustrativa pues en principio el controlador podría elegir entre todas las posiciones adyacentes o incluso elegir la posición inicial y no sólo entre las tres dibujadas. Nótese además que, en este ejemplo, para

cada posición de la partícula, existe sólo un número finito de acciones posibles (cuadrados adyacentes) que el controlador puede elegir.

Ahora supongamos que la partícula debe llegar desde una posición inicial s en la cuadrícula hasta una final h, (véase el cuadro 2(i)). Entonces el controlador necesitará determinar una serie de acciones que deriven en que la partícula alcance la posición h. En el caso de que el controlador tome acciones deterministas el punto h se puede alcanzar de una sola manera. Por otro lado, si las acciones que puede elegir el controlador son aleatorias, resultan distintas formas de llegar a la posición h. A las formas de llegar las llamamos *trayectorias*. En el Cuadro 2 se muestran dos tipos de trayectorias generadas bajo acciones aleatorias.

Consideremos ahora que en cada paso, es decir ante la acción elegida para el estado de la partícula, el sistema no sólo se mueve hacia otro estado sino que además responde de alguna forma; podríamos pensar por ejemplo que hay una penalización si la acción elegida provoca que la partícula llegue a la zona oscura de la retícula. Así esta función de respuesta del sistema de la partícula podría llamarse *función de costo por etapa*.

Desde el enfoque de los problemas de control de Markov (aunque aún de manera informal) este ejemplo puede observarse así:

Denotemos con x_t la posición en la que se encuentra la partícula en el tiempo t ($t = 0, 1, ...$, es decir tiempo discreto), al conjunto de las posiciones le llamaremos *espacio de estados*. Tomando en cuenta x_t un controlador elige una acción a_t, ir al norte, al sur, al sureste, al este etc. Como consecuencias de la acción elegida están la nueva posición de la partícula x_{t+1} y la función de respuesta del sistema $\rho(x_t, a_t)$, como la penalización de la que se habló antes; dado que esto sucede para cada etapa del proceso resulta natural llamar a esta ρ *función de respuesta por etapa*. Como la partícula está ya en su nueva posición x_{t+1}, el controlador tiene las condiciones para elegir una nueva acción a_{t+1} y continuar con la dinámica del sistema. En el contexto de los procesos de control de Markov a la sucesión de acciones tomadas se le conoce como *política* ó *estrategia*. No hay razón para pensar que hay una sola política, así que para cada política fija y cada estado x_0 también fijo, queda determinado un proceso estocástico que será llamado *proceso de control de Markov*. Ahora consideramos una función real-valuada V que mida la calidad de cada política π dado un estado inicial x_0, utilizando ρ la función de

respuesta por etapa. $V(\pi, x_0)$ será llamada *función objetivo*. Como ejemplos de este tipo de funciones tenemos el costo total, el costo promedio y la recompensa total ([6], [9], [15] y [16]). Conviene aquí aclarar la diferencia entre ρ y V. Como ya dijimos la primera es una función de respuesta por etapa mientras que en la segunda se debe considerar todo el proceso.

El problema básico de un *problema de control de Markov* (PCM), consistirá en encontrar la política que optimiza la función objetivo y también es conocido como *problema de control óptimo* (PCO). De tal manera que en esta tesis se usarán las dos expresiones para referirnos al mismo concepto; por otro lado reservaremos el uso de las iniciales PCMs para abreviar *problemas de control de Markov*.

La *función objetivo* es un ingrediente importante del problema de control, pues ésta determina el tipo de problema; así, se habla de problemas con *recompensa total*, con *recompensa descontada*, con *costo total*, etc. En particular están los que por función objetivo tienen, de hecho, una familia de funciones objetivo parametrizada por un número real llamado *coeficiente de sensibilidad al riesgo*, usualmente denotado por λ.

Desde la perspectiva de la teoría de los problemas de control de Markov, utilizar este tipo de funciones objetivo simplemente genera una familia particular de problemas, y como tales se han estudiado. En la literatura pueden encontrarse resultados acerca de problemas de control *con costo promedio sensible al riesgo* [4], *con recompensa total sensible al riesgo* [5]. Para una revisión más exhaustiva puede consultarse el artículo de Marcus et. al. [13], este es un artículo que podríamos llamar panorámico, [3] y la disertación de Liu [12] y las referencias dadas por ellos.

Históricamente los PCMs que toman en cuenta la sensibilidad al riesgo del controlador surgieron a partir de observarlos desde el enfoque de la llamada *Teoría de la Utilidad*. En ella se afirma que, bajo ciertas condiciones, es posible modelar las preferencias de un consumidor a través de una función numérica que llaman *función de utilidad*. En [17] von Neumann y Morgenstern llevaron esta idea al ámbito estocástico y encontraron que pueden modelarse preferencias entre distribuciones de probabilidad (que ellos llamaban loterías), a través de una *función de utilidad esperada*. En [10] Howard y Matheson extendieron esta idea aún más y la llevaron a la preferencia entre políticas relacionando de esta manera a la teoría

de la utilidad con la teoría de los problemas de control de Markov. Esto derivó en una extensión importante a la teoría de los problemas estudiados anteriormente pues la función objetivo de los que podríamos llamar "clásicos" modela las preferencias del consumidor de manera lineal y en ese sentido la función objetivo es una *función de utilidad neutral al riesgo*. Pero gracias a la propuesta de Howard y Matheson es posible incluir en las decisiones del controlador *su* actitud ante el riesgo a través de una función no lineal. Con actitud al riesgo se refieren a los comportamientos que tiene el controlador al momento de elegir sus acciones y se distinguen tres tipos de controladores: neutrales al riesgo, aversos al riesgo y propensos al riesgo.

En este sentido son muy interesantes los resultados presentados en la tesis [12] para el ejemplo de las trayectorias de la partícula. En este trabajo hicieron la simulación de 2000 trayectorias proponiendo funciones objetivo de tres tipos distintos: neutral al riesgo, aversa al riesgo y propensa al riesgo. Los resultados obtenidos son cualitativamente distintos, por ejemplo es notoria la dispersión de las trayectorias cuando se usa una función de utilidad propensa al riesgo. Esto no es una sorpresa, pues es intuitivamente claro que usar una función de utilidad propensa al riesgo derive en que las acciones elegidas por un controlador sean más "aventuradas" y esto a su vez, derive en que las trayectorias sean más dispersas. Hasta donde sabemos este es uno los primeros trabajos de inteligencia artificial en los que están aplicados los conceptos que desarrollaron Howard y Matheson en 1972.

La organización de este texto es la siguiente. En el Capítulo 1 se presentan con detalle los preliminares necesarios para plantear un problema de control óptimo; desde lo que es un modelo de control de Markov hasta llegar a los dos tipos de problemas de control que se discutirán en el resto del trabajo. El Capítulo 2 está dedicado a presentar cuidadosamente las dos funciones objetivo que generan los problemas de control, esto se hace necesario sobre todo para la que será llamada λ-función objetivo, pues en su definición están incluidos conceptos de tipo económico que requieren cierta cautela. En el Capítulo 3 se hallan los teoremas que respectivamente serán usados para resolver los problemas de control. Los Capítulos 4 y 5 contienen los ejemplos que fueron propuestos y resueltos con la teoría previamente descrita. En particular a lo largo del Capítulo 4 está *minuciosamente* desarrollado un ejemplo de un *PCM con recompensa total neutral al riesgo*, inclui-

das las herramientas técnicas necesarias para su solución. El Capítulo 5 contiene la propuesta y solución de un ejemplo de un PCM sensible al riesgo. Este capítulo está fuertemente apoyado en el anterior pues como ya se ha dicho el ejemplo λ-sensible al riesgo puede mirarse como una extensión del neutral al riesgo. Además en el apéndice A se pueden consultar los resultados y propiedades de teoría de la medida necesarios tanto para la construcción del proceso estocástico como para la solución a los ejemplos planteados. Mientras que en el Apéndice B se halla una discusión acerca de los elementos de la teoría de la utilidad que generaron la idea del proceso de control de Markov sensible al riesgo.

Capítulo 1

Preliminares

En este capítulo quedarán establecidos los preliminares que permiten plantear un problema de control óptimo. En la primera sección se establecerán lo que es un modelo de control de Markov (MCM) y una interpretación para este tipo de modelos. En la segunda se define lo que entenderemos por *política* para llegar a la definición de proceso de control de Markov. En la tercera sección se plantea de manera general lo que es un problema de control de Markov (PCM), o problema de control óptimo (PCO).

1.1. Modelos de Control de Markov

En esta sección se encontrarán la definición e interpretación de lo que es MCM, particularmente de modelos de control de Markov (MCMs) a tiempo discreto y con espacios, tanto de estados como de controles, finitos (véase [9]), pues los ejemplos desarrollados en la tesis no requieren más.

Definición 1.1. Un **Modelo de control de Markov** es una quíntupla

$$M := (X, A, \{A(x)|x \in X\}, Q, \rho) \tag{1.1}$$

que consiste de

1. X, un conjunto finito, al que se llamará **espacio de estados del sistema**. Los elementos $x \in X$ se llamarán **estados**.

2. A, un conjunto finito, llamado **espacio de controles** o **espacio de acciones**.

3. $\{A(x)|x \in X\}$, una familia de subconjuntos no vacíos $A(x)$ de A, donde $A(x)$ es el conjunto de controles admisibles para el estado $x \in X$. A

$$\mathbb{K} := \{(x, a) \mid x \in X, \quad a \in A(x)\}, \tag{1.2}$$

se le llamará conjunto de pares **estado-acción admisible**.

4. Q, una medida de probabilidad sobre X dado $\mathbb{K}$, se le llama también ***ley de transición***. Este nombre tiene sentido pues Q nos "da" la probabilidad condicional de que el sistema se mueva a un nuevo estado dado que se encuentra en el estado actual y se elige una acción admisible; es decir Q es de la siguiente forma (véase Apéndice A):

$$Q(B \mid x, a) := \text{Prob}(X_{t+1} \in B \mid X_t = x, A_t = a), \quad B \subset X, \tag{1.3}$$

donde $\quad t = 0, 1, 2, ...$

5. $\rho : \mathbb{K} \to \mathbb{R}$, una función que representa una respuesta del sistema, en el sentido de que $\rho(x, a)$ es resultado de haber aplicado el control a cuando el sistema estaba en el estado x.

El último elemento de la quíntupla es muy importante para la definición del problema de control óptimo. Algunos ejemplos de ρ pueden ser: el **costo por etapa** o **la recompensa por etapa** es decir el costo o recompensa que se obtienen en cada etapa del proceso como resultado de haber elegido una acción (control) dado que el estado actual es x . En particular en esta tesis se trabajará con una función de recompensa por etapa.

Interpretación.

Consideremos un sistema estocástico controlado y supongamos que el sistema puede ser observado en cada etapa. Es posible hacer una conexión entre este sistema y un modelo de control como el descrito al inicio de la sección. Es decir el modelo de control definido en (1.1) representa al sistema estocástico controlado

con espacio de estados X y de controles A, este sistema es observado en cada tiempo $t = 0, 1, ...$ Con X_t y A_t se denotarán el estado del sistema y el control (o acción) aplicado en el tiempo t, respectivamente. Así el desarrollo del sistema puede ser descrito como sigue: si el sistema está en el estado $X_t = x \in X$ en el tiempo t y el control $A_t = a \in A(x)$ es aplicado, entonces sucede lo siguiente.
(i) se obtiene una respuesta del sistema $\rho(x, a)$, como consecuencia de la acción elegida para ese estado y
(ii) el sistema se mueve al siguiente estado X_{t+1}, el cual es una variable aleatoria X-valuada con distribución $Q(\cdot \mid x, a)$ i.e.,

$$Q(B \mid x, a) := \text{Prob}(X_{t+1} \in B \mid X_t = x, A_t = a), \quad B \subset X$$

esto sucede para cada estado y cada acción admisible elegida para él así, lo que tenemos es una matriz de transición

$$Q := [q_{xy}(a)]$$

con las siguientes propiedades: para cada $y \in X$ fijo ocurre $\sum_{x \in X} q_{yx} = 1$, y además cada $q_{xy} \in [0, 1]$.

Una vez que el sistema se encuentra en el nuevo estado, se vuelven a tener las condiciones de elegir un nuevo control y el proceso se repite. Un modelo de control de Markov se caracteriza por **(i)** y **(ii)** de manera que, en cualquier tiempo, la respuesta obtenida ρ y la ley de transición dependen sólo del estado actual del sistema y de la acción elegida para ese estado y ese momento.

Con el modelo de control de Markov establecido, surge natural la siguiente pregunta ¿cómo elegir a para cada x?, más aún ¿la respuesta del sistema a cada acción elegida mejora o empeora con cada acción? ¿puede controlarse la forma de elegir? Aquí se hace importante el concepto de estrategia o política.

1.2. Políticas

En esta sección se definirá lo que es una política o estrategia en el contexto de modelos de control y con esto se podrá llegar a la definición de proceso de control de Markov.

Considerando un modelo de control como en la definición (1.1), para cada $t = 0, 1, ...$ se define el espacio H_t de *historias admisibles* hasta el tiempo t como sigue $H_0 := X$; y

$$H_t := \mathbb{K}^t \times X = \mathbb{K} \times H_{t-1} \quad \text{para} \quad t = 1, 2, ... \tag{1.4}$$

donde $\mathbb{K}$ está dado por (1.2). Un elemento genérico h_t de H_t, al que se llamará una t-historia *admisible*, o simplemente una *t-historia*, es un vector de la forma

$$h_t = (\xi_0, \alpha_0, ..., \xi_{t-1}, \alpha_{t-1}, \xi_t) \tag{1.5}$$

con $(\xi_i, \alpha_i) \in \mathbb{K}$ para $i = 0, ..., t-1$, y $\xi_t \in X$.

Definición 1.2. Una **política de control aleatorizada** - o simplemente **política**- es una sucesión $\pi = \{\pi_t, t = 0, 1, ...\}$ de medidas de probabilidad π_t sobre el conjunto de controles A dado H_t que satisfacen lo siguiente

$$\pi_t(A(x_t) \mid h_t) = 1 \quad \text{para todo} \quad h_t \in H_t, \quad t = 0, 1, ... \tag{1.6}$$

Definición 1.3. Sea $\mathbb{F}$ el conjunto de todas las funciones $f : X \to A$ tales que $f(x) \in A(x)$ para todo $x \in X$. Se llamarán **políticas deterministas estacionarias** a las políticas para las cuales existe una función $f \in \mathbb{F}$ tal que $\pi_t(\cdot \mid h_t)$ está concentrada en $f(x_t) \in A(x_t)$ para toda $h_t \in H_t$ y $t = 0, 1, ...$

El conjunto de todas las políticas es denotado por $\mathcal{P}$ y al conjunto de las deterministas estacionarias con $\mathbb{F}$; está claro que $\mathbb{F} \subset \mathcal{P}$. Nótese que estos conjuntos de políticas están asociadas al modelo, sin embargo no es costumbre indexarlas pues cargaría aún más la ya de por sí abigarrada notación. En esta tesis usaremos políticas estacionarias. Con el concepto de política definido estamos en condiciones de definir lo que llamaremos *proceso de control de Markov*.

Dada una política π y un estado inicial x_0 queda determinado un proceso estocástico cuya dinámica podemos describir como sigue.

Supongamos que en el tiempo t el proceso tiene una historia h_t y está en el estado x_t, entonces se elige una acción (posiblemente de manera aleatoria), de acuerdo a π. Sea a_t la acción dictada por π, entonces el proceso estará en el estado y

con probabilidad $Q(\{y\} \mid x_t, a_t)$. Esta dinámica junto con la política π y un estado inicial x definen todas las distribuciones finito-dimensionales $\xi_0, \alpha_0, ..., \xi_{t-1}, \alpha_{t-1}, \xi_t, t \in \mathbb{N}$, y el teorema de Ionescu Tulcea garantiza que, dados x y π se definen las sucesiones $\{X_t\}$, $\{A_t\}$ de estados y controles respectivamente (véase apéndice A). Denotamos con P_x^π y E_x^π respectivamente a las probabilidades y las esperanzas relacionadas con esta construcción.

1.3. Problema de Control Óptimo

En esta sección se planteará, en general, lo que es un problema de control óptimo o problema de control de Markov. Con esto quedan terminados los preliminares.

La idea es considerar una función del siguiente estilo

$$V : \mathcal{P} \times X \to \mathbb{R}, \tag{1.7}$$

a través de la cual mediremos el resultado obtenido a lo largo del proceso, bajo las acciones dictadas por alguna política π, y dado que el estado inicial fue un x fijo. Una función con tales características es lo que llamaremos *función objetivo*. Como regla general esta función estará relacionada con la función de respuesta por etapa.

Definición 1.4. Dados un MCM $\{(X, A, \{A(x)|x \in X\}, Q, \rho)\}$, el conjunto de políticas $\mathcal{P}$ y una función objetivo V. El *Problema de Control Óptimo* consiste en determinar $\pi^* \in \mathcal{P}$ (si es que ésta existe), tal que

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

Más aún

Definición 1.5. A la función

$$\mathcal{V}(x) = \sup_{\pi \in \mathcal{P}} V(\pi, x), \quad x \in X, \tag{1.8}$$

la llamaremos *función de valor óptimo del PCO*.

De tal manera que resulta natural la siguiente definición.

Definición 1.6. Si existe una política $\pi^* \in \mathcal{P}$ tal que

$$\mathcal{V}(x) = V(\pi^*, x) \quad \text{para todo} \quad x \in X \tag{1.9}$$

entonces a esta π^* le llamaremos *política óptima*.

El *proceso de control de Markov* junto con la función objetivo a optimizar es lo que se conoce como *Problema de Control de Markov* (PCM). En algunos textos se utilizan de manera indistinta las dos expresiones. Pero en este trabajo reservaremos la segunda para hablar del modelo de control y su estructura de políticas $\mathcal{P}$ más la función objetivo a optimizar. Y en todo caso usaremos a discreción problema de control de Markov como sinónimo de problema de control óptimo.

