BENEMÉRITA UNIVERSIDAD AUTÓNOMA
DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Nombre del proyecto: Procesos de Decisión de Markov con Recompensa Promedio: Caso Neutral y Sensible.

Director de Tesis:
Hugo Adán Crúz Suárez.

Alumna:
Alejandra Xochitl Hernández Dávila.

Grado por obtener: Maestría.

Índice general

Introducción

En este trabajo de tesis se abordan algunos conceptos relacionados a los Procesos de Decisión de Markov (PDM). Los PDMs son aplicados en problemas de control en ingeniería (véase [19]) y economía (véase [11]). A la estrategia seguida en cada periodo de tiempo se le llama política. Para evaluar la calidad de cada política se cuenta con un criterio de rendimiento o función objetivo, el problema de interés, llamado problema de control óptimo, consiste en hallar una política que optimice la función objetivo. Para resolver dicho problema es empleada la técnica de programación dinámica que permite determinar las decisiones que optimizan el comportamiento de un sistema que evoluciona a lo largo de una serie de etapas; la idea es encontrar la solución óptima de un problema descomponiéndolo en n subproblemas de una variable y resolviendo cada problema con el fin de hallar una solución global.

Se tratará el criterio de costo promedio empleado en problemas de redes de comunicación y teoría de colas (véase [9]). Este problema es resuelto usando el enfoque de descuento desvaneciente basado en el problema de costo descontado, se establecen condiciones para la existencia de una solución a la ecuación de programación dinámica que caracteriza a la ecuación de costo promedio óptima. La política que optimiza el criterio de rendimiento es conocida como política óptima y, al criterio de rendimiento evaluado en la política óptima como función de valor óptimo.

Los procesos de decisión de Markov se dividen en dos tipos: neutral y sensible al riesgo, para distinguir ambos tipos se emplea una constante λ conocida como coeficiente de sensibilidad al riesgo; si $\lambda \neq 0$ el proceso de control de Markov es sensible al riesgo (véase [1], [2] y [11]), mientras que, si

$\lambda = 0$ se trata del caso neutral al riesgo (véase [1] [5], [11], [14], [10]).

En este trabajo de tesis se aborda tanto el caso neutral como el caso sensible al riesgo para un proceso de decisión de Markov con criterio de rendimiento promedio (véase [11]). La herramienta básica para resolver este problema es programación dinámica, introducida por Richard E. Bellman (véase [4]). El principio de Programación Dinámica permite resolver problemas en los que se toman decisiones en etapas sucesivas que condicionan la evolución del sistema, afectando a las situaciones en las que el sistema se encontrará en el futuro, así como a las decisiones que serán tomadas.

En el caso neutral al riesgo es presentado el criterio de costo descontado y el criterio de costo promedio. Posteriormente, el caso de costo promedio es analizado usando una técnica de factor de descuento desvaneciente, la idea, es aproximar al caso promedio usando un problema de control óptimo descontado, para la validez de dicha aproximación se presentan condiciones necesarias para hallar una solución a los problemas. En el caso sensible al riesgo es tratado el problema de valor óptimo con el criterio de costo promedio para lo que se sigue la misma idea presentada con el factor de descuento desvaneciente ahora con espacio de estados y de acciones finitos. También es expuesto un ejemplo con dos estados y dos acciones, con lo que es posible observar que dada un costo (o recompensa) constante es posible hallar una solución al problema aunado a otras suposiciones adicionales, una de las más importantes es que la matriz de transición sea comunicante.

Los proceso de decisión de Markov sensibles al riesgo fueron estudiados por primera vez por Howard y Mathenson en 1972 (véase [13]) en dicho trabajo se considera el espacio de estados y de acciones finitos, si el proceso es comunicado, aperiódico y bajo la acción de cada política estacionaria, la ecuación de optimalidad tiene solución para un coeficiente sensible al riesgo positivo. En 1998 Cavazos-Cadena y Fernández-Gaucherand (véase [5]) demostraron que cuando el espacio de estados es finito la condición de Doeblin (véase [2], [10], [6]) garantiza la existencia de una solución de la ecuación de costo promedio óptima, mientras que Hernández-Hernández [10] se basa en la teoría de juegos estocásticos para resolver el problema de control óptimo para un coeficiente de sensibilidad al riesgo cercano a cero. Los trabajos mencionados parten de la idea de Von Neumann y Morgenstern (presentada en [20]) donde se muestra que las preferencias entre distribuciones de

probabilidad pueden modelarse, a través de una función de utilidad esperada.

La organización de esta tesis es la siguiente: en el Capítulo 1 se presenta el modelo de decisión de Markov así como conceptos básicos empleados, se describe el problema de control óptimo así como la técnica de programación dinámica empleada en la resolución de este problema; además se dan ejemplos en los que se emplea la teoría presentada en dicho capítulo, en particular se presenta un ejemplo de inventarios con horizonte finito. En el Capítulo 2 es resuelto el problema de control óptimo con el criterio de costo promedio y es aplicado al ejemplo de inventarios presentado en el Capítulo 1 ahora usando el enfoque de factor de descuento desvaneciente. En el Capítulo 3 se resuelve el problema de control óptimo de costo promedio sensible al riesgo presentando una vez más un ejemplo aplicado a la teoría de inventarios. Finalmente, se presenta un capítulo de conclusiones y un apéndice de resultados auxiliares utilizados en la tesis.

Objetivos

Establecer condiciones necesarias para garantizar la existencia de una solución del problema de control óptimo empleando la técnica de programación dinámica.

En el caso neutral al riesgo presentar un ejemplo de inventarios en donde se resuelve de problema de control óptimo con el criterio de rendimiento de costo descontado y de costo promedio.

Para el caso de un PDM sensible al riesgo establecer una condición para la existencia de una solución del problema de control óptimo con criterio de rendimiento de costo promedio para un coeficiente de sensibilidad al riesgo λ que no es "cercano" a cero.

Presentar un ejemplo aplicado a la teoría de inventarios en el caso sensible al riesgo ya que en este no existen muchos ejemplos en la bibliografía.

Capítulo 1

Preliminares.

Un Proceso de Decisión de Markov (PDM) modela un sistema dinámico cuyos estados son observados periódicamente por un controlador de forma discreta en el tiempo, el cual debe tomar una decisión, y como consecuencia se paga un costo (o se obtiene una recompensa). En este capítulo se darán conceptos fundamentales empleados en la teoría de los procesos de decisión de Markov (véase [11] y [14]).

1.1. Procesos de Decisión de Markov.

Definición 1.1.1. *Un modelo de decisión de Markov (MDM) es una quíntupla*

$$(X, A, \{A(x)|x \in X\}, Q, c).$$

Donde

- X *es un espacio de Borel llamado* espacio de estados*, los elementos* $x \in X$ *son llamados* estados*.*
- A *es un espacio de Borel llamado* espacio de acciones*, los elementos* $a \in A$ *son llamados* acciones o controles*.*
- $\{A(x)|x \in X\}$ *es una familia de subconjuntos medibles no vacíos de* $X \times A$*. El conjunto* $\mathbb{K}$ *denota las parejas estados-acciones admisibles definida por*

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

- *Q es un kernel estocástico definido en X dado $\mathbb{K}$ llamada ley de transición, es decir, para cada $(x, a) \in \mathbb{K}$, $Q(\cdot|x, a)$ es una medida de probabilidad en X y para cada $B \subset X$ medible, $Q(B|\cdot)$ es una función medible en $\mathbb{K}$, es decir, para cada $x \in X$ y $a \in A(x)$, Q representa la siguiente probabilidad condicional*

$$Q(B|x, a) = P(X_{t+1} \in B|X_t = x, A_t = a) \quad B \in \mathcal{B}(X),$$

  donde $\mathcal{B}(X)$ representa la σ-álgebra de Borel.

- *La función $c : \mathbb{K} \to \mathbb{R}$ representa la función de costo.*

Observación 1.1.1. *El calificativo de Markov en la definición anterior es usado pues la probabilidad de transición y la función de costo dependen del estado actual y de la acción seleccionada por el controlador en este estado (véase [14]).*

La dinámica que describe a este sistema estocástico es la siguiente: si al tiempo t se encuentra en el estado $x_t = x \in X$, y se aplica la acción $a_t = a \in A(x)$, entonces ocurren dos cosas: se paga un costo y el sistema se traslada a un nuevo estado x_{t+1} mediante la ley de transición $Q(\cdot|x, a)$, hecha la transición a un nuevo estado, es elegida una nueva acción y la dinámica descrita se repite (véase [14]).

De esta manera la *historia* del proceso está dada por

$$\mathbb{H}_0 := X; \quad \mathbb{H}_t := \mathbb{K}^t \times X, \quad t \geq 1$$

un elemento $h_t \in \mathbb{H}$ es de la forma $h_t = (x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$, el cual representa la historia hasta el tiempo t (véase [8]).

1.2. Políticas.

Una política o estrategia especifica la regla de decisión que será aplicada en un periodo de observación del proceso, la cual comúnmente es denotada por π.

Definición 1.2.1. *Una* política *es una sucesión* $\pi = \{\pi_t | t = 0, 1, \dots\}$ *de kerneles estocásticos* π_t *definidas sobre* A *dado* $\mathbb{H}$*, la cual satisface*

$$\pi_t(A(x_t)|h_t) = 1 \text{ para cada } h_t \in \mathbb{H} \text{ y } t = 0, 1, \dots$$

Se denotará por Π *al conjunto de todas las políticas (véase [12]).*

Una política $\pi \in \Pi$ se clasifica dependiendo de como es incorporada la información pasada y como son seleccionadas las acciones (véase [14]), Así la clasificación es la siguiente.

- *Markoviana determinista*: Para seleccionar la acción admisible el controlador sólo observa estado y la acción actual, además dicha acción es elegida con certeza.
- *Dependiente la historia determinista*: Si la elección de una acción admisible depende de la historia pasada del sistema, es representada por una sucesión de los estados previos y sus acciones, dado un estado, la acción es elegida con certeza.
- *Markoviana aleatorizada*: La acción admisible elegida es determinada por una distribución de probabilidad sobre el conjunto de acciones, además la elección sólo depende del estado actual.
- *Dependiente de la historia aleatorizada*: El controlador elige la acción admisible mediante la observación de historia y es por medio de una distribución de probabilidad.

1.3. Problemas con Horizonte Finito.

1.3.1. Criterio de Rendimiento.

Cada PDM está dotado de una función real, llamada criterio de rendimiento también llamada función objetivo, esta mide la calidad de cada política, a través de la sucesión de costos que genera. El problema de control consiste en minimizar el criterio de rendimiento.

$$J(\pi, x) = E_x^{\pi}\left[\sum_{t=0}^{N-1} c(x_t, a_t) + c_N(x_N)\right], \tag{1.3.1}$$

donde $c_N : X \to \mathbb{R}$ es una función conocida, la cual denota la función de costo terminal. Al entero positivo N se le conoce como horizonte del problema, que representa el número de etapas en el cual el sistema está operando este puede puede ser finito o infinito.

El criterio definido en (1.3.1) es conocido como como *costo total acumulado*, también se define a la *función de valor* como

$$J^*(x) := \inf_{\pi \in \Pi} J(\pi, x), \quad x \in X.$$

De esta manera, el problema de valor óptimo consiste en hallar una política π^* (conocida como política óptima) tal que

$$J(\pi^*, x) = J^*(x), \quad x \in X.$$

Una técnica para resolver este problema es empleando el Teorema de Programación Dinámica.

1.3.2. Programación Dinámica.

El Teorema de Programación Dinámica enunciado a continuación proporciona un algoritmo para hallar la función de valor V^* así como la política π^*.

Teorema 1.3.1. *Teorema de Programación Dinámica. Sean $V_0, V_1, \ldots, V_N$ funciones sobre X definidas para cada $x \in X$ como*

$$V_N(x) := c_N(x), \tag{1.3.2}$$

y para $t = N-1, N-2, \ldots, 0$

$$V_t(x) := \min_{a \in A(x)} \Big[c(x,a) + \int_X V_{t+1}(y) Q(dy|x,a) \Big]. \tag{1.3.3}$$

Suponga que cada una de estas funciones son medibles y que para cada $t = 0, 1, \ldots, N-1$ existe un selector $f_t \in \mathbb{F}$ tal que $f_t(x) \in A(x)$ y para todo $x \in X$ y $t = 0, \ldots, N-1$

$$V_t(x) = c(x, f_t) + \int V_{t+1}(y) Q(dy|x, f_t), \tag{1.3.4}$$

entonces la política $\pi^* = (f_0, \ldots, f_{N-1})$ *es óptima y la función de valor* V^* *es igual a* V_0*, es decir,*

$$V^*(x) = V_0(x) = V(\pi^*, x) \quad \textit{para toda } x \in X.$$

La demostración del Teorema 1.3.1 puede encontrarse en [12] y en [22].

La ecuación (1.3.3) es conocida como ecuación de programación dinámica (EPD). El Teorema 1.3.1 impone al modelo de control de Markov una suposición importante conocida como condición de selección medible.

Suposición 1.3.1. *Dado el modelo de control de Markov y una función medible* $u : X \to \mathbb{R}$ *tales que*

$$u^*(x) = \inf_{a \in A(x)} \left[c(x,a) + \int_X u(y) Q(dy|x,a) \right],$$

es medible y existe un selector $f \in \mathbb{F}$ *tal que la función alcanza su mínimo en* $f(x) \in A(x)$*, es decir*

$$u^*(x) = c(x, f(x)) + \int_X u(y) Q(dy|x, f(x)) \quad \forall x \in X.$$

En problemas de aplicación esta suposición puede ser verificada directamente, aunque es conveniente tener condiciones generales con las que se cumpla, gracias a los Teoremas de selección medible (véase [12]) es posible hallar una caracterización.

En algunos casos es necesario escribir la ecuación de programación dinámica en formas equivalentes.