En el capítulo siguiente se presentan dos problemas de control óptimo que tienen asociado el mismo modelo de control pero distintas funciones objetivo.

Capítulo 2

Funciones Objetivo Asociadas a la Recompensa Total

A lo largo de este capítulo tendremos fijo un proceso de control de Markov, es decir un MCM y su estructura de políticas. La idea es que a partir de ellos queden descritos completamente los dos problemas de control de Markov o problemas de control óptimo que son ejes de este trabajo. El primero de ellos fue motivado por el ejemplo de la Sección 2 del capítulo IV en [16]. Este problema tiene como función objetivo a la recompensa total esperada, así la primera sección de este capítulo está dedicada al planteamiento de este problema. Al final de la segunda sección quedará planteado otro PCM que de hecho será una familia de PCMs parametrizados por un número real $\lambda \neq 0$. Pero antes de llegar a este planteamiento deberán presentarse algunos preliminares que ayudan a entender la importancia de este tipo de función objetivo.

2.1. Caso Neutral al Riesgo

Consideremos un modelo de control

$$(X, A, \{A(x)|x \in X\}, Q, R), \tag{2.1}$$

en el que tanto A como X son finitos y donde estamos usando R para denotar la función de respuesta ρ de la que hablamos en el capítulo anterior, pues en este

modelo en particular será una función de recompensa R, por etapa que además supondremos no negativa; sea $\mathcal{P}$ el conjunto de políticas de este modelo.

Para describir el primer PCM se usará como función objetivo la esperanza de la recompensa total, es decir dados el modelo de control y el conjunto de políticas asociadas a él, consideremos la siguiente función de objetivo para una política $\pi \in \mathcal{P}$ y un estado inicial $x_0 = x \in X$

$$E_x^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \quad x \in X.$$

Aunque esta medida del funcionamiento de una política no siempre es apropiada tendremos, más adelante, condiciones en la descripción de un MCM que nos permitirán usarlo, véase [15] p. 123; denotaremos con $V(\pi, x)$ a la ***esperanza de la recompensa total ganada*** bajo π cuando el estado inicial es $x_0 = x$, es decir

$$V(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \quad x \in X.$$

En este caso se dice que tendremos un problema con horizonte infinito. Como $R \geq 0$, $V(\pi, x)$ está bien definida, aunque podría ser infinita.

Igual que en (1.8) definimos a la función de valor óptimo como sigue

$$\mathcal{V}(x) = \sup_{\pi \in \mathcal{P}} E_x^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \quad \text{para todo} \quad x \in X,$$

y del mismo modo diremos que π^* es *óptima* si

$$\mathcal{V}(x) = V(\pi^*, x), \quad \text{para todo} \quad x \in X.$$

2.2. Caso Sensible al Riesgo

El concepto de *función de utilidad* U nace en economía a partir de la necesidad de estudiar las preferencias de un consumidor. La idea básica es poder representar, a través de una función numérica U, las preferencias de un consumidor; estas

preferencias en principio no están determinadas numéricamente así que la conveniencia de usar una U es que ella sí lo es. Más aún, gracias a los trabajos de J. von Neumann y O. Morgenstern [17] es posible representar preferencias bajo condiciones de incertidumbre; en estos casos se habla de una *función de utilidad esperada*. J. von Neumann y O. Morgenstern demuestran que utilidades esperadas mayores representan situaciones más deseables. En el caso de los PCMs lo que se hace es modelar preferencias sobre las políticas a través de la función objetivo. Pero si además introducimos una función de utilidad asociada a la recompensa total ganada bajo una política dada [5], resulta natural decir que preferimos una política sobre otra observando cuál de ellas tiene mejor utilidad esperada. Más aún, a través de esta función U también es posible medir la sensibilidad al riesgo que tiene un controlador. Y en este sentido también hablamos de *preferencia* al riesgo [11],[10],[5],[17]. En este caso la función objetivo también estará relacionada con la recompensa total pero además con otros dos conceptos que a continuación definiremos.

2.2.1. Certeza Equivalente aplicada a PCMs [1]

Definición 2.1. Dado $\lambda \in \mathbb{R}$ fijo. Para todo $x \in \mathbb{R}$ sea

$$U_\lambda(x) := \begin{cases} \text{sign}(\lambda)e^{\lambda x}, & \lambda \neq 0; \\ x, & \lambda = 0. \end{cases} \tag{2.2}$$

Donde $\text{sign}(\lambda) = 1$ si $\lambda > 0$ y $\text{sign}(\lambda) = -1$ si $\lambda < 0$. Nótese que $U_\lambda(\cdot)$ es una función estrictamente creciente para cualquier valor de λ.

El segundo concepto del que hablaremos y que será de gran importancia para ayudarnos a definir lo que llamaremos la λ-función objetivo es el siguiente.

Definición 2.2. Sea Z una variable aleatoria y supongamos que el valor esperado (con respecto a cierta probabilidad, digamos Θ), de $U_\lambda(Z)$ está bien definido. **La certeza equivalente** $\mathcal{Q}$, de Z con respecto a U_λ está dada por

$$\mathcal{Q}_\lambda(Z) := \begin{cases} \frac{1}{\lambda}\ln(E[e^{\lambda Z}]), & \lambda \neq 0; \\ E[Z], & \lambda = 0. \end{cases} \tag{2.3}$$

[1] En esta sección usaremos el concepto de Certeza Equivalente como objeto matemático fuera de interpretaciones pues no se pretende estudiarlo desde el punto de vista de la Economía

En esta definición E denota la esperanza (con respecto a Θ) de manera genérica. A partir de (2.2) y (2.3) es posible verificar directamente que se cumple lo siguiente

$$U_\lambda(\mathcal{Q}_\lambda(Z)) = E[U_\lambda(Z)], \tag{2.4}$$

es decir, la función de utilidad y la certeza equivalente están fuertemente relacionadas. Lo que esto implica es que el controlador tiene la opción de intercambiar la oportunidad de obtener la recompensa aleatoria Z, por la correspondiente *certeza equivalente* $\mathcal{Q}_\lambda(Z)$, (véase [5]).

Esta relación entre función de utilidad y certeza equivalente tiene gran importancia, como se verá más adelante, en el contexto del modelo de control. Dado un estado inicial $x_0 \in X$ fijo, consideremos los procesos $\{X_t\}$ y $\{A_t\}$ generados a partir de la medida inducida por alguna política, digamos por π, y consideremos también la variable aleatoria $Y = \sum_{t=0}^{\infty} R(X_t, A_t)$; es decir Y es la recompensa total ganada (a través de todo el proceso). Para no cargar la notación no será indexada esta Y con la política π y el estado inicial x_0 que la generaron, sin embargo esto debe tenerse siempre en cuenta. Por otro lado, puesto que Y es un valor en $\mathbb{R}$, tiene sentido aplicarle U_λ. En este caso diremos que $U_\lambda(Y)$ es la *utilidad de la recompensa total ganada (bajo π)*. No obstante nuestro interés estará enfocado en algo un poco más elaborado.

Denotamos con

$$E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right]$$

a la ***utilidad esperada de la recompensa total ganada*** dado que el estado inicial es $X_0 = x$ y se está usando la política π.

Es claro que si η es otra política podemos comparar la utilidad esperada de la recompensa total obtenida bajo ella con respecto a la que se obtuvo bajo π, pues son números reales. Si en particular ocurre que

$$E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right] > E_x^\eta \left[U_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right] \tag{2.5}$$

entonces tiene sentido decir que un controlador *prefiere* a π sobre η.

La utilidad esperada de la recompensa total ganada bajo alguna política π en sí misma podría ser propuesta como función objetivo, sin embargo cambiaremos el enfoque y, a través de (2.4), obtenemos

$$E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right] = U_\lambda \left(\mathcal{Q}_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right).$$

Para aligerar la notación, regresemos a usar $Y = \sum_{t=0}^{\infty} R(X_t, A_t)$ y escribimos

$$E_x^\pi \left[U_\lambda \left(Y \right) \right] = U_\lambda \left(\mathcal{Q}_\lambda \left(Y \right) \right). \tag{2.6}$$

Esto no es más que lo escrito en (2.4) sólo que ahora guarda sentido con los elementos del modelo de control. Aunque el lado derecho de esta igualdad no parece estar relacionado con la política π recordemos que Y está relacionada con ella; además este lado de la igualdad tiene sus ventajas como veremos ahora. Supongamos que U_λ toma valores y no constantes y tomemos λ negativo así que tenemos la siguiente función:

$$U_\lambda(y) = -e^{\lambda y}; \quad y \in \mathbb{R},$$

que es claramente cóncava. Veamos qué consecuencias tiene esto sobre (2.6), para ello tenemos una herramienta básica: la *desigualdad de Jensen*. Lo que en ella se afirma es que si consideramos una función real valuada f y una variable aleatoria W entonces $E(f(W))$ es siempre más pequeña que $f[E(W)]$ si y sólo si f es cóncava. Tomemos $f = U_\lambda$ y $W = Y$. Usando esta herramienta podemos afirmar que

$$U_\lambda \left(\mathrm{E} \left[Y \right] \right) > \mathrm{E} \left[U_\lambda \left(Y \right) \right].$$

Llevando esto al contexto de los PCMs, cuando un controlador tiene este comportamiento, es decir cuando prefiere la utilidad del valor esperado con certeza sobre la utilidad esperada de una situación incierta, se le llama *averso al riesgo*. Si además recordamos que la esperanza es con respecto a la medida de probabilidad inducida por la política π y el proceso tiene como estado inicial $x \in X$, esto queda aún más preciso de la siguiente manera

$$U_\lambda \left(\mathrm{E}_x^\pi \left[Y \right] \right) > \mathrm{E}_x^\pi \left[U_\lambda \left(Y \right) \right]. \tag{2.7}$$

Nótese que esta desigualdad no está comparando lo obtenido por una política con respecto a otra. Aquí estamos estableciendo otro orden de preferencias a saber: sobre cómo prefiere medirse la calidad de la política. Esta desigualdad tiene gran importancia pues al combinar (2.7) con (2.6) obtenemos

$$U_\lambda\left(\mathrm{E}_x^\pi\left[Y\right]\right) > U_\lambda(\mathcal{Q}_\lambda(Y)) \tag{2.8}$$

mejor aún, como U_λ tiene inversa conseguimos

$$\mathrm{E}_x^\pi\left[Y\right] > \mathcal{Q}_\lambda(Y). \tag{2.9}$$

Del lado izquierdo de esta desigualdad tenemos la función objetivo del problema de control planteado en la sección anterior así que tiene sentido, al menos suponer, que del lado derecho lo que hay es otra función objetivo. De hecho gracias a la definición (2.3) la conocemos explícitamente.

$$\mathrm{E}_x^\pi\left[\sum_{t=0}^{\infty} R(X_t, A_t)\right] > \frac{1}{\lambda}\ln\left(E_x^\pi\left[e^{\lambda\left(\sum_{t=0}^{\infty} R(X_t,A_t)\right)}\right]\right). \tag{2.10}$$

Además de permitirnos describir una función objetivo, de hecho una familia de funciones objetivo, esta desigualdad nos dice que un controlador sensible al riesgo es *averso al riesgo*, cuando λ es menor que cero, pues prefiere como función objetivo algo menor que la función objetivo no modulada con U_λ.

No está por demás decir que para el caso $\lambda > 0$ se obtiene de manera análoga lo siguiente

$$\mathrm{E}_x^\pi\left[\sum_{t=0}^{\infty} R(X_t, A_t)\right] < \frac{1}{\lambda}\ln\left(E_x^\pi\left[e^{\lambda\left(\sum_{t=0}^{\infty} R(X_t,A_t)\right)}\right]\right), \tag{2.11}$$

y aquí se considerará que un controlador sensible al riesgo es *propenso* al riesgo pues prefiere como función objetivo algo mayor que la función objetivo no modulada con U_λ.

En el caso $\lambda = 0$ al controlador se le considera *neutral al riesgo*, este adjetivo resulta natural simplemente porque la función de utilidad definida en (2.2) es lineal para este caso (de hecho es la identidad). Por lo tanto el comportamiento del controlador en realidad no está modulada por una función de utilidad.

2.2.2. PCMs con Recompensa Total Sensible al Riesgo

Recordemos que tenemos un MCM fijo y $\mathcal{P}$ el conjunto de políticas del modelo; la función U_λ definida en (2.2) ahora es ingrediente fijo importante. Sea $\pi \in \mathcal{P}$ una política y, para $\lambda \neq 0$, consideremos la certeza equivalente de la utilidad esperada de la recompensa total ganada bajo π, dado que el estado inicial es $x_0 = x$, es decir

$$\mathcal{Q}_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) = \frac{1}{\lambda} \ln \left(E_x^\pi \left[e^{\lambda \sum_{t=0}^{\infty} R(X_t, A_t)} \right] \right). \tag{2.12}$$

Con ésta como función objetivo[2] se puede ya definir lo que llamaremos λ-*problema de control óptimo o problema de control de Markov* λ-*sensible al riesgo.* De hecho esta λ-función describe una familia de PCOs. En adelante la denotaremos como sigue:

$$V_\lambda(\pi, x) = \frac{1}{\lambda} \ln \left(E_x^\pi \left[e^{\lambda \sum_{t=0}^{\infty} R(X_t, A_t)} \right] \right). \tag{2.13}$$

Nótese que para todo $\lambda \neq 0$ se tiene que $e^{\lambda \sum_{t=0}^{\infty} R(X_t, A_t)} \geq 1$ simplemente porque estamos considerando a la recompensa como no negativa. Por lo tanto $V_\lambda(\pi, x) \geq 0$.
Recordemos la definición de la λ-*función de valor óptimo*

$$\mathcal{V}_\lambda(x) = \sup_\pi \{V_\lambda(\pi, x)\}, \quad x \in X, \tag{2.14}$$

y que una política π^* es λ-óptima si

$$V_\lambda(\pi^*, x) = \mathcal{V}_\lambda(x), \quad \text{para toda} \quad x \in X$$

Finalmente han quedado establecidos los dos tipos de problemas de control de Markov que se estudiarán en esta tesis. En el siguiente cuadro quedan esquematizados estos dos problemas.

[2] En [5] a esta función la llaman *recompensa total esperada* λ-*sensible al riesgo.*

Función Objetivo	Función de valor óptimo	Política Óptima	Problema de C. Óptimo
$V(\pi, x) = E_x^{\pi}\left[\sum_{t=0}^{\infty} R(X_t, A_t)\right]$	$\mathcal{V}^*(x) = \sup_{\pi} V(\pi, x)$	π^*	Neutral al Riesgo
$V_\lambda(\pi, x) = \frac{1}{\lambda}\ln\left(E_x^{\pi}\left[e^{\lambda\sum_{t=0}^{\infty} R(X_t, A_t)}\right]\right)$	$\mathcal{V}_\lambda^*(x) = \sup_{\pi} V_\lambda(\pi, x)$	π_λ^*	Sensible al Riesgo

Capítulo 3

Desigualdades de Optimalidad

En este capítulo se presentan los resultados que permitirán dar solución a los problemas de control óptimo planteados en el anterior. Demostraremos cuándo es posible afirmar que una política es óptima tanto para el modelo con función objetivo *recompensa total* como para el que tiene asociada la *recompensa total λ-sensible al riesgo*.

3.1. Caso Neutral al Riesgo

Consideremos el siguiente modelo

$$(X, A, \{A(x)|x \in X\}, Q, R), \tag{3.1}$$

en el que tanto A como X son finitos y donde R es una función de recompensa por etapa que además supondremos no negativa; sea $\mathcal{P}$ el conjunto de políticas de este modelo y como función objetivo la recompensa total

$$V(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right]. \tag{3.2}$$

El problema de control de Markov asociado con este modelo será el siguiente.

Hallar una política $\pi^* \in \mathcal{P}$ (si es que existe) tal que

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

En esta sección daremos las condiciones suficientes para afirmar que una política f es óptima es decir,

$$V(f,x) = \mathcal{V}(x), \qquad x \in X;$$

y esto lo haremos a través de la que será llamada ***desigualdad de optimalidad***.

Teorema 3.1. Sea f una política estacionaria tal que $V(f,x) < \infty$ para todo $x \in X$. Si

$$V(f,x) \geq R(x,a) + \sum_y V(f,y) q_{xy}(a), \tag{3.3}$$

para todo $x \in X$ y para todo $a \in A(x)$ entonces se cumple que

$$\mathcal{V}(\cdot) = V(f,\cdot),$$

es decir, f es óptima.