1.3.2.1. Variantes de la Ecuación de Programación Dinámica.

En esta sección se presentan algunas de las variantes de la ecuación de programación dinámica más usadas.

Modelo de Ecuación en Diferencias Considere el modelo de ecuación en diferencias

$$x_{t+1} = F(x_t, a_t, \xi_t),$$

donde $\{\xi_t\}$ es una sucesión de variables aleatorias (v.a.) independientes e idénticamente distribuidas (i.i.d) de un espacio de Borel S independiente del estado inicial x_0 con distribución común μ y $F : X \times A \times S \to X$ es una función medible conocida. En este caso, la ley de transición Q puede ser escrita como

$$Q(B|x,a) = \int_S I_B[F(x,a,s)]\mu(ds),$$

para todo $B \in \mathcal{B}(X)$, I_B representa la función indicadora del conjunto B. Por el teorema de cambio de variable (véase [3]) se tiene que si v es una función medible sobre X entonces

$$\begin{aligned} E[v(x_{t+1})|x_t = x, a_t = a] &= \int_X v(y)Q(dy|x,a), \\ &= \int_S v[F(x,a,s)]\mu(ds), \\ &= E[v(F(x,a,\xi_0))], \end{aligned}$$

así las ecuaciones (1.3.2) y (1.3.3) pueden ser reescritas como

$$V_N(x) = c_N(x), \tag{1.3.5}$$

$$\begin{aligned} V_t(x) &= \min_{a\in A(x)} \left\{ c(x,a) + \int_S V_{t+1}[F(x,a,s)]\mu(ds) \right\}, \\ &= \min_{a\in A(x)} \left\{ c(x,a) + E[V_{t+1}(F(x,a,\xi_t))] \right\}, \end{aligned} \tag{1.3.6}$$

para toda $x \in X$ y $t = N-1, N-2, \ldots, 1, 0$.

Forma hacia adelante de la EPD. Sean V_t las funciones (1.3.2)-(1.3.3) y defínase $v_t := V_{N-t}$ con $t = 0, \ldots, N$, entonces, reescribiendo la EPD en la forma hacia adelante

$$v_0(x) = c_N(x), \tag{1.3.7}$$

$$v_t(x) = \min_{a\in A(x)} \left\{ c(x,a) + \int_X v_{t-1}(y)Q(dy|x,a) \right\}, \tag{1.3.8}$$

si $t = 1, \ldots, N$. Más aún si $f_t \in \mathbb{F}$ es como en la ecuación (1.3.4), entonces $g_t := f_{N-t}$ $(t = 1, \ldots, N)$ es un minimizador de (1.3.8), en términos de las funciones v_t la conclusión del teorema de programación dinámica puede ser reescrita como $\tilde{\pi} = \{g_{N-1}, g_{N-2}, \ldots, g_1\}$ es una política óptima y la función de valor es $V^*(\cdot) = v_N(\cdot) = V(\tilde{\pi}, \cdot)$ así

$$v_N(x) = \inf_{\pi \in \Pi} V(\pi, x), \tag{1.3.9}$$

para todo $x \in X$, las funciones v_t en (1.3.7) y (1.3.8) son llamadas funciones de *iteración de valores*.

Costo Descontado. Suponga que en lugar del criterio de costo acumulado dado en (1.3.1), el costo esperado es de la forma

$$V^{\alpha}(x) := E_x^{\pi}\left[\sum_{t=0}^{N-1} \alpha^t c(x_t, a_t) + \alpha^N c_N(x_N)\right], \tag{1.3.10}$$

donde $0 < \alpha < 1$ es un número dado llamado *factor de descuento*. El modelo de control puede ser visto como un modelo no estacionario con X, A y Q fijos y un costo que varía por etapa, $c_t(x, a) := \alpha^t c(x, a)$. Entonces, de (1.3.2) y (1.3.10), se obtiene la ecuación de programación dinámica

$$V_N^{\alpha}(x) = \alpha^N c_N(x)$$

y para $t = N-1, N-2, \ldots, 0$

$$V_t^{\alpha}(x) = \min_{a \in A(x)}\left[\alpha^t c(x, a) + \int_X V_{t+1}^{\alpha} Q(dy|x, a)\right], \quad x \in X.$$

Reescribiendo esta ecuación en términos de las funciones $J_t^{\alpha}(\cdot) := \alpha^{-1} V_t^{\alpha}(\cdot)$, $t = 0, \ldots, N$ para obtener

$$J_N^{\alpha}(x) = c_N(x)$$

$$J_t^{\alpha}(x) = \min_{a \in A(x)}\left[c(x, a) + \alpha \int_X J_{t+1}^{\alpha} Q(dy|x, a)\right]. \tag{1.3.11}$$

El teorema de programación dinámica sigue siendo válido cuando las funciones V_t^{α} son remplazadas por J_t^{α}; esto es, la política $\pi^* = \{f_0, \ldots, f_{N-1}\}$ con $f_t \in \mathbb{F}$ un minimizador de (1.3.11) es óptima para el criterio de costo descontado dado y $V^{\alpha}(\pi^*, x) = J_0^{\alpha}(x)$.

Maximización de Recompensas En lugar de un costo por etapa, considere una *recompensa* r, entonces, el criterio de costo total acumulado se convierte en el criterio de recompensa total esperada definido como

$$J(\pi, x) := E_x^{\pi}\left[\sum_{t=0}^{N-1} r(x_t, a_t) + r_N(x_N)\right],$$

donde r_N es una función de recompensa terminal dada. En este caso, el problema de control consiste en hallar una política π^* que maximice el criterio $J(\pi, x)$, así que la función de valor $J^*(x)$ es

$$J^*(x) := \sup_{\pi \in \Pi} J(\pi, x), \quad x \in X.$$

Y haciendo los cambios necesarios para que las suposiciones dadas en el caso de costos sean válidas, se tiene que las ecuación de programación dinámica están dadas por

$$\begin{aligned} J_N(x) &= r_N(x) \\ J_t(x) &= \max_{a \in A(x)} \Big\{ r(x, a) + \int_X J_{t+1} Q(dy|x, a) \Big\}, \quad t = N-1, \ldots, 0. \end{aligned}$$

Algunos criterios de rendimiento pueden ser resueltos cuando el horizonte del problema es infinito, en particular el criterio de costo descontado.

1.4. Problemas con Horizonte Infinito.

En esta Sección será descrito el problema de costo descontado con horizonte infinito, el cual es resuelto empleando el método de aproximaciones sucesivas. Para aplicar dicho método es necesario que se cumplan las Suposiciones 1.4.1 y 1.4.2 enunciadas posteriormente

1.4.1. Costo Total Descontado.

El problema consiste en minimizar la esperanza total del costo descontado con horizonte infinito.