Demostración. Sea $X_0 = x \in X$. Dada una política π cualquiera, por la propiedad de Markov, (véase (A.6)) tenemos para $t \geq 0$ lo siguiente

$$\begin{aligned} E_x^\pi[V(f,X_{t+1}) \mid h_t, a_t] &= \sum_y V(f,y) q_{x_t y}(a_t) \\ &= \{R(X_t,A_t) + \sum_y V(f,y) Q(dy \mid x_t, a_t)\} - R(X_t,A_t) \\ &\leq V(f,X_t) - R(X_t,A_t) \quad \text{por hipótesis,} \end{aligned}$$

es decir,

$$R(X_t,A_t) \leq V(f,X_t) - E_x^\pi[V(f,X_{t+1}) \mid h_t, a_t], \quad t = 0,1,2,...$$

en lo que sigue $V(f,\cdot) = W(\cdot)$; sumemos desde $t = 0$ hasta $n-1$

$$\sum_{t=0}^{n-1} R(X_t,A_t) \leq \sum_{t=0}^{n-1} W(X_t) - \sum_{t=0}^{n-1} E_x^\pi[W(X_{t+1}) \mid h_t, a_t],$$

luego reordenamos los índices de la primera suma del lado derecho es decir,

$$\sum_{t=0}^{n-1} R(X_t,A_t) \leq \sum_{t=0}^{n-1} W(X_{t+1}) + W(X_0) - W(X_n) - \sum_{t=0}^{n-1} E_x^\pi[W(X_{t+1}) \mid h_t, a_t],$$

al reescribir

$$\sum_{t=0}^{n-1} R(X_t, A_t) - W(X_0) + W(X_n) \leq \sum_{t=0}^{n-1} W(X_{t+1}) - \sum_{t=0}^{n-1} E_x^{\pi}[W(X_{t+1}) \mid h_t, a_t].$$

Ahora tomamos la esperanza, usamos del Apéndice A la Proposición A.1(a) y obtenemos

$$E_x^{\pi}[\sum_{t=0}^{n-1} R(X_t, A_t)] - E_x^{\pi}[W(X_0)] + E_x^{\pi}[W(X_n)] \leq \sum_{t=0}^{n-1} \{E_x^{\pi}[W(X_{t+1})] - E_x^{\pi}[W(X_{t+1})]\},$$

i.e,

$$E_x^{\pi}[\sum_{t=0}^{n-1} R(X_t, A_t)] - E_x^{\pi}[W(X_0)] + E_x^{\pi}[W(X_n)] \leq 0.$$

Como la recompensa es no negativa tenemos que $E_x^{\pi}[W(X_n)]$ es positivo, así que podemos escribir, a partir de la desigualdad anterior,

$$E_x^{\pi}[\sum_{t=0}^{n-1} R(X_t, A_t)] \leq E_x^{\pi}[W(X_0)] = W(X_0) = W(x).$$

Al tomar el límite cuando $n \to \infty$ y regresando a la notación $V(f, \cdot) = W(\cdot)$ esto es igual a

$$V(\pi, x) \leq V(f, x),$$

pero esto lo hicimos para cualquier π, en particular tendremos

$$\mathcal{V}(x) \leq V(f, x).$$

Por otro lado sabemos que se cumple trivialmente la otra desigualdad, así obtenemos finalmente que

$$\mathcal{V}(x) = V(f, x), \quad x \in X,$$

es decir, f es óptima. □

3.2. Caso Sensible al Riesgo

Con el mismo modelo de control de Markov como base, pero ahora tomando por función objetivo la recompensa total sensible al riesgo:

$$V_\lambda(\pi, x) = \frac{1}{\lambda}\ln\left(E_x^\pi\left[e^{\lambda\sum_{t=0}^{\infty}R(X_t,A_t)}\right]\right), \quad \lambda \neq 0, \quad \pi \in \mathcal{P}, x \in X.$$

El problema de control óptimo que se tiene ahora consiste en hallar la política π^* tal que

$$V_\lambda(\pi^*, x) = \sup_\pi V_\lambda(\pi, x), \quad x \in X, \tag{3.4}$$

recordemos que a

$$\mathcal{V}_\lambda(x) = \sup_\pi V_\lambda(\pi, x), \quad x \in X,$$

se le llama λ-función de valor óptimo. Como ha sido el caso esta vez necesitamos elaborar un poco más las herramientas necesarias para abordar el problema. En primer lugar recuérdese la definición de función de utilidad U_λ, que se presentó en el Capítulo 2, para $z \in \mathbb{R}$

$$U_\lambda(z) := \begin{cases} \text{sign}(\lambda)e^{\lambda z}, & \lambda \neq 0; \\ z, & \lambda = 0. \end{cases} \tag{3.5}$$

Nótese que $U_\lambda(\cdot)$ es una función estrictamente creciente y que, para $\lambda \neq 0$, tiene la siguiente propiedad

$$U_\lambda(z + c) = e^{\lambda c}U_\lambda(z), \quad z, c \in \mathbb{R}, \tag{3.6}$$

además, de la definición de $V_\lambda(\cdot)$ y (3.5) se obtiene que

$$U_\lambda(V_\lambda(\pi, x)) = E_x^\pi\left[U_\lambda\left(\sum_{t=0}^{\infty}R(X_t, A_t)\right)\right]. \tag{3.7}$$

Teorema 3.2. Sea f una política estacionaria. Si $V_\lambda(f,x) < \infty$ para todo x y

$$U_\lambda(V_\lambda(f,x)) \geq e^{\lambda(R(x,a))}\left[\sum_y U_\lambda(V_\lambda(f,y))q_{xy}(a)\right] \tag{3.8}$$

para todo $x \in X$ y para todo $a \in A(x)$, entonces $\mathcal{V}_\lambda(\cdot) = V_\lambda(f,\cdot)$ y por lo tanto f es óptima.

La demostración de este teorema es consecuencia del siguiente lema que, por razones de espacio, será demostrado usando alternadamente la notación de integrales y de sumas.

Lema 3.1. Con las mismas hipótesis del Teorema 3.2, para toda $n \in \mathbb{N}$ y para cualquier política $\pi \in \mathcal{P}$ se cumple que

$$U_\lambda(V_\lambda(f,x)) \geq E_x^\pi\left[e^{\sum_{t=0}^{n} R(X_t,A_t)} U_\lambda\left(V_\lambda(f,X_{n+1})\right)\right], \quad x \in X. \tag{3.9}$$

Demostración. En efecto sea $\pi \in \mathcal{P}$ cualquier política. A lo largo de esta prueba denotaremos a $V_\lambda(f,\cdot)$ con $W(\cdot)$. La demostración se hará por inducción. Así para $n = 0$ tenemos lo siguiente

$$E_x^\pi[e^{\lambda(R(X_0,A_0))}U_\lambda(W(X_1))] = \\ = \int_X\int_A\int_X e^{\lambda R(x_0,a_0)}U_\lambda(W(x_1))Q(dx_1 \mid x_0,a_0)\pi_0(da_0 \mid x_0)\nu(dx_0)$$

donde ν es la medida concentrada en $X_0 = x$. Observemos que la integral interna

$$\int_X e^{\lambda R(x_0,a_0)}U_\lambda(W(x_1))Q(dx_1 \mid x_0,a_0) = \sum_{x_1} e^{\lambda R(x_0,a_0)}U_\lambda(W(x_1))q_{x_0,x_1}(a_0),$$

por (3.8), es menor o igual que $U_\lambda(W(x))$ así

$$E_x^\pi\left[e^{\lambda R(X_0,A_0)}U_\lambda(W(X_1))\right] \leq \int_X\int_A U_\lambda(W(x))\pi_0(da_0 \mid x_0)\nu(dx_0) = U_\lambda(W(x))$$

Así hemos probado la base de inducción pues obtuvimos que

$$U_\lambda(W(x)) \geq E_x^\pi\left[e^{\lambda R(X_0,A_0)}U_\lambda(W(X_1))\right].$$

Ahora supongamos que para algún $n \geq 1$ es cierto lo siguiente

$$U_\lambda(W(x)) \geq E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda(W(X_{n+1})) \right].$$

Probaremos que esto vale para $n+1$. En efecto, usando la Propiedad A.2 del apéndice A calculemos la siguiente esperanza condicional

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n+1} R(X_t, A_t)} U_\lambda(W(X_{n+2})) \mid h_{n+1}, a_{n+1} \right] =$$

$$= \int_X e^{\lambda \sum_{t=0}^{n+1} R(x_t, a_t)} U_\lambda(W(x_{n+2})) Q(dx_{n+2} \mid x_{n+1}, a_{n+1}) \tag{3.10}$$

y si reescribimos

$$= e^{\lambda \sum_{t=0}^{n} R(x_t, a_t)} \int_X e^{\lambda R(x_{n+1}, a_{n+1})} U_\lambda(W(x_{n+2})) Q(dx_{n+2} \mid x_{n+1}, a_{n+1})$$

una vez más, observemos con cuidado que la integral es:

$$\int e^{\lambda R(x_{n+1}, a_{n+1})} U_\lambda(W(x_{n+2})) Q(dx_{n+2} \mid x_{n+1}, a_{n+1}) =$$

$$= \sum_{x_{n+2}} e^{\lambda R(x_{n+1}, a_{n+1})} U_\lambda(W(x_{n+2})) q_{x_{n+1}, x_{n+2}}(a_{n+1})$$

que, usando (3.8), es menor o igual a $U_\lambda(W(x_{n+1}))$. Así podemos escribir la ecuación (3.10) como sigue

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n+1} R(X_t, A_t)} U_\lambda(W(X_{n+2})) \mid h_{n+1}, a_{n+1} \right] \leq e^{\lambda \sum_{t=0}^{n} R(x_t, a_t)} U_\lambda(W(x_{n+1})),$$

para quitar el condicional del lado izquierdo, integramos de manera conveniente ambos lados de la desigualdad obteniendo

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n+1} R(X_t, A_t)} U_\lambda(W(X_{n+2})) \right] \leq E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda(W(X_{n+1})) \right]$$

y de la hipótesis de inducción se sigue

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n+1} R(X_t, A_t)} U_\lambda(W(X_{n+2})) \right] \leq U_\lambda(W(x)).$$

Por lo tanto hemos probado que para toda n se cumple que

$$U_\lambda(V_\lambda(f,x)) \geq E_x^\pi \left[e^{\sum_{t=0}^{n} R(X_t, A_t)} U_\lambda \left(V_\lambda(f, X_{n+1}) \right) \right].$$

□

Demostración del Teorema 3.2. Notemos que, debido a que $V_\lambda(f, X_k)$ es positiva y U_λ es creciente se sigue que

$$U_\lambda(V_\lambda(f, X_k)) \geq U_\lambda(0)$$

para cualquier $k \geq 1$, así

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda \left(V_\lambda(f, X_{n+1}) \right) \right] \geq E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda(0) \right]. \tag{3.11}$$

Por otro lado de la propiedad (3.6) tenemos que

$$e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda(0) = U_\lambda \left(\sum_{t=0}^{n} R(X_t, A_t) \right).$$

Combinando estos dos últimos hechos se sigue que

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} R(X_t, A_t)} U_\lambda \left(V_\lambda(f, X_{n+1}) \right) \right] \geq E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{n} R(X_t, A_t) \right) \right], \tag{3.12}$$

y usando (3.9) y (3.12) obtenemos para toda $n \in \mathbb{N}$ y para cualquier $\pi \in \mathcal{P}$ y $x \in X$ que

$$U_\lambda(V_\lambda(f,x)) \geq E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{n} R(X_t, A_t) \right) \right]. \tag{3.13}$$

Ahora consideremos los siguientes casos

$\star$ $\lambda > 0$. En este caso, como la recompensa es no negativa, tenemos

$$0 \leq U_\lambda \left[\sum_{t=0}^{n} R(X_t, A_t) \right] \nearrow U_\lambda \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right],$$

así, al tomar el límite cuando n tiende a infinito en (3.13), por el teorema de convergencia monótona obtenemos

$$U_\lambda(V_\lambda(f, x)) \geq E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right) \right], \tag{3.14}$$

o usando (3.7)

$$U_\lambda(V_\lambda(f, x)) \geq U_\lambda \left(V_\lambda(\pi, x) \right). \tag{3.15}$$

$\star$ $\lambda < 0$. En este caso, usando las propiedades de $U_\lambda(\cdot)$, y debido a que la recompensa es no negativa, tenemos que

$$U_\lambda(R(X_0, A_0)) \leq U_\lambda \left[\sum_{t=0}^{n} R(X_t, A_t) \right] \nearrow U_\lambda \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right] \leq 0.$$

Tomando el límite cuando n tiende a infinito en (3.13) el teorema de convergencia dominada aseguran, también para este caso, que ocurre

$$U_\lambda(V_\lambda(f, x)) \geq U_\lambda \left(V_\lambda(\pi, x) \right). \tag{3.16}$$

Así, para cualquier λ se ha establecido la desigualdad (3.15). Y debido a que U_λ es creciente se tiene que $V_\lambda(f, x) \geq V_\lambda(\pi, x)$ para cualquier π, entonces

$$V_\lambda(f, \cdot) \geq \mathcal{V}_\lambda(\cdot).$$

Es decir, f es óptima.

□

Observación 3.1. En [5] y [16] se presentan condiciones que garantizan el cumplimiento de la igualdad en (3.3) y (3.8).

Capítulo 4

Un Ejemplo: Caso Neutral al Riesgo

En este capítulo se detallará la solución a un ejemplo, con el cual se ilustra el problema de control de Markov con recompensa total como función objetivo (véase la sección 2.1 del capítulo anterior). Este ejemplo será llamado *ejemplo neutral al riesgo*). Originalmente planteado por Ross en [16], su importancia en esta tesis radica en ofrecer la solución a un tipo de problema de control de Markov con recompensa total pues este tipo de funciones objetivo no son fáciles de trabajar debido a la posible divergencia de la serie involucrada al sumar la recompensa por etapa. Sin embargo Ross propone una recompensa binaria lo cual le permitirá reducir el problema a una caminata aleatoria. Nuestra principal aportación en este capítulo es ofrecer ***una solución detallada del ejemplo***.

4.1. Planteamiento del Ejemplo

Consideremos el siguiente modelo de decisión de Markov.

Modelo 4.1. Para un entero positivo fijo N y un número $p \in (0,1)$, los cinco elementos del modelo quedarán descritos como sigue:

- $X := \{0, 1, 2, ..., N\}$, **el espacio de estados del sistema**.
- $A := \{0, 1, 2, ..., [N/2]\}$, **el espacio de controles**, donde $[z]$ es la parte entera de z.

- $A(x)$: Para cada $x \in X$, $A(x) = \{1, 2, ..., \min\{x, N - x\}\}$.
- Entre los estados definimos la siguiente ley de transición $Q = (q_{xy}(a))$ para $x \in X$ y $a \in A(x)$:
  1. $q_{x,x+a}(a) = p$
  2. $q_{x,x-a}(a) = q = 1 - p$
  3. $q_{N,0}(a) = 1$
  4. $q_{0,0}(a) = 1$, (nótese que esta condición implica que el cero es **absorbente**).
- Como función de respuesta definimos la recompensa por etapa como sigue:

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

Con el modelo descrito tomemos la siguiente función objetivo:

$$V(\pi, x) = E_x^{\pi} \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right),$$

es decir, la *recompensa* ***total*** *esperada* bajo una política π dado que el sistema comienza en x. El PCM asociado al modelo 4.1 con esta función objetivo es el siguiente: hallar, si es que existe, una política π^* tal que

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

y para resolverlo utilizaremos el Teorema 3.1 del capítulo anterior. Antes de iniciar propiamente el desarrollo de la solución es necesario detenernos a analizar las siguientes consecuencias sobre la sucesión de estados (véase el apéndice A), que resultan de la descripción del modelo.

De $q_{N,0}(a) = 1$ se sigue que

$$\{X_t = N\} \subset \{X_{t+k} = 0\}, \tag{4.1}$$

para todo $t = 0, 1, 2, ...$y para todo $k = 0, 1, 2, ...$y

$$\{X_m = N\} \cap \{X_n = N\} = \emptyset, \tag{4.2}$$

para todo par $m \neq n$.

Por otro lado, que el proceso alcance el estado N en el tiempo t implica que en todos los $k < t$, no ha alcanzado ni al cero ni al mismo N es decir,

$$\{X_t = N\} \subset \bigcap_{k<t} \left(\{X_k \neq N\} \cap \{X_k \neq 0\}\right). \tag{4.3}$$

Además notemos que la recompensa en la etapa j guarda la siguiente relación con el estado del sistema en j, para todo $j \in \mathbb{N}$

$$\{R(X_j, A_j) = 1\} = \{X_j = N\}, \quad \text{y} \quad \{R(X_j, A_j) = 0\} = \{X_j \neq N\}. \tag{4.4}$$

Para $i \in \mathbb{N}$, sea $B_i = \bigcup_{t=0}^{i} \{X_t = N\}$, el conjunto de trayectorias que alcanzan a N en algún t entre 0 e i; claramente $B_i \subset B_{i+1}$ para todo $i \in \mathbb{N}$. Al conjunto de trayectorias que alcanzan a N en algún tiempo $t \in \{0, 1, 2, ...\}$, lo denotamos con $B_\infty = \bigcup_{t=0}^{\infty} \{X_t = N\}$.

Lema 4.1. Bajo las condiciones del modelo para todo $m \in \mathbb{N}$ ocurre lo siguiente.

$$P\left[\sum_{t=0}^{m} R(X_t, A_t) = 1\right] = P[B_m]. \tag{4.5}$$

Demostración. A lo largo de esta prueba usaremos la siguiente notación $R_t = R(X_t, A_t)$; por inducción, veamos primero el caso $m = 0$

$$\{R_0 = 1\} = \{X_0 = N\} = B_0,$$

como el otro valor que puede tomar la recompensa es el cero se sigue que

$$\{R_0 = 0\} = \{X_0 \neq N\} = (B_0)^c.$$

Con lo cual hemos probado la base de la inducción, ahora supongamos que es cierto lo siguiente

$$\left\{\sum_{t=0}^{m-1} R_t = 0\right\} = (B_{m-1})^c \tag{4.6}$$

Usando la observación (4.4) para $i = m$ obtenemos

$$\{R_m = 0\} = \{X_m \neq N\},$$

entonces

$$\left\{\sum_{t=0}^{m} R_t = 0\right\} = \left\{\sum_{t=0}^{m-1} R_t + R_m = 0\right\} = (B_{m-1})^c \cap \{X_m \neq N\} = (B_m)^c$$

por lo cual

$$\left\{\sum_{t=0}^{m} R_t = 1\right\} = B_m$$

□

Lema 4.2. $P\left[\sum_{t=0}^{\infty} R(X_t, A_t) = 1\right] = P[B_\infty]$, donde $P = P_x^\pi$ para cualquier política π y un estado inicial x, $x \neq 0$, (véase Apéndice A).

Demostración. Por inducción puede probarse que para todo $m \in \mathbb{N}$

$$P\left[\sum_{t=0}^{m} R(X_t, A_t) = 1\right] = P[B_m]. \tag{4.7}$$

De la consecuencia (4.1) listada arriba, y de que el modelo dicta recompensa uno sólo para el estado N, se sigue que

$$\left\{\sum_{t=0}^{m} R(X_t, A_t) = 1\right\} = \left\{\sum_{t=0}^{\infty} R(X_t, A_t) = 1\right\} \quad \text{para todo} \quad m \in \mathbb{N};$$

y como $A_i \to A_\infty$, al tomar el límite cuando $m \to \infty$ en (4.7) obtenemos directamente que

$$P\left[\sum_{t=0}^{\infty} R(X_t, A_t) = 1\right] = P\left[\bigcup_{t=0}^{\infty} \{X_t = N\}\right].$$

□

Observación 4.1. Ahora, como los conjuntos $\{X_t = N\}$ son ajenos se sigue que

$$P\left[\sum_{t=0}^{\infty} R(X_t, A_t) = 1\right] = \sum_{t=0}^{\infty} P\left[\{X_t = N\}\right].$$

y, del hecho de que el estado N de alcanzarse, se alcanza una sola vez, concluimos que el único otro valor posible para la serie es cero. De aquí se desprende que

$$\begin{aligned} V(\pi, x) &= E_x^{\pi}\left[\sum_{t=0}^{\infty} R(X_t, A_t)\right] \\ &= \sum_{k=0,1} k P_x^{\pi}\left[\sum_{t=0}^{\infty} R(X_t, A_t) = k\right] \\ &= P_x^{\pi}\left(\sum_{t=0}^{\infty} R(X_t, A_t) = 1\right). \end{aligned} \tag{4.8}$$

es decir

$$V(\pi, x) = \sum_{t=0}^{\infty} P_x^{\pi}\left(X_t = N\right). \tag{4.9}$$

Así el problema de control óptimo que consiste en *maximizar la esperanza de la recompensa total ganada* es **equivalente** a *maximizar la probabilidad de que el sistema alcance el valor N antes que a cero.*

4.2. Solución

Para hallar la solución buscada primero la idea es encontrar cómo se ve la desigualdad de optimalidad (3.1) bajo las condiciones del modelo descrito al principio de este capítulo. Recordemos que N es arbitrario pero fijo, x se usará para denotar el estado actual del sistema y a para denotar el control aplicado a esta x. Distinguiremos tres grupos de estados $\{0\}$, $\{N\}$ y $\{1, ..., N-1\}$. Para los dos primeros, y cualquier política estacionaria f, la desigualdad (3.3) siempre se cumple, veamos cada caso.

Caso x=0. Basta notar que los siguientes hechos ocurren bajo cualquier política estacionaria: el único control para cero es el cero mismo i.e., $A(0) = \{0\}$, así al único estado al que el sistema puede moverse en un paso es también al mismo cero así $q_{00}(0) = 1$, por otro lado la descripción del modelo dicta recompensa cero para este estado (i.e. $R(0,0) = 0$), así pues lo que debe verificarse es lo siguiente.

$$V(f,0) \geq V(f,0)q_{00}(0).$$

Lo cual es claramente cierto.

Caso x=N. Basta notar que los siguientes hechos ocurren bajo cualquier política estacionaria: el único control para el estado N es el cero i.e., $A(N) = \{0\}$, y por definición del modelo al único estado al que el sistema puede moverse desde N en un paso es al cero i.e, $q_{N0}(0) = 1$, por otro lado la descripción del modelo dicta recompensa uno para este estado ($R(N,0) = 1$), de modo tal que la desigualdad por verificar queda como sigue:

$$V(f,N) \geq 1 + V(f,0).$$

Pero es claro que si el estado inicial es N, la recompensa total ganada, bajo cualquier política es uno, de donde $V(f,N) = 1$ y que si el estado inicial es cero, la recompensa total ganada, bajo cualquier política (en particular bajo cualquier estacionaria), es cero i.e., $V(f,0) = 0$. Una vez más lo que obtenemos es la igualdad.

Ahora veamos qué pasa con el tercer conjunto de estados.

Observación 4.2. Sea f una política estacionaria. Bajo las condiciones del Modelo 4.1, para el conjunto de estados $\{1, ..., N-1\}$ la Desigualdad de Optimalidad (3.3), toma la siguiente forma

$$V(f,x) \geq pV(f,x+a) + qV(f,x-a) \quad \text{para} \quad a \leq \min\{x, N-x\}. \tag{4.10}$$

Esto se sigue de los siguientes hechos: para los estados en cuestión, el modelo dicta recompensa cero. Así (3.3) se reduce en primer lugar a lo siguiente

$$V(f,x) \geq \sum_y V(f,y)q_{xy}(a) \quad \text{para toda} \quad a \in A(x). \tag{4.11}$$

Un segundo hecho es que para cada a elegido por la política f hay sólo dos estados y a los que el sistema puede moverse desde x, a saber $y = x + a$ y $y = x - a$, con las probabilidades definidas por la ley de transición Q dada en el modelo, es decir $q_{x,x+a}(a) = p$ y $q_{x,x-a}(a) = q$, de modo que la serie de (4.11) tiene sólo dos sumandos, es decir

$$\begin{aligned}\sum_y V(f,y)q_{xy}(a) &= \sum_{y=x-a,x+a} q_{x,y}(a)V(f,y) \\ &= pV(f,x+a) + qV(f,x-a).\end{aligned}$$

Ahora, para un estado x, el modelo permite que sus controles sean a lo más $\min\{x, N-x\}$, así que (4.11) queda finalmente como sigue.

$$V(f,x) \geq pV(f,x+a) + qV(f,x-a) \quad \text{para} \quad a \leq \min\{x, N-x\}.$$

En resumen hemos obtenido que, bajo las condiciones del modelo, para afirmar que una política estacionaria f es óptima, basta verificar que su correspondiente $V(f,x)$ cumple con (4.10) en el conjunto de estados $\{1, 2, ..., N-1\}$.

Ahora bien, la idea es dividir el problema en dos casos, a saber: $p \geq q$ y $p \leq q$. En cada caso se propone una política estacionaria y se probará que la esperanza de la recompensa total obtenida bajo ella cumple con (4.10).

Caso $p \geq q$

Se define la política ***tímida*** τ como la que siempre elige el control $a = 1$, es decir

$$\tau(x) = 1 \quad \text{para todo} \quad x \in X.$$

Claramente τ es estacionaria, bajo esta política ocurre que el sistema se aproxima o se aleja de su objetivo N con pasos de tamaño uno. Ahora nos será útil la Observación 4.1 que derivó en la ecuación (4.9) y así para τ tenemos que

$$V(\tau,x) = \sum_{t=0}^{\infty} P_x^{\tau}\left(X_t = N\right). \tag{4.12}$$

Observemos que bajo esta política el proceso $\{X_t\}$ se comporta como una caminata aleatoria con paso de tamaño uno, con la siguiente matriz de transición.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0 & 1 & 2 & 3 & ... & N-2 & N-1 & N \end{array} \\ \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ \mathbf{q} & 0 & \mathbf{p} & 0 & ... & 0 & 0 & 0 \\ 0 & \mathbf{q} & 0 & \mathbf{p} & ... & 0 & 0 & 0 \\ 0 & 0 & \mathbf{q} & 0 & ... & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & ... & \mathbf{q} & 0 & \mathbf{p} \\ \mathbf{1} & 0 & 0 & 0 & ... & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Y calcular la probabilidad de que este proceso llegue alguna vez a N antes que a cero, dado que se inicia en x es lo que tenemos en el siguiente resultado.

Lema 4.3. La probabilidad de que, dado que el estado inicial es x, el proceso alcance N antes de llegar al cero es:

$$V(\tau, x) = \sum_{t=0}^{\infty} P_x^{\tau}\left(X_t = N\right) = \begin{cases} \dfrac{1-\left(\frac{q}{p}\right)^x}{1-\left(\frac{q}{p}\right)^N}, & p \neq \frac{1}{2}; \\ \dfrac{x}{N}, & p = \frac{1}{2}, \end{cases} \tag{4.13}$$

donde $q = 1 - p$.

Demostración. Sea ϱ_x la probabilidad de que el proceso alcance el estado N antes que al 0, dado que el estado inicial es x. Por la definición del modelo tenemos que $\varrho_0 = 0$ y $\varrho_N = 1$. El proceso tiene la siguiente dinámica: desde el estado x con probabilidad q llegará al estado $x-1$ y con probabilidad p llegará al estado $x+1$. Entonces debe ser claro que existe la siguiente relación entre las $\varrho_x{}'s$.

$$p\varrho_x = p\varrho_{x+1} + q\varrho_{x-1}, \quad \text{donde} \quad 1 \leq x \leq N-1.$$

Como $p + q = 1$ esta igualdad podemos escribirla como

$$p(\varrho_x - \varrho_{x+1}) = q(\varrho_{x-1} - \varrho_x),$$

o bien

$$\varrho_x - \varrho_{x+1} = \frac{q}{p}(\varrho_{x-1} - \varrho_x),$$

de aquí no es difícil obtener que

$$\varrho_x - \varrho_{x+1} = \left(\frac{q}{p}\right)^x (\varrho_0 - \varrho_1),$$

es decir,

$$\varrho_x - \varrho_{x+1} = -\varrho_1 \left(\frac{q}{p}\right)^x. \tag{4.14}$$

Por otro lado, usando una suma telescópica y las condiciones iniciales tenemos

$$\sum_{x=0}^{N-1}(\varrho_x - \varrho_{x+1}) = \varrho_0 - \varrho_N = -1,$$

combinando estas dos últimas ecuaciones obtenemos

$$\varrho_1 \sum_{x=0}^{N-1} \left(\frac{q}{p}\right)^x = 1, \tag{4.15}$$

cuando $p \neq q$ esta última expresión toma la forma

$$\varrho_1 \frac{\left(\frac{q}{p}\right)^N - 1}{\frac{q}{p} - 1} = 1;$$

es decir,

$$\varrho_1 = \frac{\frac{q}{p} - 1}{\left(\frac{q}{p}\right)^N - 1}. \tag{4.16}$$

Además para cualquier z tal que $1 \leq z \leq N$ podemos escribir

$$\sum_{x=0}^{z-1}(\varrho_x - \varrho_{x+1}) = \varrho_0 - \varrho_z = -\varrho_z$$

o bien por (4.14)

$$\varrho_1 \sum_{x=0}^{z-1} \left(\frac{q}{p}\right)^x = \varrho_z.$$

Usando el valor de ϱ_1 calculado en (4.16) y calculando la suma llegamos a

$$\varrho_1 \frac{\left(\frac{q}{p}\right)^z - 1}{\frac{q}{p} - 1} = \varrho_z. \tag{4.17}$$

Aquí usamos otra vez el valor obtenido para ϱ_1 en (4.16) y obtenemos

$$\varrho_z = \frac{1 - \left(\frac{q}{p}\right)^N}{1 - \frac{q}{p}}.$$

Probando así el resultado para $p \neq q$. En el caso $p = q$ las ecuaciones (4.15) y (4.17) toman la forma

$$\varrho_1 = \frac{1}{N}; \qquad z\varrho_1 = \varrho_z$$

de donde

$$\varrho_z = \frac{z}{N}.$$

Así se ha obtenido el resultado para los dos casos. □

Ahora podemos plantear el siguiente resultado.

Teorema 4.1. Si $p \leq \frac{1}{2}$, la política tímida τ maximiza la esperanza de la recompensa total ganada.

Demostración. Debe probarse que (4.13), que es una forma de escribir la esperanza de la recompensa total ganada bajo τ dado que se inicia en x, cumple con (4.10) que es la versión adecuada de la desigualdad de optimalidad (3.1) bajo las condiciones del modelo.
En el caso $p = q = \frac{1}{2}$ es fácil verificar que $V(\tau, x) = \frac{x}{N}$ sí satisface (4.10) y por lo tanto τ en este caso es óptima.
Cuando $p > \frac{1}{2}$ debe probarse que

$$\frac{1-\left(\frac{q}{p}\right)^{x}}{1-\left(\frac{q}{p}\right)^{N}} \geq p\left[\frac{1-\left(\frac{q}{p}\right)^{x+a}}{1-\left(\frac{q}{p}\right)^{N}}\right]+q\left[\frac{1-\left(\frac{q}{p}\right)^{x-a}}{1-\left(\frac{q}{p}\right)^{N}}\right]$$

es decir,

$$\left(\frac{q}{p}\right)^{x} \leq p\left(\frac{q}{p}\right)^{x+a}+q\left(\frac{q}{p}\right)^{x-a}$$

equivalentemente

$$1 \leq p\left(\frac{q}{p}\right)^{a}+q\left(\frac{p}{q}\right)^{a}$$

reescribiendo esto

$$1 \leq p\left[\left(\frac{q}{p}\right)^{a}+\left(\frac{p}{q}\right)^{a-1}\right].$$

Ahora debe notarse que esto sí se vale para $a = 1$, simplemente porque

$$1 \leq p\left(\frac{q}{p}+1\right) = 1;$$

y para los a mayores que uno, también es cierto usando el Lema 4.5 para la expresión entre corchetes haciendo $r = \frac{q}{p}$. Por lo tanto

$$V(\tau, x) \geq pV(\tau, x+a)+qV(\tau, x-a), \qquad x \in \{1, ..., N-1\}$$

Por lo tanto podemos afirmar que τ es óptima cuando $p > q$. □

Así, se ha probado que cuando $p \geq q$ la política óptima es τ.

Caso $p \leq q$

Sea α la siguiente política

$\alpha(x) = x$ si $x \leq N/2$.

$\alpha(x) = N - x$ si $x \geq N/2$.

A diferencia de la política tímida que en cada paso elige la acción 1 , esta política elige una acción más agresiva, pues cuando el sistema está en un estado $x \leq N/2$, α elige la acción que le permitiría avanzar el doble del estado actual y cuando sistema está en $x \geq N/2$, su elección es todo lo necesario para que, de avanzar, llegue al objetivo N; por esta razón llamaremos a esta política ***audaz***.

Recordemos que con $V(\alpha, x)$ se denota la esperanza de la recompensa total ganada bajo α y dado que el estado inicial es x y por la Observación 4.1 y por la ecuación (4.9) podemos representarla con

$$V(\alpha, x) = \sum_{t=0}^{\infty} P_x^{\alpha} \left(X_t = N \right). \tag{4.18}$$

Ahora, en lugar de considerar la recompensa total supongamos que tenemos un número finito n de intentos permitido para alcanzar el objetivo N y entonces usaremos la siguiente notación

$$V^n(\alpha, x) = P_x^{\alpha,n} \left(\bigcup_{t=0}^{n} \{X_t = N\} \right) = \sum_{t=0}^{n} P_x^{\alpha,n} \left(X_t = N \right), \tag{4.19}$$

donde $P_x^{\alpha,n} \left(X_t = N \right)$ denota la probabilidad condicional de que el proceso alcance el estado N (antes que el cero), en n pasos dado que se inició en el estado x y, por supuesto, estamos bajo la política α. Notemos que el teorema de convergencia monótona garantiza que

$$\lim_{n \to \infty} V^n(\alpha, x) = V(\alpha, x). \tag{4.20}$$

Bajo α, la dinámica del proceso depende de que el estado actual del sistema sea menor o mayor que $N/2$, analicemos esto con cuidado. Cuando $x \leq N/2$ la estrategia dicta que con probabilidad p el sistema se mueve hacia $2x$ y hacia cero con probabilidad q. Por otro lado si $x \geq N/2$, el control elegido deriva en que el sistema se mueva hacia N con probabilidad p y hacia $2x - N$ con probabilidad q. Si además tomamos en cuenta a n, el número de intentos permitidos, está claro que después de cada intento queda uno menos por hacer pero ahora desde el estado al que se llegó. La dinámica del proceso queda entonces descrito con la siguiente ecuación en diferencias. Para $n \in \mathbb{N}$ definimos

$$V^n(\alpha, x) = \begin{cases} pV^{n-1}(\alpha, 2x), & x \leq \frac{N}{2}; \\ p + qV^{n-1}(\alpha, 2x - N), & x \geq \frac{N}{2}. \end{cases} \tag{4.21}$$

Las condiciones iniciales que siguen también tienen sentido en el contexto del modelo a saber, para todo $n \in \mathbb{N}$

* $V^n(\alpha, 0) = 0$, dado que el estado inicial es cero, es claro que la recompensa es cero para cualquier número de intentos permitidos.

* $V^n(\alpha, N) = 1$, si el estado inicial es uno, el siguiente sólo puede ser cero y entonces para cualquier $n \geq 0$ la esperanza es uno.

* $V^0(\alpha, x) = 0$, si no hay intentos permitidos entonces la esperanza es cero para $x < N$.

Lema 4.4. Cuando $p \leq \frac{1}{2}$, para cada $n > 0$ la política audaz maximiza la probabilidad de alcanzar el estado N en un tiempo $n \geq 0$.

Demostración. Debemos probar que la expresión obtenida en (4.21) para la esperanza de la recompensa ganada bajo α dado que se inicia en x y se tienen n intentos permitidos cumple con la condición (4.10)

$$V(\alpha, x) \geq pV(\alpha, x + a) + qV(\alpha, x - a) \quad \text{para} \quad a \leq \min\{x, N - x\},$$

que es la versión adecuada de la desigualdad de optimalidad (3.1), bajo las condiciones del modelo. Pero además debemos tomar en cuenta los n intentos que se tienen, esto se convierte en

$$V^n(\alpha, x) - pV^{n-1}(\alpha, x + a) - qV^{n-1}(\alpha, x - a) \geq 0 \tag{4.22}$$

para $\quad a \leq \min\{x, N - x\}$.
Pero esto es lo que se afirma en el Lema 4.6. □

Entonces hemos probado que cuando $p \leq 1/2$ la estrategia audaz (α) maximiza la probabilidad de alcanzar N para cualquier n. Cuando el número de intentos no está limitado tendremos el siguiente resultado.

Teorema 4.2. Cuando $p \leq \frac{1}{2}$ la política audaz es la que maximiza la probabilidad de alcanzar N.