Dado un modelo de control de Markov $(X, A, \{A(x)|x \in X\}, Q, c)$ el criterio a minimizar es

$$V^{\alpha}(\pi, x) := E_x^{\pi}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right] \quad \pi \in \Pi, x \in X,$$

con $\alpha \in (0, 1)$ un factor de descuento dado. Una política π^* que satisface

$$V^{\alpha}(\pi^*, x) = \inf_{\pi \in \Pi} V^{\alpha}(\pi, x) = V_*^{\alpha},$$

para todo $x \in X$ se le llama política óptima (o α-descontada óptima) y V_*^{α} es llamada función de valor óptimo (o función de valor α-descontada óptima).

Bajo el supuesto de que la función de costo es no negativa se define el *n-ésima* costo descontado

$$V_n^{\alpha}(\pi, x) = E_x^{\pi}\left[\sum_{t=0}^{n-1} \alpha^t c(x_t, a_t)\right].$$

Por el Teorema de convergencia monótona (véase [3]) se tiene

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

Se dice que una función medible $v : X \to \mathbb{R}$ es solución de la ecuación de programación dinámica si para cada estado x se satisface

$$v(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int_X v(y) Q(dy|x, a)\right], \tag{1.4.1}$$

la función de valor óptimo V_*^{α} es una solución de la ecuación (1.4.1), es decir,

$$V_*^{\alpha}(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int_X V_*^{\alpha}(y) Q(dy|x, a)\right] \quad \forall x \in X, \tag{1.4.2}$$

para mostrar esta afirmación son empleadas las funciones de iteración de valores (véase [12]) definida como

$$v_n(x) := \min_{a \in A(x)} \left[c(x, a) + \alpha \int_X v_{n-1}(y) Q(dy|x, a)\right], \tag{1.4.3}$$

para toda $x \in X$ y $n = 1, 2, \ldots$ con $v_0 = 0$, v_n es la función de valor óptimo del *n-ésimo* costo descontado, esto es

$$v_n(x) = \inf_{\pi \in \Pi} V_n(\pi, x),$$

pues

$$V_*^\alpha(x) = \lim_{n \to \infty} v_n(x).$$

Haciendo a n tender a infinito en (1.4.3) se obtiene (1.4.2) si se cumple el intercambio de límite con el mínimo. Este procedimiento es conocido como método de aproximaciones sucesivas, para poder aplicarlo es necesario que se cumplan las siguientes suposiciones:

Suposición 1.4.1. *(a) La función de costo c es semicontinua inferiormente (l.s.c), no negativa e inf-compacta en* $\mathbb{K}$.

(b) Q es fuertemente continua.

Suposición 1.4.2. *Existe una política* $\pi \in \Pi$ *tal que* $V^\alpha(\pi, x) < \infty$ *para cada estado* $x \in X$.

Bajo estas condiciones también se cumple el siguiente teorema.

Teorema 1.4.1. *Bajo las Suposiciones 1.4.1 y 1.4.2*

1. *La función de valor óptimo* V_*^α *es la solución de la EPD para el costo descontado, esto es, para cada* $x \in X$

$$V_*^\alpha(x) = \min_{a \in A(x)} \left\{ c(x, a) + \alpha \int V_*^\alpha(y) Q(dy|x, a) \right\}, \tag{1.4.4}$$

   si u es otra solución de la EPD entonces $u \geq V_*^\alpha$.

2. *Existe un selector* f^* *tal que* $f^*(x) \in A(x)$ *con el que se alcanza el mínimo de la EPD para el costo descontado, es decir, para cada* $x \in X$

$$V_*^\alpha(x) = r(x, f^*) + \alpha \int V_*^\alpha(y) Q(dy|x, f^*). \tag{1.4.5}$$

3. *Si π^* es una política tal que $V^\alpha(\pi^*, \cdot)$ es una solución de la EPD para el costo descontado y satisface que para cada $x \in X$*

$$\lim_{n\to\infty} \alpha^n E_x^\pi \left[V^\alpha(\pi^*, x_n) \right] = 0, \tag{1.4.6}$$

entonces $V^\alpha(\pi^*, \cdot) = V_*^\alpha(\cdot)$ por lo tanto π^* es óptima.

4. *Si existe una política óptima π^* entonces existe una política que es determinista estacionaria.*

Para la demostración del teorema véase [11], [12] y [22].

En la siguiente sección se presentarán dos ejemplos de aplicación.

1.5. Ejemplos.

En esta Sección se presentan dos ejemplos importantes, el primero es usado en ingeniería y economía, conocido como lineal cuadrático o LQ (véase [11]), el segundo es aplicado a teoría de inventarios (véase [14]).

1.5.1. Problema Lineal Cuadrático (LQ).

El problema LQ o lineal cuadrático consisten de un sistema lineal con un costo cuadrático. Considere un sistema definido por la siguiente ecuación en diferencias

$$x_{t+1} = \gamma x_t + \beta a_t + \xi_t,$$

para toda $t = 0, 1, \ldots$ donde γ y β son constantes conocidas, la función de costo está dada por

$$c(x, a) = qx^2 + ra^2,$$

donde q y r son constantes reales tales que $q \geq 0$ y $r > 0$ y $\{\xi_t\}$ es una sucesión de v.a. i.i.d. tomando valores en $\mathbb{R}$ con función de densidad continua, independiente del estado inicial x_0 con media cero y varianza σ^2 finita.

Sea $X = A = A(x) = \mathbb{R}$, se busca una política que minimice el criterio de rendimiento

$$J(\pi, x) = E_x^{\pi}\left[\sum_{t=0}^{N-1}(qx_t^2 + ra_t^2) + q_N x_N^2\right],$$

donde $q_N \geq 0$.

Observe que las Suposiciones 1.4.1 y 1.4.2 se cumplen (véase [22]) así que es posible aplicar el algoritmo de programación dinámica. De las ecuaciones (1.3.5) y (1.3.6) se tiene que

$$\begin{aligned} J_N(x) &= q_N x^2, \\ J_t(x) &= \min_A \left\{ qx^2 + ra^2 + E[V_{t+1}(\gamma x_t + \beta a_t + \xi_t)] \right\}, \end{aligned}$$

para todo $x \in \mathbb{R}$ y $t = N-1, \ldots, 0$, para obtener V_{N-1} se toma $t = N-1$ en las ecuaciones anteriores. En general por inducción hacia atrás es posible demostrar que la política $\pi^* = \{f_0, \ldots, f_{N-1}\}$ está dada por

$$f_t(x) = G_t x \quad \text{con} \quad G_t = -(r + K_{t+1}\beta^2)^{-1} K_{t+1}\gamma\beta.$$

K_t se calcula de forma recursiva con $K_N = q_N$, es dada por

$$K_t = \left(1 - \frac{r + K_{t+1}\beta^2}{K_{t+1}\beta^2}\right) K_{t+1}\gamma^2 + q.$$

Por lo tanto la función de valor está dada por

$$J^*(x) = K_0 x^2 + \sigma^2 \sum_{n=1}^{N} K_n.$$

1.5.2. Inventarios.

Cada mes el encargado de un almacén lleva a cabo el inventario de cierto producto, basado en esta información decide ordenar o no más producto, con el que se enfrenta a un costo asociado con almacenar el producto o las ganancias perdidas por no ser capaz de cubrir la demanda del cliente. El objetivo del administrador es maximizar el beneficio obtenido. La demanda del producto es aleatoria con distribución de probabilidad conocida. El modelo obedece las siguientes condiciones:

1. La decisión de ordenar pedido adicional es hecha al inicio del periodo y se entrega inmediatamente.

2. La demanda del producto se recibe durante todo el periodo pero todas las ordenes son cumplidas al final del mes.

3. Si la demanda excede el inventario el cliente es enviado a otra parte por el producto faltante, es decir, no hay pedidos pendientes.