Demostración. Por la observación (4.20) al tomar el siguiente límite

$$\lim_{n\to\infty}[V^n(\alpha,x)-pV^{n-1}(\alpha,x+a)-qV^{n-1}(\alpha,x-a)\geq 0] \quad \text{para} \quad a\leq \min\{x,N-x\},$$

obtenemos

$$V(\alpha,x)-pV(\alpha,x+a)-qV(\alpha,x-a)\geq 0 \quad \text{para} \quad a\leq \min\{x,N-x\},$$

pero esto implica que α es óptima. □

Concluimos pues habiendo hallado dos políticas óptimas para el ejemplo de PCM planteado al inicio del capítulo, cada una de ellas en los casos particulares $p \geq q$ y $p \leq q$. Como ya se dijo a este ejemplo lo llamamos *ejemplo neutral al riesgo* para distinguirlo del ejemplo que será discutido en el siguiente capítulo.

4.3. Demostraciones de Resultados Auxiliares

En esta sección se encontrarán los resultados técnicos que fueron usados para sustentar la soluciones que fueron presentadas en la anterior.

Lema 4.5. Si $r \in (0,1]$, entonces $h(x) = (r)^x + (\frac{1}{r})^{x-1}$, $x \in [1,\infty)$ es creciente.

Demostración. En efecto, basta calcular la derivada de h con respecto a x y verificar que es positiva

$$\begin{aligned} h'(x) &= (r)^x \log(r) + \left(\frac{1}{r}\right)^{x-1} \log\left(\frac{1}{r}\right) \\ &= -(r)^x \log\left(\frac{1}{r}\right) + \left(\frac{1}{r}\right)^{x-1} \log\left(\frac{1}{r}\right) \\ &= \left[\left(\frac{1}{r}\right)^{x-1} - (r)^x\right] \log\left(\frac{1}{r}\right) \end{aligned}$$

pero esto es mayor que cero debido a porque $x \geq 1$ y $r \in (0,1]$. Por lo tanto h es creciente. □

La siguiente es la demostración de lo afirmado en el Teorema 4.2. Por simplicidad prescindiremos en la notación de α, para ello denotamos con $\mathbb{V}^k(z)$ a la $V^n(\alpha, x)$ de la sección anterior. Hecha la aclaración consideremos la siguiente ecuación en diferencias.

$$\mathbb{V}^k(z) = \begin{cases} p\mathbb{V}^{k-1}(2z), & z \leq \frac{N}{2}; \\ p + q\mathbb{V}^{k-1}(2z - N), & z \geq \frac{N}{2}, \end{cases} \tag{4.23}$$

con las siguientes condiciones iniciales y de frontera:

$$\mathbb{V}^k(0) = 0, \quad \mathbb{V}^k(N) = 1, \quad k \geq 0, \quad \mathbb{V}^0(z) = 0, \quad z < N.$$

Lema 4.6. Si $p \leq \frac{1}{2}$ entonces para todo $n > 0$,

$$\mathbb{V}^{k+1}(x) - q\mathbb{V}^k(x - a) - p\mathbb{V}^k(x + a) \geq 0, \quad a \leq \min\{x, N - x\}. \tag{4.24}$$

Demostración. En efecto. Para $k = 0$, usando las condiciones de frontera puede verificarse directamente que

$$\mathbb{V}^1(x) - q\mathbb{V}^0(x - a) - p\mathbb{V}^0(x + a) = \mathbb{V}^1(x) \geq 0$$

sucede tanto para $x \leq N/2$ como para $x \geq N/2$.

Ahora supongamos que vale lo siguiente

$$\mathbb{V}^k(l) - q\mathbb{V}^{k-1}(l - m) - p\mathbb{V}^{k-1}(l + m) \geq 0, \quad m \leq \min\{l, N - l\}. \tag{4.25}$$

`CASO 1.` $x + a \leq \frac{N}{2}$.

En este caso ocurre también lo siguiente $x - a \leq \frac{N}{2}$, $x \leq \frac{N}{2}$, con estas suposiciones y usando (4.23) adecuadamente para x, $x + a$ y $x - a$ en (4.24) obtenemos

$$p\mathbb{V}^k(2x) - p(p(\mathbb{V}^{k-1}(2x + 2a)) - q(p\mathbb{V}^{k-1}(2x - 2a)),$$

es decir

$$p[\mathbb{V}^k(2x) - p\mathbb{V}^{k-1}(2x+2a) - q\mathbb{V}^{k-1}(2x-2a)],$$

pero esto es mayor que cero pues $p > 0$, y

$$\mathbb{V}^k(2x) - p\mathbb{V}^{k-1}(2x+2a) - q\mathbb{V}^{k-1}(2x-2a)$$

también, basta tomar $l = 2x$ y $m = 2a$ en la hipótesis de inducción (4.25).

CASO 2. $x - a \geq \frac{N}{2}$.

En este caso se tiene además que $x+a \geq \frac{N}{2}$ y también que $x \geq \frac{N}{2}$; con estas suposiciones y una vez más usando (4.23) adecuadamente para x, $x+a$, y $x-a$ en (4.24) obtenemos

$$p + q\mathbb{V}^k(2x-N) - p(p + q(\mathbb{V}^{k-1}(2x+2a-N)) - q(p + q\mathbb{V}^{k-1}(2x-2a-N))$$

es decir

$$p - p^2 - qp + q[\mathbb{V}^k(2x-N) - p\mathbb{V}^{k-1}(2x+2a-N) - q\mathbb{V}^{k-1}(2x-2a-N)].$$

Ahora nótese que los tres primeros sumandos se reducen a cero, así: $p - p^2 - qp = 0$ y lo que está entre corchetes es mayor que cero haciendo $l = 2x - N$ y $m = 2a$ en la hipótesis de inducción (4.25).

CASO 3. $x \leq \frac{N}{2} \leq x + a$.

Recuérdese que $a \leq x$, además es claro que $x - a \leq x \leq \frac{N}{2}$.
Usando adecuadamente la definición (4.23) vemos que, lo que ha de probarse es

$$p\mathbb{V}^k(2x) - p(p + q(\mathbb{V}^{k-1}(2(x+a) - N))) - q(p\mathbb{V}^{k-1}(2(x-a))) \geq 0$$

en el lado izquierdo tenemos

$$p[\mathbb{V}^k(2x) - p - q\mathbb{V}^{k-1}(2x+2a-N) - qp\mathbb{V}^{k-1}(2x-2a)].$$

Ahora nótese que $2x \geq x + a \geq \frac{N}{2}$ así que puede continuarse como sigue

$$p[p + q\mathbb{V}^{k-1}(4x-N) - p - q\mathbb{V}^{k-1}(2x+2a-N) - q\mathbb{V}^{k-1}(2x-2a)]$$

o equivalentemente

$$q[p\mathbb{V}^{k-1}(4x-N)-p\mathbb{V}^{k-1}(2x+2a-N)-p\mathbb{V}^{k-1}(2x-2a)].$$

Ahora nótese que $4x-N=2(2x-\frac{N}{2})$ y que $2x-\frac{N}{2}\leq\frac{N}{2}$, por lo tanto la última expresión es equivalente a

$$q[\mathbb{V}^{k}(2x-N/2)-p\mathbb{V}^{k-1}(2x+2a-N)-p\mathbb{V}^{k-1}(2x-2a)].$$

Basta probar que

$$\mathbb{V}^{k}(2x-N/2)-p\mathbb{V}^{k-1}(2x+2a-N)-p\mathbb{V}^{k-1}(2x-2a)\geq 0. \qquad (4.26)$$

Pero esto es cierto para cualquiera de los dos siguientes casos. Cuando $a\geq\frac{N}{4}$ como $p<q$ se tiene que (4.26) vale al menos

$$\mathbb{V}^{k-1}(2x-\frac{N}{2})-p\mathbb{V}^{k-1}(2x+2a-N)-q\mathbb{V}^{k-1}(2x-2a)$$

pero esto sí es mayor que cero, basta usar $l=2x-\frac{N}{2}$ y $m=2a-\frac{N}{2}$ en la hipótesis de inducción. Ahora en el caso $a\leq\frac{N}{4}$ (4.26) vale al menos

$$\mathbb{V}^{k-1}(2x-\frac{N}{2})-q\mathbb{V}^{k-1}(2x+2a-N)-p\mathbb{V}^{k-1}(2x-2a)$$

para ver que esto sí es mayor que cero, basta usar $l=2x-\frac{N}{2}$ y $m=\frac{N}{2}-2a$ en la hipótesis de inducción.

CASO 4. $x-a\leq\frac{N}{2}\leq x$

De la suposición se sigue que $\frac{N}{2}\leq x+a$, así que usando adecuadamente (4.23) en el lado izquierdo de (4.24) se obtiene

$$p+q\mathbb{V}^{k}(2x-N)-p(p+q\mathbb{V}^{k-1}(2x+2a-N))-q(p\mathbb{V}^{k-1}(2x-2a))$$

para probar que esto es mayor que cero primero nótese que de las suposiciones se obtiene que $x\leq\frac{3N}{4}$ y de esto último se desprende que $2x-N\leq\frac{N}{2}$; así usando esta nueva información en la última expresión se obtiene

$$pq + qp\mathbb{V}^{k-1}(4x - 2N) - pq\mathbb{V}^{k-1}(2x + 2a - N)) - pq\mathbb{V}^{k-1}(2x - 2a) \quad (4.27)$$

Hacemos un paréntesis para notar que $2x - \frac{N}{2} \geq \frac{N}{2}$ (simplemente porque $x \geq \frac{N}{2}$), por lo tanto

$$\mathbb{V}^k(2x - \frac{N}{2}) = p + q\mathbb{V}^{k-1}(2(2x - \frac{N}{2}) - N) = p + q\mathbb{V}^{k-1}(4x - 2N)$$

es decir

$$\mathbb{V}^k(2x - \frac{N}{2}) - p = q\mathbb{V}^{k-1}(4x - 2N).$$

Cerramos el paréntesis y sustituyendo en (4.27) esta nueva información, se obtiene

$$pq + p(\mathbb{V}^k(2x - \frac{N}{2}) - p) - pq\mathbb{V}^{k-1}(2x + 2a - N) - pq\mathbb{V}^{k-1}(2x - 2a),$$

es decir,

$$p(q - p) + p[\mathbb{V}^k(2x - \frac{N}{2}) - q\mathbb{V}^{k-1}(2x + 2a - N) - q\mathbb{V}^{k-1}(2x - 2a)].$$

El primer sumando de esta expresión es claramente positivo pues $p < q$, resta probar que lo que está en el interior de los corchetes es positivo. Cuando $a \geq \frac{N}{4}$ esto vale al menos

$$\mathbb{V}^k(2x - \frac{N}{2}) - p\mathbb{V}^{k-1}(2x + 2a - N) - q\mathbb{V}^{k-1}(2x - 2a)$$

lo cual es positivo tomando en la hipótesis de inducción $l = 2x - \frac{N}{2}$ y $m = 2a - \frac{N}{2}$. El caso $a \geq \frac{N}{4}$ es análogo.

Finalmente se ha conseguido probar que para cualquier caso se verifica que

$$\mathbb{V}^{k+1}(x) - q\mathbb{V}^k(x - a) - p\mathbb{V}^k(x + a) \geq 0, \quad a \leq \min\{x, N - x\}.$$

□

Capítulo 5

Un Ejemplo: Caso Sensible al Riesgo

En el presente capítulo se encuentra la aportación más importante de este trabajo: la extensión al ejemplo del problema de control propuesto en el capítulo anterior, a esta extensión la llamaremos *ejemplo sensible al riesgo*. La idea es mantener el modelo tal cual y proponer una función objetivo que es de hecho una familia de funciones objetivo parametrizada por un número real $\lambda \neq 0$. Técnicamente entonces estamos hablando también de una familia de ejemplos, pero siempre nos referiremos a la extensión en singular. La función objetivo que genera este nuevo ejemplo de problema de control óptimo está fuertemente relacionada con la anterior pues también tiene que ver con la recompensa total, será de hecho llamada *función objetivo sensible al riesgo*. Sin embargo una diferencia fundamental es el hecho de considerar una *función de utilidad* para plantearla y, a través de ella, considerar la actitud que tiene el controlador en el momento de elegir el control.

5.1. Planteamiento del Ejemplo

Consideremos el Modelo 4.1 descrito en el capítulo anterior y suponiendo $\lambda \neq 0$, tomemos la siguiente función objetivo:

$$V_\lambda(\pi, x) = \frac{1}{\lambda} ln \left(E_x^\pi \left[e^{\lambda \left(\sum_{t=0}^{\infty} R(X_t, A_t) \right)} \right] \right),$$

a la cual llamamos en el Capítulo 2 *recompensa total esperada sensible al riesgo* o λ-*recompensa* ***total*** *esperada* bajo una política π dado que el sistema comienza en x. Recordemos que el problema de control de Markov asociado al modelo con esta función objetivo es el siguiente: hallar, si es que existe, una política π^* tal que

$$V_\lambda(\pi^*, x) = \sup_{\pi \in \mathcal{P}} V_\lambda(\pi, x), \quad x \in X.$$

A partir de ahora nos referiremos a este ejemplo como el ejemplo de PCM sensible al riesgo. Para resolverlo recordemos primero que la función objetivo $V_\lambda(\pi, x)$ está relacionada con la certeza equivalente $\mathcal{Q}$ de la variable aleatoria $Y = \sum_{t=0}^{\infty} R(X_t, A_t)$ es decir,

$$V_\lambda(\pi, x) = \mathcal{Q}_\lambda(Y)$$

por lo tanto de (2.6) se sigue que

$$U_\lambda(V_\lambda(\pi, x)) = E_x^\pi [U_\lambda(Y)]. \tag{5.1}$$

Por otro lado, las observaciones hechas en el capítulo anterior para $V(\pi, x)$ se siguen manteniendo, en particular de (4.8) se sigue que

$$V(\pi, x) = 1 - P_x^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) = 0 \right], \tag{5.2}$$

y

$$V(\pi, x) = \sum_{t=0}^{\infty} P_x^\pi (X_t = N). \tag{5.3}$$

5.2. Solución

La idea es usar el Teorema 3.2, para ello notemos que para los estados 0 y N la desigualdad (3.8) se cumple para cualquier política estacionaria, este hecho quedará explicado en la siguiente observación.

Observación 5.1. Para una política estacionaria f, si $x = 0$ ó N entonces

$$U_\lambda(V_\lambda(f, x)) \geq e^{\lambda R(x,a)} \left[\sum_y U_\lambda(V_\lambda(f, y)) q_{xy}(a) \right],$$

para todo $a \in A(x)$.

Esto se sigue directamente de aplicar las condiciones del Modelo 4.1 para cada x.

x=0. Para este estado la recompensa es cero i.e. $e^{\lambda R(0,a)} = 1$, además el único elemento del conjunto $\{A(0)\}$ es el mismo cero así que la suma del lado derecho de la desigualdad tiene un único sumando por lo tanto lo que debe probarse es

$$U_\lambda(V_\lambda(f,0)) \geq U_\lambda(V_\lambda(f,0))q_{00}(0),$$

pero $q_{00}(0) = 1$, por lo cual obtenemos la igualdad.

x=N. En este caso la recompensa es uno, el único control permitido es el cero y, por las condiciones del modelo, el único estado al que se permite moverse desde N es al cero así lo que hay que probar es

$$U_\lambda(V_\lambda(f,N)) \geq e^\lambda U_\lambda(V_\lambda(f,0))q_{N0}(0);$$

otra consecuencia del modelo es que $q_{N0}(0) = 1$, usando esto y la definición de V_λ la última expresión se reescribe como

$$U_\lambda\left(\frac{1}{\lambda}ln\left(E_N^f\left[e^{\lambda\sum_{t=0}^{\infty}R(X_t,A_t)}\right]\right)\right) \geq e^\lambda U_\lambda\left(\frac{1}{\lambda}ln\left(E_0^f\left[e^{\lambda\sum_{t=0}^{\infty}R(X_t,A_t)}\right]\right)\right).$$

Pero es claro que $\sum_{t=0}^{\infty}R(X_t,A_t)$ suma uno cuando el estado inicial es N, y cero cuando el estado inicial es cero (véase 4.1), entonces

$$U_\lambda\left(\frac{1}{\lambda}ln\left(E_N^f\left[e^\lambda\right]\right)\right) \geq e^\lambda U_\lambda\left(\frac{1}{\lambda}ln\left(E_0^f\left[1\right]\right)\right)$$

finalmente, es claro que

$$U_\lambda(1) \geq e^\lambda U_\lambda(0) = U_\lambda(1).$$

Así tenemos que para cualquier estrategia estacionaria, los estados 0 y N cumplen con la desigualdad (3.8).

Por otro lado, para el resto de los estados analicemos lo siguiente.

Observación 5.2. Para cualquier política estacionaria g, la desigualdad (3.8) toma la siguiente forma para los estados $\{1, 2, ..., N-1\}$

$$U_\lambda(V_\lambda(g,x)) \geq pU_\lambda(V_\lambda(g,x+a))+qU_\lambda(V_\lambda(g,x-a)), \quad a \leq \min\{x, N-x\}. \quad (5.4)$$

Esto se sigue trivialmente de los siguientes hechos: el modelo prescribe para estos estados recompensa cero, de tal manera que (3.8) se reduce a:

$$U_\lambda(V_\lambda(g,x)) \geq \sum_y U_\lambda(V_\lambda(g,y))q_{xy}(a),$$

y por otro lado es claro que desde el estado actual x, hay sólo dos estados y a los que el sistema puede moverse, a saber $y = x + a$ y $y = x - a$ con las probabilidades descritas por el modelo. Así, la última expresión se puede reescribir como afirmamos.

Finalmente, del mismo modo que ocurrió con el caso neutral al riesgo, hemos obtenido que, bajo las condiciones del modelo, para asegurar que una política estacionaria f es óptima, basta verificar que su correspondiente $V(f,x)$ cumple con (5.4) para todos los controles tales que $a \leq \min\{x, N-x\}$.

De manera similar a lo hecho en el capítulo anterior, distinguiremos dos casos con respecto al valor que tome p.

CASO $p \geq q$.

Sea τ, la política tímida del capítulo anterior; usando (5.1) y la definición de la función U_λ calculamos $U_\lambda(V_\lambda(\tau,x))$:

$$\begin{aligned} U_\lambda(V_\lambda(\tau,x)) &= \mathrm{E}_x^\tau \left[\mathrm{sign}(\lambda) e^{\lambda \sum_{t=0}^{\infty} R(X_t,A_t)} \right] \\ &= \mathrm{sign}(\lambda) \left(\sum_{k=0,1} e^{\lambda k} P_x^\tau \left[\sum_{t=0}^{\infty} R(X_t,A_t) = k \right] \right) \\ &= \mathrm{sign}(\lambda) \left(e^\lambda P_x^\tau \left[\sum_{t=0}^{\infty} R(X_t,A_t) = 1 \right] + P_x^\tau \left[\sum_{t=0}^{\infty} R(X_t,A_t) = 0 \right] \right). \end{aligned}$$

Se sigue de las igualdades dadas en (5.2) y (5.3) que

$$U_\lambda(V_\lambda(\tau, x)) = \text{sign}(\lambda)[(e^\lambda - 1)V(\tau, x) + 1]. \tag{5.5}$$

Donde $V(\tau, x)$ es la esperanza de la recompensa total ganada bajo τ cuando el sistema comienza en x en el caso neutral al riesgo. Para el conjunto de estados $\{1, ..., N-1\}$ tenemos el siguiente resultado.

Lema 5.1. $U_\lambda(V_\lambda(\tau, x))$ cumple con la desigualdad (5.4) es decir,

$$U_\lambda(V_\lambda(\tau, x)) \geq pU_\lambda(V_\lambda(\tau, x + a)) + qU_\lambda(V_\lambda(\tau, x - a)),$$

$a \leq \min\{x, N - x\}$.

Demostración. Es necesario verificar los dos casos de λ:

$\lambda > 0$**:** en efecto, por el Teorema 4.1 sabemos que

$$V(\tau, x) \geq pV(\tau, x + a) + qV(\tau, x - a)$$

debido a la condición sobre λ, sucede que $(e^\lambda - 1) > 0$ así obtenemos

$$(e^\lambda - 1)V(\tau, x) \geq p(e^\lambda - 1)V(\tau, x + a) + q(e^\lambda - 1)V(\tau, x - a)$$

usando el hecho de que $p + q = 1$ se consigue

$$(e^\lambda - 1)V(\tau, x) + 1 \geq ((e^\lambda - 1)V(\tau, x + a) + 1)p + ((e^\lambda - 1)V(\tau, x - a) + 1)q$$

es decir

$$(e^\lambda - 1)V(\tau, x) + 1 \geq ((e^\lambda - 1)V(\tau, x + a) + 1)p + ((e^\lambda - 1)V(\tau, x - a) + 1)q$$

lo cual prueba que

$$U_\lambda(V_\lambda(\tau, x)) \geq U_\lambda(V_\lambda(\tau, x + a))p + U_\lambda(V_\lambda(\tau, x - a))q$$

para este caso de λ.

$\lambda < 0$**:** Procedemos de manera análoga, por (4.1) sabemos que

$$V(\tau, x) \geq pV(\tau, x + a) + qV(\tau, x - a)$$

debido a la condición de λ sucede que $(e^{\lambda} - 1) < 0$ así obtenemos

$$(e^{\lambda} - 1)V(\tau, x) \leq p(e^{\lambda} - 1)V(\tau, x + a) + q(e^{\lambda} - 1)V(\tau, x - a)$$

usando el hecho de que $p + q = 1$ se consigue

$$-\{(e^{\lambda}-1)V(\tau, x)+1\} \geq -\{((e^{\lambda}-1)V(\tau, x+a)+1)p+((e^{\lambda}-1)V(\tau, x-a)+1)q\}$$

es decir

$$-[(e^{\lambda}-1)V(\tau, x)+1] \geq -[((e^{\lambda}-1)V(\tau, x+a)+1)p+((e^{\lambda}-1)V(\tau, x-a)+1)q]$$

lo cual prueba que

$$U_{\lambda}(V_{\lambda}(\tau, x)) \geq U_{\lambda}(V_{\lambda}(\tau, x + a))p + U_{\lambda}(V_{\lambda}(\tau, x - a))q$$

para λ negativa.

□

De este resultado y de (5.2) se sigue inmediatamente que

$$U_{\lambda}(V_{\lambda}(\tau, x)) \geq U_{\lambda}(\mathcal{V}_{\lambda}(x)), \quad x \in X.$$

Tenemos finalmente el siguiente resultado.

Teorema 5.1. Para el λ-ejemplo si $p \geq q$, entonces la política τ es óptima.

Demostración. Esto se sigue de usar la observación (5.1) para τ y del Lema 5.1. □

CASO $p \leq q$

Sea α, la política audaz del capítulo anterior; la idea es la misma, calcular $U_\lambda(V_\lambda(\alpha, x))$ y obtener

$$U_\lambda(V_\lambda(\alpha, x)) = \text{sign}(\lambda)[(e^\lambda - 1)V(\alpha, x) + 1]. \tag{5.6}$$

Donde $V(\alpha, x)$ es la esperanza de la recompensa total ganada bajo α cuando el sistema comienza en x en el caso neutral al riesgo. Para el conjunto de estados $\{1, ..., N-1\}$ tenemos el siguiente.

Lema 5.2. $U_\lambda(V_\lambda(\alpha, x))$ cumple con la desigualdad (5.4) es decir,

$$U_\lambda(V_\lambda(\alpha, x)) \geq pU_\lambda(V_\lambda(\alpha, x + a)) + qU_\lambda(V_\alpha(\tau, x - a)),$$

$a \leq \min\{x, N - x\}$.

Demostración. Esta demostración es totalmente análoga al caso de τ. □

De este resultado y de la observación (5.1) se sigue que cuando $p \leq q$ podemos afirmar lo siguiente

$$U_\lambda(V_\lambda(\alpha, x)) \geq U_\lambda(\mathcal{V}_\lambda(x)), \quad x \in X$$

y obtenemos el teorema correspondiente.

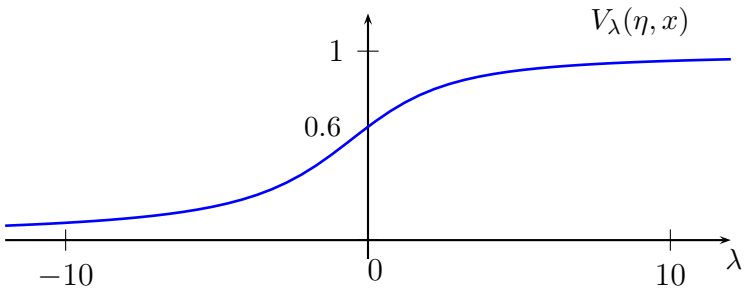
Teorema 5.2. Para el λ- ejemplo del problema de control óptimo la política α es óptima cuando $p \leq q$.

Demostración. Esto se sigue aplicar del Lema 5.1 para α y del Lema 5.2. □

Observación 5.3. Finalmente podemos afirmar lo siguiente. Las políticas que resultaron óptimas para el ejemplo neutral al riesgo también son óptimas en el caso sensible al riesgo. Una vez más es importante notar que esto es consecuencia de las condiciones impuestas en el modelo, que en el caso particular del sensible al riesgo, derivaron en la forma para la λ-función objetivo obtenida mediante la siguiente ecuación

$$U_\lambda(V_\lambda(\eta, x)) = \text{sign}(\lambda)[(e^\lambda - 1)V(\eta, x) + 1], \quad x \in X, \quad \eta \in \mathcal{P}.$$

Gracias a esta ecuación es fácil ver algunas propiedades de $V_\lambda(\eta, x), \quad x \in X, \quad \eta \in \mathcal{P}$. En particular podemos demostrar que ésta es una función creciente en λ, para cualquiera de las dos políticas: τ y α, (de hecho para cualquier política). También es posible verificar que cuando λ tiende a infinito $V_\lambda(\eta, x)$ tiende a 1, mientras que cuando λ tiende a menos infinito $V_\lambda(\eta, x)$ tiende a 0. Por otro lado el límite cuando λ tiende a cero es la función objetivo del neutral al riesgo, i.e. $V(\eta, x)$. La siguiente es una gráfica típica del comportamiento de $V_\lambda(\eta, x)$, con $V(\eta, x) = 0.6$



Conclusiones

Usar un parámetro λ y una *función de utilidad* para plantear los *problemas de control óptimo λ-sensibles al riesgo*, de manera que el caso $\lambda = 0$ resulte *neutral al riesgo*, ha sido la manera clásica de abordar el tema de los problemas de control óptimo sensibles al riesgo. Sin embargo en este trabajo se parte del lado opuesto: se conoce la solución del ejemplo a un problema de control óptimo con recompensa total que no ha sido abordado desde la perspectiva sensible al riesgo (y en ese sentido lo consideramos *neutral al riesgo*), y buscamos extenderlo a un problema de control óptimo λ-*sensible al riesgo* de tal manera que el caso $\lambda = 0$ se reduzca al ya conocido. Podríamos decir que nuestra primera pregunta al iniciar el trabajo de tesis fue: ¿es posible extender la solución conocida (la neutral) al caso sensible al riesgo? es decir ¿qué ocurre con las políticas que dan solución en el problema que no toma en cuenta el riesgo cuando se plantea el sensible al riesgo? ¿funcionarán de la misma manera? La respuesta a estas preguntas constituyen la principal aportación de este trabajo.

Antes de resolver la pregunta original fue necesario comprender a fondo la solución al ejemplo conocido. De esta manera resultó otra aportación de esta tesis que consistió en describir exhaustivamente esta solución conocida. En primer lugar resultó necesario clarificar una fuerte afirmación de Ross [16], en el sentido de que la esperanza de la recompensa total ganada es equivalente calcular la probabilidad de que el sistema alcance a un estado especial N (fijo), antes que al cero. Esta afirmación, aunque cierta incluso desde el punto de vista intuitivo, no es trivial y su demostración requiere algunos detalles que Ross no ofreció en su solución. Por otro lado, cuando hablamos de *la* solución al caso neutral al riesgo nos debemos referir al par de soluciones verificadas en [16] pues, para ser resuelto, el problema debió ser dividido en dos casos, a partir de ciertos parámetros p y q descritos en el modelo. El primero p, es la probabilidad con la que se avanza sobre el espacio de estados en la dinámica del sistema. Mientras que $q = 1 - p$ es la de retroceder.

Así, cuando $p \geq q$ se encuentra que la política llamada *tímida* τ, es óptima. Y respectivamente cuando $p \leq q$ es la política *audaz* α la óptima.

A manera de síntesis se presentan los cuadros finales. El primero (5.1), es básicamente un esquema de los dos problemas de control de Markov que se ejemplificaron en este trabajo y el resultado que les da solución respectivamente. Los cuadros subsecuentes contienen el modelo que generó los ejemplos que ilustran la teoría del cuadro (5.1); en cada caso se presenta la solución obtenida a través de: una versión apropiada de la desigualdad de optimalidad y de la representación que induce el modelo de la función objetivo. Cabe señalar que el caso $p \geq q$ (cuadro 5.2), permitió dar la forma explícita de la función de valores óptimos de la política solución, mientras que en el otro caso, aunque sabemos cuál es la política óptima explícita, la función valores óptimos no lo es. Esto último abre la posibilidad de establecer como problema abierto la búsqueda de las condiciones en el modelo que permitan obtener de forma explícita la función de valor óptimo para el caso $p \leq q$.

Como puede verse en los cuadros (5.2) y (5.3), la misma política da solución tanto al ejemplo neutral al riesgo como al sensible al riesgo. Es decir, hay una especie de herencia hacia el modelo sensible al riesgo, pues las políticas que funcionaron para el neutral, funcionan también para el sensible. Este resultado depende fuertemente de que el modelo tiene un único estado absorbente. Es interesante notar que las políticas mencionadas dan solución al caso extendido sin importar cuál sea valor de λ, es decir no hay incidencia del valor (ni siquiera del signo), de este parámetro en la solución al caso extendido.

En general se espera encontrar un comportamiento distinto de las soluciones cuando se utiliza una función de utilidad. Pero está claro que no ocurrió tal efecto en el caso estudiado en este trabajo. Esto es resultado del tipo de función de utilidad con el que se trabajó, pues las propiedades que tiene la función exponencial permitieron conectar ambas soluciones.

En este sentido, cabe notar que el hecho de que la función de utilidad tenga forma exponencial, es consecuencia de considerar constante al coeficiente de sensibilidad al riesgo (véase [8]). De tal manera que un problema abierto a este tema de tesis sería proponer una función de utilidad que no considere constante esta sensibilidad y que por lo tanto no tenga forma exponencial.

PROBLEMAS	
Recompensa Total Neutral al Riesgo	Recompensa Total Sensible al Riesgo
$V(\pi,x)=E_x^{\pi}\left[\sum_{t=0}^{\infty}R(X_t,A_t)\right]\ x\in X,\ \pi\in\mathcal{P}.$	$V_\lambda(\pi,x)=\frac{1}{\lambda}\ln\left(E_x^{\pi}\left[e^{\lambda\sum_{t=0}^{\infty}R(X_t,A_t)}\right]\right)\ x\in X,\ \pi\in\mathcal{P}.$
SOLUCIONES	
Desigualdad de Optimalidad (DO)	λ-Desigualdad de Optimalidad (λ-DO)
Sea $f\in\mathbb{F}$ tal que.	Sea $f\in\mathbb{F}$ tal que.
* $V(f,x)<\infty,\ x\in X,$ * $V(f,x)\geq R(x,a)+\sum_y V(f,y)q_{xy}(a),$ $x\in X,\ a\in A(x)$	* $V_\lambda(f,x)<\infty,\ x\in X,$ * $U_\lambda(V_\lambda(f,x))\geq e^{\lambda(R(x,a))}\left[\sum_y U_\lambda(V_\lambda(f,y))q_{xy}(a)\right],$ $x\in X,\ a\in A(x)$
entonces f es óptima y $\mathcal{V}(\cdot)=V(f,\cdot)$.	entonces f es óptima y $\mathcal{V}_\lambda(\cdot)=V_\lambda(f,\cdot)$.

Cuadro 5.1: Teoría

<table>
<tr><td colspan="2">EJEMPLOS ($p \geq q$)</td></tr>
<tr><td colspan="2">$\diamond\ X := \{0,1,2,...,N\} \quad \diamond\ A := \{0,1,2,...,[N/2]\} \quad \diamond\ A(x) = \{1,2,...,\min\{x, N-x\}\}$
$\diamond\ q_{x,x+a}(a) = p, \quad q_{x,x-a}(a) = q = 1-p, \quad q_{N,0}(a) = q_{0,0}(a) = 1, \quad \diamond\ R(x,a) = 0, \quad x \neq N, \quad R(N,a) = 1.$</td></tr>
<tr><td>Recompensa Total Neutral al Riesgo
$V(\pi, x) = \sum_{t=0}^{\infty} P_x^{\pi}\{X_t = N\}, x \in X, \pi \in \mathcal{P}$</td><td>Recompensa Total Sensible al Riesgo
$U_\lambda(V_\lambda(\pi, x)) = \text{sign}(\lambda)[(e^\lambda - 1)V(\pi, x) + 1], x \in X, \pi \in \mathcal{P}$</td></tr>
<tr><td colspan="2">SOLUCIONES</td></tr>
<tr><td>DO. $f \in \mathbb{F}$ es óptima si
$V(f,x) \geq pV(f, x+a) + qV(f, x-a)$
$x \in X,\ a \leq \min\{x, N-x\}.$</td><td>$\lambda$-DO. $f \in \mathbb{F}$ es óptima si
$U_\lambda(V_\lambda(f,x)) \geq pU_\lambda(V_\lambda(f, x+a)) + qU_\lambda(V_\lambda(f, x-a))$
$x \in X,\ a \leq \min\{x, N-x\}.$</td></tr>
<tr><td colspan="2">La política tímida τ es óptima en ambos ejemplos y,
$V(\tau,x) = \frac{1-\left(\frac{q}{p}\right)^x}{1-\left(\frac{q}{p}\right)^N} \quad \text{y} \quad U_\lambda(V_\lambda(\tau,x)) = \text{sign}(\lambda)\left[(e^\lambda - 1)\left[\frac{1-\left(\frac{q}{p}\right)^x}{1-\left(\frac{q}{p}\right)^N}\right] + 1\right], \quad p \neq \frac{1}{2}$
$V(\tau,x) = \frac{x}{N} \quad \text{y} \quad U_\lambda(V_\lambda(\tau,x)) = \text{sign}(\lambda)\left[(e^\lambda - 1)\left[\frac{x}{N}\right] + 1\right], \quad p \neq \frac{1}{2}$</td></tr>
</table>

Cuadro 5.2:

<table>
<tr><td colspan="2">EJEMPLOS ($p \leq q$)</td></tr>
<tr><td colspan="2">$\diamond\ X := \{0,1,2,...,N\} \quad \diamond\ A := \{0,1,2,...,[N/2]\} \quad \diamond\ A(x) = \{1,2,...,\min\{x, N-x\}\}$
$\diamond\ q_{x,x+a}(a) = p, \quad q_{x,x-a}(a) = q = 1-p, \quad q_{N,0}(a) = q_{0,0}(a) = 1, \quad \diamond\ R(x,a) = 0, \quad x \neq N, \quad R(N,a) = 1.$</td></tr>
<tr><td>Recompensa Total Neutral al Riesgo
$V(\pi, x) = \sum_{t=0}^{\infty} P_x^{\pi}\{X_t = N\}, x \in X, \pi \in \mathcal{P}$</td><td>Recompensa Total Sensible al Riesgo
$U_\lambda(V_\lambda(\pi, x)) = \text{sign}(\lambda)[(e^\lambda - 1)V(\pi, x) + 1], x \in X, \pi \in \mathcal{P}$</td></tr>
<tr><td colspan="2">SOLUCIONES</td></tr>
<tr><td>DO. $f \in \mathbb{F}$ es óptima si
$V(f,x) \geq pV(f, x+a) + qV(f, x-a)$
$x \in X,\ a \leq \min\{x, N-x\}.$</td><td>$\lambda$-DO. $f \in \mathbb{F}$ es óptima si
$U_\lambda(V_\lambda(f,x)) \geq pU_\lambda(V_\lambda(f, x+a)) + qU_\lambda(V_\lambda(f, x-a))$
$x \in X,\ a \leq \min\{x, N-x\}.$</td></tr>
<tr><td colspan="2">La política audaz α :
$\alpha(x) = x$, cuando $x \leq N/2$, $\alpha(x) = N - x$, cuando $x \geq N/2$
es óptima en ambos ejemplos.</td></tr>
</table>

Cuadro 5.3:

Apéndice A

Propiedades Básicas de Procesos de Control de Markov

Este apéndice está basado en literatura dedicada a los procesos de control de Markov [6],[9]. Y está dividido en dos partes, la primera contiene resultados básicos de esperanza condicional y la segunda las pruebas de las propiedades básicas de los procesos de control de Markov. Antes de la primera sección tenemos el siguiente resultado de teoría de la medida.