4. Los ingresos, costos, y la distribución de la demanda no varía con el periodo.

5. El producto es vendido únicamente en unidades enteras.

6. El almacén tiene capacidad de M unidades.

Se denotará por x_t la cantidad de inventario al inicio del mes t, como a_t al número de unidades ordenadas en el mes t y, por D_t a la demanda aleatoria en este mes. Suponga que la probabilidad de la demanda está dada por $p_y = P(D_t = y)$, con $y = 0, 1, 2, \ldots$. El inventario en la época de decisión $t+1$, x_{t+1} está relacionado con el inventario en el periodo t, x_t a través del sistema de ecuaciones

$$\begin{aligned} x_{t+1} &= \text{máx}\{x_t + a_t - D_t, 0\} \\ &\equiv [x_t + a_t - D_t]^+. \end{aligned}$$

Ya que no son permitidos pedidos pendientes el nivel del inventario no puede ser negativo, así que si $x_t + a_t - D_t < 0$, el nivel del inventario en el siguiente periodo de decisión será 0.

El costo por ordenar a unidades en cualquier periodo es $L(a)$ y, se supondrá que está compuesto por un costo fijo $K > 0$ por realizar el pedido y un costo variable $c(a)$ que se incrementa con la cantidad ordenada, es decir,

$$L(a) = \begin{cases} K + c(a) & , \text{ si } a > 0, \\ 0 & , \text{ si } a = 0. \end{cases}$$

Sea h una función no decreciente que representa el valor del costo por mantener un inventario de x unidades en un mes. Si la demanda es de j unidades y el inventario es suficiente para cubrirla entonces, el administrador

recibe una ganancia de $f(j)$, se supondrá que $f(0) = 0$.

En este modelo el costo depende del estado del sistema y de la siguiente época de decisión; es decir,

$$c(x_t, a_t, x_{t+1}) = L(a_t) + h(x_t + a_t) - f([x_t + a_t - x_{t+1}]^+)$$

Aunque es más conveniente trabajar con $c(x_t, a_t)$. Así que se calculará $F_t(x)$ que es el valor esperado de los ingresos recibidos en el periodo t cuando el inventario previo al ingreso del pedido del cliente es de x unidades.

- si el inventario x excede a la demanda j, el valor del ingreso es $f(j)$ que ocurre con probabilidad p.
- si la demanda excede al inventario, el valor del ingreso es $f(x)$ que ocurre con probabilidad $q_x = \sum_{j=x}^{\infty} p_j$.

Entonces

$$F(x) = \sum_{j=x}^{x-1} f(j)p_j + f(x)q_x$$

La formulación del modelo de control de Markov es la siguiente

- Épocas (o periodos) de decisión: $T = \{1, 2, \ldots, N\}$ para $N \leq \infty$.
- Estados: (cantidad de inventario al inicio del periodo t) $X = \{0, 1, 2, \ldots, M\}$.
- Acciones: (cantidad de producto que puede ser ordenada) $A = \{0, 1, 2, \ldots, M\}$.
- Acciones admisibles: (la cantidad de producto ordenado en el periodo t) $A(x) = \{0, 1, 2, \ldots, M - x\}$.
- Costo esperado: (costos de pedido y mantenimiento menos el ingreso esperado)

$$\begin{aligned} c(x, a) &= L(a) + h(x + a) - F(x + a) \\ c_N(x) &= g(x) \quad t = N. \end{aligned}$$

- Probabilidades de transición:

$$p_t(j|s, a) = \begin{cases} 0 & , \text{ si } M \geq j > x + a, \\ p_{x+a-j} & , \text{ si } M \geq x + a \geq j > 0, \\ q_{x+a} & , \text{ si } M \geq x + a \text{ y } j = 0. \end{cases}$$

1.5.2.1. Ejemplo numérico: Inventarios.

Ejemplo 1.5.1. *Suponga que se tiene los siguientes valores:*
$K = 5$, $c(a) = 10a$, $g(a) = 0$, $h(x) = 9x$, $M = 3$, $N = 4$.

La capacidad del inventario es de tres (o menos) unidades, la distribución de la demanda está dado por

$$p_y = \begin{cases} \frac{1}{3} & , \quad si\ y = 1, 2, \\ \\ \frac{1}{6} & , \quad si\ y = 0, 3, \end{cases}$$

El ingreso esperado cuando se tienen x unidades almacenadas antes de recibir una orden está dado por

x	$F(x)$
0	*0*
1	$\frac{25}{3}$
2	$\frac{40}{3}$
3	*15*

ahora, calculando el costo $c(x, a) = L(a) + h(x + a) - F(x + a)$

x/a	$c_t(x, a)$			
	0	*1*	*2*	*3*
0	*0*	*47/3*	*89/3*	*47*
1	*2/3*	*32/3*	*32*	$\times$
2	*14/3*	*27*	$\times$	$\times$
3	*12*	$\times$	$\times$	$\times$

Las $\times$ representan las acciones no admisibles para el estado x.

La matriz de transición es

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 5/6 & 1/6 & 0 & 0 \\ 1/2 & 1/3 & 1/6 & 0 \\ 1/6 & 1/3 & 1/3 & 1/6 \end{bmatrix}$$

Como el espacio de estados es finito y dada la matriz de recompensas se satisface la Suposición 1.4.1 trivialmente. Por otro lado la Suposición 1.4.2 se cumple pues, tomando $f(x) = 0$ para cada estado se tiene

$$\begin{aligned} V^{\alpha}(f,x) &= E_x^{\pi}\left[\sum_{t=0}^{\infty}\alpha^t c(x_t,a_t)\right] \\ &= E_x^{\pi}\left[\sum_{t=0}^{\infty}\alpha^t\right] \\ &= E_x^{\pi}\left[\frac{1}{1-\alpha}\right] \\ &= \frac{1}{1-\alpha} < \infty. \end{aligned}$$

Por lo tanto es posible aplicar las ecuaciones de programación dinámica (1.3.3) y (1.3.2) con las que se obtiene

$$\begin{aligned} V_4(x) &= 0, \\ V_3(x) &= \min_{a\in A(x)}\left\{c(x,a)+\sum_{y\in X}Q(y|x,a)V_4(y)\right\}, \end{aligned}$$

calculando $V_3(x)$ para cada estado

$$\begin{aligned} V_3(0) &= 0, \\ V_3(1) &= 2/3, \\ V_3(2) &= 14/3, \\ V_3(3) &= 12. \end{aligned}$$

Análogamente para $V_2(x)$

$$\begin{aligned} V_2(x) &= \min_{a\in A(x)}\left\{c(x,a)+\sum_{y\in X}Q(y|x,a)V_3(y)\right\}, \\ V_2(0) &= 0, \\ V_2(1) &= 14/18, \\ V_2(2) &= 102/18, \\ V_2(3) &= 284/18. \end{aligned}$$

Finalmente para $t = 1$

$$V_1(x) = \min_{a \in A(x)} \left\{ c(x,a) + \sum_{y \in X} Q(y|x,a) V_2(y) \right\},$$

$$\begin{aligned} V_1(0) &= 0, \\ V_1(1) &= 86/108 = 0.796296296, \\ V_1(2) &= 636/108 = 5.888888889, \\ V_1(3) &= 1808/108 = 16.740740741. \end{aligned}$$

Otro problema importante es el de recompensa promedio a largo plazo el cual se presenta en el siguiente capítulo.

Capítulo 2

Costo Promedio.

2.1. Planteamiento del Problema

En este capítulo se estudiará el problema de costo promedio a largo plazo que se describe a continuación.

Sea $(X, A, \{A(x)|x \in X\}, Q, c)$ un proceso de decisión de Markov y

$$V_n(\pi, x) = E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right],$$

el costo esperado en el *n-ésimo* periodo cuando es usada la política π y es dado el estado inicial $x_0 = x$. El costo promedio esperado usando $\pi \in \Pi$ y cuando $x_0 = x$ está dado por

$$V(\pi, x) := \limsup_{n\to\infty} \frac{V_n(\pi, x)}{n}. \tag{2.1.1}$$

El problema es hallar una política π^* tal que para cada estado x

$$V(\pi^*, x) = \inf_{\pi\in\Pi} V(\pi, x) = V^*(x), \tag{2.1.2}$$

a la política π^* se le conoce como *óptima* y la función $V^*(\cdot)$ es conocida como *función de valor*.

Para hallar una solución a este problema es necesario que la Suposición 1.4.1 se cumpla es decir

a) La función de costo es semicontinua inferiormente, no negativa e inf-compacta sobre $\mathbb{K}$.

b) Q *es fuertemente continua.*

2.2. Tripleta Canónica.

Sea $h : X \to \mathbb{R}$ una función medible conocida, $V_n(\pi, x, h)$ el costo total esperado en el $n-$*ésimo* periodo con costo terminal h cuando es usada la política π y es dado el estado inicial $x_0 = x$ y $V_0(\pi, x, h) := h(x)$, y para $n \geq 1$

$$V_n(\pi, x, h) := E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t, a_t) + h(x_n)\right] = V_n(\pi, x) + E_x^{\pi}[h(x_n)]. \quad (2.2.1)$$

La función de valor está dada por

$$\begin{aligned} V_n^*(x, h) &:= \inf_{\pi \in \Pi} V_n(\pi, x, h), \quad (2.2.2) \\ V_n^*(x, h) &:= \inf_{\pi \in \Pi} V_n(\pi, x) \quad \text{si } h(\cdot) \equiv 0. \end{aligned}$$

Definición 2.2.1. *Sean* ρ *y* h *funciones medibles con valores en los reales definida en* X *y* $f \in \mathbb{F}$ *un selector dado, se dice que* (ρ, h, f) *es una tripleta canónica si para cualquier* $x \in X$ *y* $n = 0, 1, \ldots$ *satisface*

$$V_n(\hat{f}, x, h) = V_n^*(x, h) = n\rho(x) + h(x), \quad (2.2.3)$$

donde $\hat{f}$ *es una política determinista estacionaria.*

Observación 2.2.1. *En la primera igualdad de (2.2.3),* $\hat{f}$ *es una política óptima para el* n-ésimo *periodo con costo terminal* h*, de la segunda igualdad se tiene que la* n-ésima *función de valor es* $n\rho(x) + h(x)$*, para,* $n = 0, 1, \ldots$.

En general, no es sencillo hallar una tripleta canónica así que es necesario aplicar el siguiente teorema como resultado auxiliar el cual proporciona una caracterización de esta.

Teorema 2.2.1. *(ρ, h, f) es una tripleta canónica si y sólo si para todo $x \in X$*

(a) $\rho(x) = \text{ínf}_{a\in A(x)} \int_X \rho(y)Q(dy|x,a)$.

(b) $\rho(x) + h(x) = \text{ínf}_{a\in A(x)} \left[c(x,a) + \int_X h(y)Q(dy|x,a) \right]$.

(c) $f(x) \in A(x)$, a) y b) alcanzan su mínimo es decir

$$\begin{aligned} \rho(x) &= \int_X \rho(y)Q(dy|x,f), \\ \rho(x) + h(x) &= c(x,f) + \int_X h(y)Q(dy|x,f) \end{aligned}$$

Las expresiones en c) son conocidas como *ecuaciones canónicas*.

Demostración. Supongamos que (ρ, h, f) es una tripleta canónica, de las ecuaciones de programación dinámica se tiene

$$V^*_{n+1} = \min_{a\in A(x)} \left[c(x,a) + \int V^*_n(y,h)Q(dy|x,a) \right], \tag{2.2.4}$$

sustituyendo V_n y V_{n+1} de la ecuación (2.2.3) en (2.2.4) se obtiene

$$(n+1)\rho(x) + h(x) = \min_{a\in A(x)} \left[c(x,a) + \int (n\rho(y) + h(y))Q(dy|x,a) \right], \tag{2.2.5}$$

para obtener (b) se toma $n = 0$ en (2.2.5).

Ahora, se mostrará (a); multiplicando la ecuación (2.2.5) por $\frac{1}{n}$ y después haciendo a n tender a infinito se obtiene

$$\begin{aligned} \rho(x) &= \lim_{n\to\infty} \min_{a\in A(x)} \left[\frac{c(x,a)}{n} + \int \left(\rho(y) + \frac{h(y)}{n} \right) Q(dy|x,a) \right], \\ &= \min_{a\in A(x)} \left[\int \rho(y)Q(dy|x,a) \right], \end{aligned}$$

es decir, se cumple (a).

Note que cualquier política determinista estacionaria $\hat{f}$ que satisface la ecuación (2.2.3) también cumple a) y b) con lo que se demuestra (c)

Ahora suponga que con $\hat{f}(x) \in A(x)$ alcanza su mínimo (a) y (b) es decir, se cumple (c), se mostrará que ρ, h y f satisfacen (2.2.3), para concluir que (ρ, h, f) es una tripleta canónica.