Teorema A.1. Sean $(\Omega, \mathfrak{F})$ y $(\Omega_0, \mathfrak{F}_0)$ espacios medibles, sea $T : (\Omega, \mathfrak{F}) \to (\Omega_0, \mathfrak{F}_0)$ un mapeo medible, y μ una medida de probabilidad sobre $\mathfrak{F}$. Definimos $\mu_0 = \mu T^{-1}$ sobre $\mathfrak{F}_0$ como

$$\mu_0(A) = \mu(T^{-1}(A)), \qquad A \in \mathfrak{F}_0.$$

Si $f : (\Omega_0, \mathfrak{F}_0) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ y $A \in \mathfrak{F}_0$, entonces

$$\int_{T^{-1}A} f(T(\omega))d\mu(\omega) = \int_A f(\omega)d\mu_0(\omega).$$

En el sentido de que si una de las integrales existe, entonces la otra también, y las dos integrales son iguales.

A.1. Resultados de Esperanza Condicional

Definición A.1. Sea $(\Omega, \mathcal{F}, \Theta)$ un espacio de probabilidad, $\mathcal{H}$ una sub-σ-álgebra de $\mathcal{F}$ y Z una v.a. $\mathcal{F}$-medible. Si Z es Θ-integrable entonces la *esperanza condi-*

cional de Z dada $\mathcal{H}$ denotada por $E(Z \mid \mathcal{H})$, es cualquier función W sobre Ω tal que

(a) W es $\mathcal{H}$-medible, y

(b) $\int_B W dQ = \int_B Z dQ$ para cada $B \in \mathcal{H}$.

Si V es un conjunto en $\mathcal{F}$, la esperanza condicional de V dada $\mathcal{H}$ se define como $\Theta(V \mid \mathcal{H}) := E(I_V \mid \mathcal{H})$.

Proposición A.1. Sean Z, Z' v. a. sobre $(\Omega, \mathcal{F}, Q)$ y $\mathcal{H}$ y $\mathcal{H}'$ sub σ-álgebras de $\mathcal{F}$, si $\mathcal{H} \subset \mathcal{H}'$ entonces

(a) $E[E(Z \mid \mathcal{H}) \mid \mathcal{H}'] = E[E(Z \mid \mathcal{H}') \mid \mathcal{H}] = E(Z \mid \mathcal{H})$.

(b) Si Z es $\mathcal{H}$- medible, entonces $E(ZZ' \mid \mathcal{H}) = ZE(Z' \mid \mathcal{H})$; en particular $E(Z \mid \mathcal{H}) = Z$.

A.2. Propiedades

Sea $\{(X, A, \{A(x) | x \in X\}, Q, \rho)\}$ un modelo de control con su estructura de políticas $\mathcal{P}$ (véase [9]). Donde X y A denotan los espacios de estados y de controles respectivamente, los cuales se suponen espacios de Borel (i.e. subconjuntos medibles de espacios métricos separables y completos), $A(x)$ los controles admisibles para el estado x, Q la ley de transición del modelo, y ρ la función de respuesta. Sea

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

el conjunto de pares estado-acción admisible.
Sea H_t el espacio de *historias admisibles* hasta el tiempo t con $H_0 := X$; y

$$H_t := \mathbb{K}^t \times X = \mathbb{K} \times H_{t-1} \quad \text{para} \quad t = 1, 2, ... \tag{A.1}$$

$h_t \in H_t$ tiene la forma $h_t = (\xi_0, \alpha_0, ..., \xi_{t-1}, \alpha_{t-1}, \xi_t)$ con $(\xi_i, \alpha_i) \in \mathbb{K}$ para todo $i = 0, ..., t-1$; y $\xi_t \in X$.

Dada una política $\pi \in \mathcal{P}$ (véase [9]), y una medida de probabilidad ν sobre X que llamaremos inicial. Construiremos el siguiente espacio de probabilidad: $\Omega_\infty =$

$(XA)^\infty = (XA)\cdot(XA)\cdot(XA)\cdots$ [nótese que $H_\infty = \mathbb{K}^\infty \subseteq \Omega$], la correspondiente σ-álgebra producto $\mathfrak{F}$ en Ω y tomamos P_ν^π la medida de probabilidad sobre $(\Omega, \mathfrak{F})$. Esta P_ν^π existe por el Teorema de Ionescu Tulcea (véase [2]) y coincide con las medidas marginales, es decir cada vez que consideramos ya sea $(XA)^t$ o bien $(XA)^tX$ para algún $t \in \mathbb{N}$ (cada vez que "recortamos" o paramos el proceso a un número finito). Para ver más claramente esto consideremos las siguientes elementos aleatorios.

$$X_t : \Omega_\infty \to X, \quad \text{dada por} \quad X_t(x_0, a_0, x_1, ...x_t, a_t, ...) = x_t, \quad \text{y}$$

$$A_t : \Omega_\infty \to A, \quad \text{dada por} \quad A_t(x_0, a_0, x_1, ...x_t, a_t, ...) = a_t.$$

Es decir para cada t, el proceso toma uno de los valores de X que indexaremos con el tiempo x_t, así pues estamos mirando las sucesiones de variables aleatorias de los estados $\{X_t\}$ y de las acciones $\{A_t\}$.

Aquí probaremos algunas de las propiedades que satisface la medida de probabilidad P_ν^π, sin embargo es necesario primero hacer una aclaración sobre la notación de las medidas, para aligerarla usaremos P en lugar de P_ν^π, aunque para no confundir la medida total sobre Ω_∞ con las medidas marginales, para $t \in \mathbb{N}$, usaremos P^{2t} para medir sobre $(XA)^t$ cuando detenemos el proceso en A_{t-1} y P^{2t+1} para medir sobre $(XA)^tX$ cuando lo detenemos en X_t y usaremos la siguiente notación

$$dP^{2t} =$$

$$\pi_{t-1}(da_{t-1} \mid h_{t-1})Q(dx_{t-1} \mid x_{t-2}, a_{t-2})\cdots Q(dx_1 \mid x_0, a_0)\pi_0(da_0 \mid x_0)\nu(dx_0) \quad \text{(A.2)}$$

y

$$dP^{2t+1} =$$

$$Q(dx_t \mid x_{t-1}, a_{t-1})\pi_{t-1}(da_{t-1} \mid h_{t-1})\cdots Q(dx_1 \mid x_0, a_0)\pi_0(da_0 \mid x_0)\nu(dx_0), \quad \text{(A.3)}$$

por supuesto $dP^0 = d\nu$.

NOTAS.

(1) Dada una v.a. W, $\sigma(W)$ representa a la sigma álgebra generada por ésta. De esta manera si usamos la Definición A.1 para los, recién definidos, Ω_∞, $\mathfrak{F}$ y P_ν^π y tomamos $\mathcal{H} = \sigma(X_0, ..., X_t, A_t)$ entonces, para un $B \in \sigma(X_0, ..., X_t, A_t)$ lo siguiente

$$\int_B I_V dP = \int_B E[I_V \mid \sigma(X_0, ..., X_t, A_t)]dP \tag{A.4}$$

se cumple para cualquier $V \in \mathfrak{F}$ y además

$$P(V \mid \sigma(X_0, ..., X_t, A_t)) := E(I_V \mid \sigma(X_0, ..., X_t, A_t))$$

(2) En el resto del apéndice tendremos $B \in \sigma(X)$ y $D \in \sigma(A)$

Propiedad A.1. $P_\nu^\pi(X_0 \in B) = \nu(B)$ c.s. con respecto a P.

Demostración. En efecto, basta calcular

$$P_\nu^\pi(X_0 \in B) = \int_\Omega I_{\{X_0 \in B\}}(\omega)dP_\nu^\pi = \int_X I_B(x_0)dP^0(x_0) = \int_B d\nu(x_0) = \nu(B).$$

La segunda desigualdad se debe al Teorema de Cambio de Variable A.1, para $\Omega = \Omega_\infty$, $\Omega_0 = X$, $\mu = P_x^\pi$, $\mu_0 = P^0$ y $T = X_0$

□

Propiedad A.2. $P_\nu^\pi(X_{t+1} \in B \mid H_t, A_t) = Q(B \mid X_t, A_t)$.

Demostración. Sea $C \in \sigma(X_0, A_0, ..., X_t, A_t)$, consideremos el conjunto $\{(X_0, A_0, ..., X_t, A_t) \in C\}$ entonces,

$$\int_{\{(X_0, A_0, ..., X_t, A_t) \in C\}} P_\nu^\pi(X_{t+1} \in B \mid H_t, A_t)dP_\nu^\pi \tag{A.5}$$

por (A.4) esta integral es igual a

$$\int\limits_{\{(X_0,A_0,...,X_t,A_t)\in C\}} I_{\{X_{t+1}\in B\}} dP_\nu^\pi$$

es decir,

$$\int\limits_{\Omega} I_{\{\{X_{t+1}\in B\}\cap\{(X_0,A_0,...,X_t,A_t)\in C\}\}} dP_\nu^\pi;$$

como $\{\{X_{t+1} \in B\} \cap \{(X_0, A_0, ..., X_t, A_t) \in C\}\}$ es la imagen inversa de un elemento de la $\sigma(X_0, A_0, ..., X_t, A_t, X_{t+1})$, usamos cuidadosamente el Teorema de Cambio de Variable con la proyección para medir en $(XA)^{t+1}X$ con $P^{2(t+1)+1}$ y obtenemos

$$\begin{aligned}
&= \int\limits_{(XA)^{(t+1)}} \int\limits_{X} I_C(x_0, ..., a_t) I_B(x_{t+1}) dP^{2(t+1)+1} \\
&= \int\limits_{(XA)^{(t+1)}} \int\limits_{X} I_C(x_0, ..., a_t) I_B(x_{t+1}) Q(dx_{t+1} \mid x_t, a_t) dP^{2(t+1)} \\
&= \int\limits_{C} Q(B \mid x_t, a_t) dP^{2(t+1)}.
\end{aligned}$$

Usamos nuevamente el Teorema de Cambio de Variable para regresar a la medida en Ω_∞, y así esta última expresión es igual a

$$\int\limits_{\{(X_0,A_0,...,X_t,A_t)\in C\}} Q(B \mid X_t, A_t) dP_\nu^\pi. \tag{A.6}$$

Como las integrales (A.5) y (A.6) son iguales, obtenemos finalmente la igualdad deseada.

$$P_\nu^\pi(X_{t+1} \in B \mid H_t, A_t) = Q(B \mid X_t, A_t)$$

□

Propiedad A.3. $P_\nu^\pi(A_0 \in D \mid X_0) = \pi_0(D \mid X_0)$ casi seguramente c.r. P.

Demostración. Sea $\overline{D} \in \sigma(X_0)$, consideremos el conjunto $\{X_0 \in \overline{D}\}$

$$\int\limits_{\{X_0 \in \overline{D}\}} P_\nu^\pi(A_0 \in D \mid X_0) dP_\nu^\pi = \int\limits_{\{X_0 \in \overline{D}\}} I_{\{A_0 \in D\}} dP_\nu^\pi = \int\limits_{\Omega} I_{\{\{A_0 \in D\} \cap \{X_0 \in \overline{D}\}\}} dP_\nu^\pi$$

al tomar la medida en XA vía el TCV esto es igual a

$$\int\limits_{XA} I_D(a_0) I_{\overline{D}}(x_0) dP^2 = \int\limits_{\overline{D}} \pi_0(D \mid x_0) dP^0(x_0) = \int\limits_{\{X_0 \in \overline{D}\}} \pi_0(D \mid X_0) dP_\nu^\pi$$

probando así que $P_\nu^\pi(A_0 \in D \mid X_0) = \pi_0(D \mid X_0)$. □

Propiedad A.4. $P_\nu^\pi(A_t \in D \mid H_t) = \pi_t(D \mid H_t)$ casi seguramente c.r. P.

Demostración. Sea $\overline{D} \in \sigma(X_0, ..., A_{t-1}, X_t)$ y consideremos los conjuntos $\{(X_0, ..., A_{t-1}, X_t) \in \overline{D}\}$ y $\{\{A_t \in D\} \cap \{(X_0, ..., A_{t-1}, X_t) \in \overline{D}\}\} = W$. Calculemos de sobre todo Ω_∞.

$$\int\limits_{\{(X_0,...,A_{t-1},X_t) \in \overline{D}\}} P_\nu^\pi(A_t \in D \mid H_t) dP_\nu^\pi = \int\limits_{\{(X_0,...,A_{t-1},X_t) \in \overline{D}\}} I_{\{A_t \in D\}} dP_\nu^\pi = \int\limits_{\Omega} I_{\{W\}} dP_\nu^\pi$$

usando el TCV con la proyección en los valores, pasamos a la medida marginal $dP^{2(t+1)}$ y esta integral la escribimos como sigue

$$\int\limits_{(XA)^{2(t+1)}} I_D I_{\overline{D}} dP^{2(t+1)} = \int\limits_{(XA)^{2(t)}X} \int\limits_{D} I_{\overline{D}}(x_0, a_0, ..., x_t) \pi_t(da_t \mid h_t) dP^{2t+1}$$

es decir

$$= \int\limits_{\overline{D}} \pi_t(D \mid h_t) dP^{2t+1} = \int\limits_{\{(X_0,...,A_{t-1},X_t) \in \overline{D}\}} \pi_t(D \mid H_t) dP_\nu^\pi$$

probando así que $P_\nu^\pi(A_t \in D \mid X_0) = \pi_t(D \mid H_t)$. □

Antes de probar la propiedad de Markov que cumple el proceso necesitamos algunas definiciones.

Definición A.2. Con Φ denotamos el conjunto de todos los kérneles estocásticos φ en $\mathfrak{P}(A \mid X)$ (el espacio de las medidas de probabilidad condicionales de X dado A) tales que $\varphi(A(x) \mid x) = 1$ para todo $x \in X$

Definición A.3. Sea Φ como en la definición anterior, $\mathbb{F}$ el conjunto de todas las funciones medibles de $X \to A$ tales que $f(x) \in A(x)$ para todo $x \in X$. Y ρ y Q como en la definición del MCM. Definimos para cada $x \in X$

$$\rho(x, \varphi) := \int_A \rho(x, a)\varphi(da \mid x)$$

y

$$Q(\cdot \mid x, \varphi) := \int_A Q(x, a)\varphi(da \mid x)$$

Más aún si f pertenece a $\mathbb{F}$ esto se convierte en

$$c(x, f) = \rho(x, f(x)) \quad \text{y} \quad Q(B \mid x, f) = Q(B \mid x, f(x)) \tag{A.7}$$

Propiedad A.5. $P^\pi_\nu(X_{t+1} \in B \mid H_t) = \int_A Q(B \mid X_t, a_t)\pi_t(da_t \mid H_t)$

Demostración. En efecto, como $\sigma(X_0, A_0, ..., X_t) \subset \sigma(X_0, ..., X_t, A_t)$, por la Proposición A.1 tenemos que

$$\begin{aligned} E^\pi_\nu[I_{\{X_{t+1}\in B\}} \mid H_t] &= E^\pi_\nu[E^\pi_\nu[I_{\{X_{t+1}\in B\}} \mid H_t, A_t] \mid H_t] \\ &= E^\pi_\nu[P^\pi_\nu(\{X_{t+1} \in B\} \mid H_t, A_t) \mid H_t] \\ &= E^\pi_\nu[Q(B \mid X_t, A_t) \mid H_t] \end{aligned} \tag{A.8}$$

la última igualdad se debe a la Propiedad (A.2). Ahora nótese que para todo $C \in \sigma(X_0, ..., X_t)$

$$\begin{aligned}\int_C E_\nu^\pi[Q(B \mid X_t, A_t) \mid H_t]dP_\nu^\pi &= \int_C Q(B \mid X_t, A_t)dP_\nu^\pi \\ &= \int_C \int_A Q(B \mid X_t, a_t)dP_\nu^\pi(A_t \in A \mid H_t)dP_\nu^\pi \\ &= \int_C \int_A Q(B \mid X_t, a_t)\pi_t(da_t \mid H_t)dP_\nu^\pi\end{aligned}$$

(la última igualdad se debe a la propiedad (A.4)). De aquí se sigue que

$$E_\nu^\pi[Q(B \mid X_t, A_t) \mid H_t] = \int_A Q(B \mid X_t, a_t)\pi_t(da_t \mid H_t)$$

o bien, por A.8

$$P_\nu^\pi[X_{t+1} \in B \mid H_t] = \int_A Q(B \mid X_t, a_t)\pi_t(da_t \mid H_t)$$

□

Observación A.1. A manera de corolario notemos lo siguiente. Si en particular se tiene una política estacionaria esto se reduce a

$$P_\nu^\pi(X_{t+1} \in B \mid H_t) = Q(B \mid X_t, f(X_t)). \tag{A.9}$$

Propiedad A.6. Si π es política de Markov entonces el proceso $\{X_t\}$ cumple con lo siguiente: para cada $B \in \mathfrak{B}(X)$ y $t = 0, 1, 2, ...,$

$$P_\nu^\pi(X_{t+1} \in B \mid X_0, ..., X_t) = P_\nu^\pi(X_{t+1} \in B \mid X_t). \tag{A.10}$$

Demostración. Se probará que ambos lados de la igualdad son iguales a $Q(B \mid X_t, f(X_t)$. Primero notemos que, por la proposición anterior

$$P_\nu^\pi(X_{t+1} \in B \mid H_t) = \int_A Q(B \mid X_t, A_t)\pi_t(da_t \mid H_t)$$

para cualquier π y para todo $B \in \mathfrak{B}(X)$ y por la proposición anterior se tiene que cuando es de Markov ocurre

$$P_\nu^\pi(X_{t+1} \in B \mid H_t) = Q(B \mid X_t, f(X_t)). \tag{A.11}$$

Ahora bien, como $\sigma(X_0, X_1, ...X_t) \subset \sigma(X_0, A_0, ...X_t)$, usamos la Proposición A.1(**a**) para el lado izquierdo de la igualdad A.10 obtenemos

$$P_\nu^\pi(X_{t+1} \in B \mid X_0, ..., X_t) = E_\nu^\pi[P_\nu^\pi(X_{t+1} \in B \mid H_t) \mid X_0, ..., X_t],$$

se sigue de (A.11) que

$$= E_\nu^\pi[Q(B \mid X_t, f(X_t)) \mid X_0, ..., X_t],$$

y de la Proposición A.1(**b**) se sigue que

$$= Q(B \mid X_t, f(X_t)).$$

Por otro, lado procediendo de manera similar para el lado derecho de la igualdad (A.10) obtenemos

$$P_\nu^\pi(X_{t+1} \in B \mid X_t) = E_\nu^\pi[P_\nu^\pi(X_{t+1} \in B \mid H_t) \mid X_t]$$

y por (A.11)

$$= E_\nu^\pi[Q(B \mid X_t, f(X_t)) \mid X_t]$$

usando propiedades de esperanza condicional

$$= Q(B \mid X_t, f(X_t)).$$

□

Apéndice B

Sensibilidad al Riesgo

Este apéndice está basado en los trabajos de Fishburn [7], Arrow [1], Pratt [14],[8].

La siguiente es una breve discusión acerca de la naturaleza y propiedades de lo que fue llamado *coeficiente de sensibilidad al riesgo*. Iniciarla requiere de contar con la definición de función de utilidad, de hecho es necesario ir más atrás; sin embargo aquí simplemente supondremos lo siguiente: hay un orden de preferencias sobre las alternativas que puede elegir un controlador, y este orden lo denotaremos por $\succcurlyeq$.

Definición B.1. Una función $u : W \to \mathbb{R}$ se dice que representa un *orden* de preferencia $\succcurlyeq$ definida sobre el conjunto de alternativas W si

$$\text{para todo} \quad x, y \in W, \qquad x \succcurlyeq y \quad \text{si y sólo si} \quad u(x) \geq u(y). \tag{B.1}$$

Es decir, u representa las preferencias del controlador si y sólo si, dadas dos alternativas, u le asigna un número real mayor a la alternativa que el controlador prefiere. A una función de utilidad sobre alternativas la llamaremos de tipo *elemental*.

Pareciera que esta función captura de manera fiel las preferencias de un controlador sin embargo hay algo que aún no está tomado en cuenta: la actitud del controlador ante el riesgo. Veamos un ejemplo de actitud ante el riesgo.

Un jugador debe elegir entre los juegos A y B, en el juego A gana 35 pesos con probabilidad $\frac{1}{2}$ o no obtiene nada con la misma probabilidad, en el juego B gana con probabilidad 1 (con certeza) 12 pesos.

	Probabilidad	Recompensa	R. Esperada
A	1/2 1/2	\$ 35.00 \$ 0.00	\$ 17.50
B	1	\$ 12.00	\$ 12.00

Cuadro B.1: Un ejemplo de actitud al riesgo

En este ejemplo tomamos como función de utilidad a la recompensa esperada. A los jugadores que eligen la opción B, ganar seguro a pesar de que la esperanza del juego A es mayor. Es decir los que eligen la que tiene menor recompensa esperada se les llama *aversos al riesgo*. Para estos jugadores parecería que la recompensa esperada como función de utilidad no está funcionando muy bien. Es aquí donde entran las ideas de von Neumann y Morgenstern: pensar en una función de utilidad sobre las probabilidades más que sobre la recompensa.

John von Neumann y Oscar Morgenstern probaron [17] que pueden representarse también preferencias entre distribuciones de probabilidad a través de funciones de utilidad a las que llamaron *funciones de utilidad esperada*. En este apéndice tendremos solamente la definición.

Definición B.2. Sea Δ el conjunto de las distribuciones de probabilidad definidas sobre un conjunto W. Existe una función $U : \Delta \to \mathbb{R}$ que modela un orden de preferencias ($\succeq$) en Δ es decir

$$\text{para todo} \quad P, Q \in \Delta, \qquad P \succeq Q \quad \text{si y sólo si} \quad U(P) \geq U(Q). \tag{B.2}$$

La actitud de los jugadores puede explicarse intuitivamente argumentando que tienen miedo de arriesgarse, que son aversos al riesgo. Esta idea no suena tan mal pero por otro lado no es una constante de comportamiento, pues las personas siguen comprando billetes de lotería sabiendo que el precio que pagan por ellos es mucho más grande que el valor esperado de la lotería. Es para tratar de entender estas evidencias, aparentemente contradictorias, que se estudian las actitudes al

riesgo. Las *funciones de utilidad esperada* de von Neumann y Morgenstern son un primer paso en este sentido pero sus alcances son mayores.

Aversión al riesgo y certeza equivalente

En [1] Arrow realiza una discusión acerca de las actitudes al riesgo que deriva en el concepto de lo que llamará *certeza equivalente*.

Supongamos que contamos con la función de utilidad u que modela las preferencias de un consumidor y que esta función tiene segunda derivada. Sea Y_0 la riqueza del consumidor y supongamos que se le ofrece la oportunidad de ganar o perder la cantidad h con igual probabilidad, $1/2$; o bien la alternativa de quedarse con su riqueza inicial Y_0. Esto es bastante parecido al primer ejemplo con la diferencia de que en este caso ambos juegos tienen la misma recompensa esperada, a saber Y_0.

	Probabilidad	Recompensa
A	$1/2$ $1/2$	$Y_0 + h$ $Y_0 - h$
B	1	Y_0

Cuadro B.2: Otro ejemplo de actitud ante el riesgo

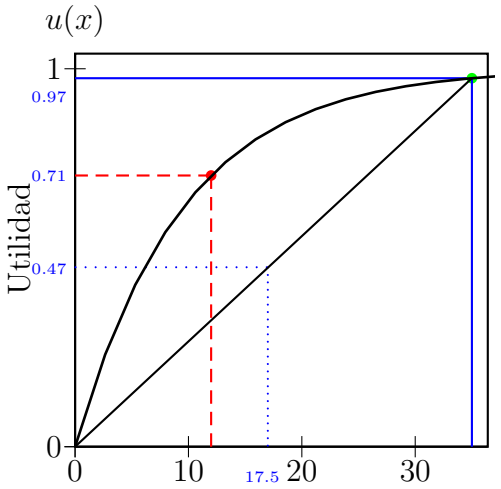
A pesar de este dato un consumidor que le teme al riesgo (es averso al riesgo), por definición preferirá quedarse con lo seguro, i.e. su función de utilidad debe cumplir con

$$u(Y_0) > \frac{1}{2}u(Y_0 + h) + \frac{1}{2}u(Y_0 - h) \tag{B.3}$$

a partir de aquí, con algunas cuentas de por medio, puede deducirse que u debe ser cóncava.

De tal modo que si proponemos una función de utilidad cóncava para el jugador de la primera tabla tendremos una idea más clara de por qué llaman aversos al riesgo a los jugadores que eligen la opción B. Nuestra propuesta de u es $1-0.9^x$ que

es claramente cóncava. Esta función asigna $u(0) = 0$, $u(12) = 0.71$ y $u(35) = 0.97$ es decir ordena las preferencias de manera racional. [1]



Pero por otro lado tenemos que la utilidad esperada del juego A es de 0.47 mientras que la esperada del juego B es otra vez 0.71. Así ya no resulta tan extraño que la mayoría (sic) de los jugadores elijan el juego B.

	Probabilidad	Recompensa	R. Esperada	Utilidad	U. Esperada
A	1/2 1/2	35.00 0.00	17.50	0.97 0.00	0.47
B	1	12.00	12.00	0.71	0.71

Cuadro B.3: Un ejemplo de actitud al riesgo

Ahora ya está más claro que cuando al controlador se le asocia una función de utilidad esperada cóncava, entonces será considerado averso al riesgo. De la figura también podemos mirar lo siguiente. Si la función de utilidad del controlador es una función afín creciente, entonces le dará lo mismo elegir el juego A o B. Simplemente porque en tal caso $u(12) < u(17.5)$. Es decir $E^A[u(R)] = u(E^A[R])$.

[1] racional en el sentido de que asigna un valor mayor a los objetos que más prefiere el jugador.

Regresando a la discusión sobre la concavidad, sabemos que el valor numérico de $u''(Y)$ nos da información en este sentido, sin embargo este valor no es suficiente para medir la aversión al riesgo de un consumidor. Pratt [14] introdujo una medida de la aversión al riesgo de la siguiente manera: pensemos en que el jugador está dispuesto a pagar una cantidad p con tal de evitar entrar en el juego que le parece riesgoso. Por supuesto esta cantidad debe ser *razonable*, ¿qué tan razonable? A esta cantidad p le llamó *premio de riesgo* del juego A. Que el jugador esté dispuesto a pagar para evitar el juego es equivalente a la siguiente condición

$$u(E^A(R) - p^A) = E^A[u(R)]. \tag{B.4}$$

Donde $E^A(R)$ es la esperanza de la recompensa ganada en el juego A. Para aligerar la notación pongamos $E^A(R) = \mathcal{R}^A$. Haciendo el desarrollo del lado izquierdo de (B.4)

$$u(\mathcal{R}^A - p^A) = u(\mathcal{R}^A) - p^A u'(\mathcal{R}^A) + o[p^A] \tag{B.5}$$

donde $o[\alpha]/\alpha \to 0$, cuando $\alpha \to 0$.

Mientras que del lado derecho obtenemos

$$E^A[u(R)] = u(\mathcal{R}^A) + \frac{1}{2}\sigma_A^2 u''(\mathcal{R}^A) + E^A[o(R - \mathcal{R}^A)^2] \tag{B.6}$$

combinando las tres últimas ecuaciones obtenemos

$$p^A u'(\mathcal{R}^A) = -\frac{1}{2}\sigma_A^2 u''(\mathcal{R}^A) + o[p^A] + E^A[o[(R - \mathcal{R}^A)]^2]. \tag{B.7}$$

De donde se deduce que p^A es proporcional hasta el primer orden (es decir, localmente) a la varianza de la recompensa, con factor de proporcionalidad igual a $1/2$. Finalmente llega a obtener una medida de aversión al riesgo que llama *coeficiente de aversión al riesgo*.

$$r(Y) = \frac{u''(Y)}{u'(Y)} = \frac{d}{dY} log(u'(Y)) \tag{B.8}$$

En general este cociente puede ser una función de Y. Esta clase de medida de aversión al riesgo "genera" funciones de utilidad que son conocidas como de *Aversión Absoluta al Riesgo*, ARA por sus siglas en inglés. Si además, como es el caso de este trabajo, el cociente es constante entonces se llama medida de tipo CARA (*Constant Absolute Risk-Aversion*). Existen otros tipos de medida de aversión al riesgo. En economía han surgido distintos tipos de funciones de utilidad que son de gran ayuda para deducir resultados analíticos, son utilizadas porque resultan de fácil manipulación pero no hay razón obvia para creer que representa actitudes de algún consumidor o controlador en el mundo real [8].

Nótese que si $r(\cdot) > 0$ entonces el controlador es averso al riesgo y será propenso al riesgo en el otro caso. Y que $r(Y) = 0$ implica que u es lineal, es decir el controlador es neutral al riesgo.

Así se ha obtenido una manera de medir al riesgo del controlador que en general podría ser una función de la riqueza que él tenga. Es posible verificar que cuando no depende la sensibilidad al riesgo de la riqueza actual sino que es una constante γ, ocurre que la función de utilidad del controlador debe ser de la siguiente forma

$$r(Y) = \gamma \Rightarrow u_\gamma(Y) := \begin{cases} \text{sign}(\gamma)e^{\gamma Y}, & \gamma \neq 0; \\ Y, & \gamma = 0. \end{cases} \tag{B.9}$$

Una última e importante consecuencia, derivada de la condición impuesta en (B.4), es el concepto de *certeza equivalente*. Este nombre no sorprende si recordamos que la condición impuesta tiene como propósito hallar la cantidad que estaría dispuesto a pagar el jugador con tal de no entrar en un juego que le parece riesgoso, es decir hallar una especie de balance entre los dos juegos. Así, llamaremos *certeza equivalente* ($\mathcal{Q}$) a lo que resulte en (B.4) para cada función de utilidad propuesta, es decir,

$$u(\mathcal{Q}(Y)) = E[u(Y)]. \tag{B.10}$$

En el caso de (B.9) se tiene como certeza equivalente a

$$\mathcal{Q}_\gamma(Y) := \begin{cases} \frac{1}{\gamma}\ln(E[e^{\gamma Y}]), & \gamma \neq 0; \\ E[Y], & \gamma = 0. \end{cases} \tag{B.11}$$

Bibliografía

[1] Arrow K, en: *Essays on the Bearing-Risk*, Markham, Chicago, 1971.

[2] Ash R., *Probability and Measure Theory*, Academic Press, India, 2008.

[3] Barz C. , *Risk-Averse Capacity Control in Revenue Management*, Springer, Berlin Heidelberg, 2007.

[4] Cavazos-Cadena R. and Fernández-Gaucherand E., *Controlled Markov chains with risk-sensitive criteria: average cost, optimality equations and optimal solutions*, Mathematical Methods of Operations Research (2000), no. 43, 121–139.

[5] 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 **2** (2000), no. 27, 167–185.

[6] Feinberg E. and Shwartz A., *Handbook of Markov Decision Processes: Methods and Applications*, Kluwer, 2001.

[7] Fishburn P., *Utility theory*, Management Science **14** (1968), no. 5, 335–378.

[8] Gollier C., *The Economics of Risk and Time*, MIT Press, 2002.

[9] Hernández-Lerma O. and Lasserre J. B., *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, 1988.

[10] Howard R. and Matheson J., *Risk-sensitive Markov decision processes*, Managment Science **18** (1972), no. 7, 356–369.

[11] Howard R., *Decision analysis: practice and promise*, Management Science **6** (1988), no. 34, 678–700.

[12] Liu Y., Ph.D. thesis, *Decision-Theoretic Planning under Risk-Sensitive Planning Objectives*, College of Computing, Georgia Institute of Technology, Atlanta, Abril 2005.

[13] Marcus S., Fernández-Gaucherand E., Hernández-Hernández D. , Coraluppi S., and Fard P., *Risk-sensitive Markov decision processes* en: *Systems and Control in the Twenty-First Century*, Eds. Byrnes Ch., Datta B., Gilliam D., and Martin C., Birkhäuser, 1997.

[14] Pratt J, *Risk aversion in the small and in the large*, Econometrica **32** (1964), 122–136.

[15] Puterman M., *Markov Decision Processes*, Wiley, New York, 1994.

[16] Ross S., *Introduction to Stochastic Dynamic Programming*, Academic Press, 1983.

[17] von Neumann J. and Morgenstern O. , *Theory of Games and Economic Behavior*, Springer-Verlag, 1944.

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Fecha : 12/08/2008
Página : 1/1

CONSTANCIA DE PRESENTACION DE EXAMEN DE GRADO

La Universidad Autónoma Metropolitana extiende la presente CONSTANCIA DE PRESENTACION DE EXAMEN DE GRADO de MAESTRA EN CIENCIAS (MATEMÁTICAS) de la alumna MARIA SOLEDAD ARRIAGA , matrícula 205181405, quien cumplió con los 132 créditos correspondientes a las unidades de enseñanza aprendizaje del plan de estudio. Con fecha veinte de agosto del 2008 presentó la DEFENSA de su EXAMEN DE GRADO cuya denominación es:

PROBLEMAS DE CONTROL DE MARKOV CON RECOMPENSA TOTAL ESPERADA EN ESPACIOS FINITOS. CASOS NEUTRAL Y SENSIBLE AL RIESGO

Cabe mencionar que la aprobación del Examen de Grado tiene un valor de 60 créditos y el programa consta de 192 créditos.

El jurado del examen ha tenido a bien otorgarle la calificación de:

Aprobar

JURADO

Presidente

DR. JUAN GONZALEZ HERNANDEZ

Secretario

DR. JOSE RAUL MONTES DE OCA MACHORRO

Vocal

DR. EVGUENI GORDIENKO

UNIDAD IZTAPALAPA
Coordinación de Sistemas Escolares
Av. San Rafael Atlixco 186 Col. Vicentina, Del. Iztapalapa CP 09340 México, DF Apodo. Postal 555-320-9000
Tels. 5804-4880 y 4883 Fax 5804-4876