De (c) se obtiene

$$\rho(x_0) + h(x_0) = c(x_0, \hat{f}) + \int h(x_1) Q(dx_1|x_0, \hat{f})$$

$$\begin{aligned}
&= c(x_0, \hat{f}) + \int \left[c(x_1, \hat{f}) + \int h(x_2) Q(dx_2|x_1, \hat{f}) - \rho(x_1) \right] Q(dx_1|x_0, \hat{f}) \\
&= c(x_0, \hat{f}) + \int c(x_1, \hat{f}) Q(dx_1|x_0, \hat{f}) + \int \int h(x_2) Q(dx_2|x_1, \hat{f})(x_1) Q(dx_1|x_0, \hat{f}) - \\
&\quad \int \rho(x_1) Q(dx_1|x_0, \hat{f}) \\
&= c(x_0, \hat{f}) + E[c(x_1, \hat{f})] + E_{x_0}^{\hat{f}}[h(X_2)] - \int \rho(x_1) Q(dx_1|x_0, \hat{f}), \qquad (2.2.6)
\end{aligned}$$

sustituyendo $\rho(x_1)$ en la ecuación (2.2.6)

$$\begin{aligned}
2\rho(x_0) + h(x_0) &= E[c(x_0, \hat{f}) + c(x_1, \hat{f})] + E_{x_0}^{\hat{f}}[h(X_2)], \\
&= E\left[\sum_{t=0}^{2} 2c(x_t, \hat{f}) \right] + E_{x_0}^{\hat{f}}[h(X_2)], \\
&= V_2(\hat{f}, x) + E_{x_0}^{\hat{f}}[h(X_2)], \\
&= V_2(\hat{f}, x, h).
\end{aligned}$$

En general

$$n\rho(x) + h(x) = V_n(\hat{f}, x) + E_{x_0}^{\hat{f}}[h(X_2)] = V_n(\hat{f}, x, h). \qquad (2.2.7)$$

Resta demostrar que para todo $n \geq 0$, $x \in X$

$$V_n(\hat{f}, x, h) = V^*(x, h), \tag{2.2.8}$$

esto será demostrado por inducción sobre n:

para $n = 0$ se cumple (2.2.8), ahora supóngase que se cumple para $n > 0$, sustituyendo $V_n(\hat{f}, x, h)$ de (2.2.4) en la ecuación (2.2.7)

$$\begin{aligned} V^*_{n+1}(x,h) &= \min_{a\in A(x)} \left[c(x,a) + \int_X (n\rho(y) + h(y)) Q(dy|x,a) \right], \\ &\geq \min_{a\in A(x)} \left[c(x,a) + \int_X h(y) Q(dy|x,a) \right] + n \min_{a\in A(x)} \int_X \rho(y) Q(dy|x,a), \\ &= (n+1)\rho(x) + h(x), \\ &= V_{n+1}(\hat{f}, x, h), \end{aligned}$$

por lo tanto $V^*_{n+1}(x,h) \geq V_{n+1}(\hat{f}, x, h)$, y de (2.2.2) se tiene $V^*_{n+1}(x,h) \leq V_{n+1}(\hat{f}, x, h)$ de estas dos desigualdades se concluye que $V^*_{n+1}(x,h) = V_{n+1}(\hat{f}, x, h)$. □

Una suposición importante sobre h de una tripleta canónica (ρ, h, f) es que para toda política $\pi \in \Pi$ y $x \in X$:

$$\lim_{n\to\infty} \frac{E^\pi_x[h(X_n)]}{n} = 0, \tag{2.2.9}$$

con lo que se obtiene que la política determinista estacionaria $\hat{f}$ es óptima y la función de valor es ρ. Esto se muestra en el Teorema 2.2.2 en donde también se emplean los conceptos de política *CP-óptima* y *F-fuertemente CP-óptima* (véase [7]).

Definición 2.2.2. *Se dice que una política π^* es*

- fuertemente CP-óptima *si* $V(\pi^*, x) \leq \lim\inf_{n\to\infty} \frac{V_n(\pi,x)}{n}$.
- F-fuertemente CP-óptima $\lim_{n\to\infty} \left[V^*_n(x) - \frac{V_n(\pi^*,x)}{n} \right] = 0$.

donde $V^*_n(x) = \inf_{\pi\in\Pi} V_n(\pi, x) = V^*_n(x, 0)$.

Teorema 2.2.2. *Sea (ρ, h, f) una tripleta canónica.*

(a) Si h satisface (2.2.9) entonces $\hat{f}$ es CP-óptima y ρ es la función de valor, de hecho, para cada $x \in X$ se cumple

$$V^*(x) = \rho = V(\hat{f}, x) = \lim_{n\to\infty} \frac{V_n(\hat{f}, x)}{n} \tag{2.2.10}$$

(b) Más aún, si h satisface para cada $x \in X$ la ecuación

$$\lim_{n\to\infty} \sup_{\pi\in\Pi} \frac{E_x^\pi[h(X_n)]}{n} = 0, \tag{2.2.11}$$

entonces es F-fuertemente CP-óptima y fuertemente CP-óptima y de (2.2.10)

$$V^*(x) = \lim_{n\to\infty} \frac{V_n^*(\hat{f}, x)}{n}$$

Demostración. (a) Observe que de las ecuaciones (2.2.1) y (2.2.2) se tiene que $V_n(\pi, x, h) = V_n(\pi, x) + E_x^\pi[h(X_n)] \geq \text{ínf}_{\pi\in\Pi} V_n(\pi, x, h) = V_n^*(x)$, así, sustituyendo $V_n(\pi, x, h)$ (ver ecuación (2.2.3)

$$n\rho(x) + h(x) = V_n^*(x) \leq V_n(\pi, x) + E_x^\pi[h(X_n)],$$

ahora, multiplicando esta expresión por $\frac{1}{n}$ y tomando límite superior cuando n tiende a infinito

$$\begin{aligned} \frac{n\rho(x) + h(x)}{n} &\leq \frac{V_n(\pi, x)}{n} + \frac{E_x^\pi[h(X_n)]}{n}, \\ \liminf_{n\to\infty} \rho(x) + \frac{h(x)}{n} &\leq \liminf_{n\to\infty} \frac{V_n(\pi, x)}{n} + \liminf_{n\to\infty} \frac{E_x^\pi[h(X_n)]}{n}, \end{aligned}$$

ya que se tiene la condición (2.2.9) entonces ρ cumple la desigualdad

$$\rho(x) \leq \liminf_{n\to\infty} \frac{V_n(\pi, x)}{n} = V(\pi, x),$$

es decir, $\rho(x)$ es una cota superior, lo que implica que $\rho(x) \leq V^*(x)$. Por otro lado, de (2.2.3) y de la definición del *n-ésimo* costo total esperado se tienen las siguientes identidades

$$V_n(\hat{f}, x, h) = V_n(\hat{f}, x) + E_x^{\hat{f}}[h(X_n)] = n\rho(x) + h(x), \tag{2.2.12}$$

multiplicando por $\frac{1}{n}$ y tomando límite superior en ambas igualdades se obtiene

$$\begin{aligned} \limsup_{n\to\infty} \frac{V_n(\hat{f}, x, h)}{n} &= \limsup_{n\to\infty} \frac{V_n(\hat{f}, x)}{n} + \limsup_{n\to\infty} \frac{E_x^{\hat{f}}[h(X_n)]}{n} = \rho(x) \\ \limsup_{n\to\infty} \frac{V_n(\hat{f}, x, h)}{n} &= V(\hat{f}, s, x). \end{aligned} \tag{2.2.13}$$

Ahora, tomando límite inferior en (2.2.12) y multiplicando por $\frac{1}{n}$

$$\begin{aligned} \liminf_{n\to\infty} \frac{V_n(\hat{f}, x, h)}{n} &= \liminf_{n\to\infty} \frac{V_n(\hat{f}, x)}{n} + \liminf_{n\to\infty} \frac{E_x^{\hat{f}}[h(X_n)]}{n} = \rho(x) \\ \liminf_{n\to\infty} \frac{V_n(\hat{f}, x, h)}{n} &= V(\hat{f}, s, x). \end{aligned} \tag{2.2.14}$$

Igualando las dos expresiones, (2.2.13) y (2.2.14) se obtiene ρ

$$\rho(x) = \liminf_{n\to\infty} \frac{V_n(\hat{f}, x)}{n} = \limsup_{n\to\infty} \frac{V_n(\hat{f}, x)}{n} = \lim_{n\to\infty} \frac{V_n(\hat{f}, x)}{n},$$

entonces, para cada estado x se cumple

$$\rho(x) = V(\hat{f}, x) = \lim_{n\to\infty} \frac{V_n(\hat{f}, x)}{n},$$

por lo tanto se ha demostrado a).

(b) Note que, de la primera igualdad de (2.2.3) se tiene

$$V_n^*(x, h) = V_n(\hat{f}, x) + E_x^{\hat{f}}[h(X_n)], \tag{2.2.15}$$

por otro lado, de las ecuaciones (2.2.1) y (2.2.2) se obtiene

$$\begin{aligned} V_n^*(x,h) &= \inf_{\pi\in\Pi}(V_n(\pi,x)+E_x^\pi[h(X_n)]), \\ &\leq V_n^*(x)+\sup_{\pi\in\Pi}E_x^\pi[h(X_n)], \end{aligned}$$

entonces

$$\begin{aligned} 0\leq V_n^*(x,h)=V_n(\hat{f},x)+E_x^{\hat{f}}[h(X_n)] &\leq V_n^*(x)+\sup_{\pi\in\Pi}E_x^\pi[h(X_n)] \\ 0\leq V_n(\hat{f},x)-V_n^*(x) &\leq \sup_{\pi\in\Pi}E_x^\pi[h(X_n)]-E_x^{\hat{f}}[h(X_n)]. \end{aligned}$$

Como h satisface (2.2.11) entonces, multiplicando por $1/n$ y tomando límite

$$0\leq \lim_{n\to\infty}\left(\frac{V_n(\hat{f},x)-V_n^*(x)}{n}\right)\leq \lim_{n\to\infty}\left(\sup_{\pi\in\Pi}\frac{E_x^\pi[h(X_n)]}{n}-\frac{E_x^{\hat{f}}[h(X_n)]}{n}\right)=0. \tag{2.2.16}$$

se concluye que $\hat{f}$ es F-fuerte CP-óptima. Para probar que $\hat{f}$ es óptima se aplicara la ecuación (2.2.15) para obtener

$$V_n(\hat{f},x)+E_x^{\hat{f}}[h(X_n)]\leq V_n(\pi,x)+E_x^\pi[h(X_n)],$$

y de (2.2.11)

$$\liminf_{n\to\infty}V_n(\hat{f},x)\leq \liminf_{n\to\infty}\frac{V_n(\pi,x)}{n}.$$

Como el lado izquierdo es igual a $V(\hat{f},x)$ (ver ecuación 2.2.10) se sigue que $\hat{f}$ es óptima.

□

En el caso cuando $\rho(\cdot)$ es una constante, se denota como $\rho(x)=\rho^*$ para cada estado x las ecuaciones se reducen a

$$\begin{aligned} \rho^*+h(x) &= \inf_{a\in A(x)}\left[c(x,a)+\int h(y)Q(dy|x,a)\right], \qquad (2.2.17) \\ \rho^*+h(x) &= c(x,f)+\int h(y)Q(dy|x,f), \end{aligned}$$

la ecuación (2.2.17) es conocida como *ecuación de costo promedio óptima.*

Observe que del Teorema 2.2.1 (ρ^*, h, f) es una tripleta canónica si y sólo si la pareja (ρ^*, h) satisface la ecuación de costo promedio óptima, en algunos casos es suficiente con tener la solución (ρ^*, h) de la *desigualdad de costo promedio óptima* definida como

$$\rho^* + h(x) \geq \inf_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy|x,a) \right], \tag{2.2.18}$$

y un selector f tal que para cada estado $x \in X$ se cumple

$$\rho^* + h(x) \geq c(x,f) + \int h(y) Q(dy|x,f). \tag{2.2.19}$$

Lema 2.2.1. *(a) Si (2.2.19) se cumple y si* $\lim_{n\to\infty} \frac{E_x^{\hat{f}}[h(X_n)]}{n} \geq 0$ *entonces para cada estado* $x \in X$ *se cumple*

$$\rho^* \geq V_n(\hat{f}, x). \tag{2.2.20}$$

(b) Si la desigualdad en (2.2.19) se invierte, es decir,

$$\rho^* + h(x) \leq c(x,f) + \int h(y) Q(dy|x,f),$$

y si para todo $x \in X$

$$\lim_{n\to\infty} \frac{E_x^{\hat{f}}[h(X_n)]}{n} \leq 0,$$

entonces

$$\rho^* \leq V_n(\hat{f}, x).$$

Demostración. (a) Por iteración de (2.2.19) (análogo a 2.2.7) se cumple para todo $n \geq 1$

$$n\rho^* + h(x) \geq V_n(\hat{f}, x) + E_x^{\hat{f}}[h(X_n)], \tag{2.2.21}$$

dividiendo entre n y tomando límite inferior se obtiene (2.2.20).

(b) La prueba es análoga a a).

□

En la siguiente sección se darán condiciones que garantizan la existencia de políticas óptimas por medio de factor de descuento desvaneciente con la que es posible hallar una solución de la desigualdad de costo promedio óptima.

2.3. Factor de Descuento Desvaneciente.

Se dará un enfoque basado en el problema de costo descontado con un factor de descuento variable $\alpha \in (0,1)$ cercano a uno. Recordando a la función de valor V^* definida como

$$\begin{aligned} V^*_\alpha(\pi, x) &:= \sup_{\pi\in\Pi} V^\alpha(\pi, x), \qquad (2.3.1)\\ &= \sup_{\pi\in\Pi} E^\pi_x\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right] = \sum_{t=0}^{\infty} \alpha^t E^\pi_x[c(x_t, a_t)]. \end{aligned}$$

El intercambio entre esperanza y sumatoria es posible gracias al teorema de convergencia monótona (véase [3]).

Existe una relación entre el costo descontado $V^\alpha(\pi, x)$ y el costo promedio $V(\pi, x)$, tomando $c_t = E^\pi_x c(x_t, a_t)$, de la primera desigualdad en el Lema A.0.3, del Apéndice A, para toda x se cumple

$$\lim_{\alpha\to 1}\sup(1-\alpha)V^\alpha(\pi, x) \leq V(\pi, x),$$

entonces, de la ecuación (2.3.1) resulta

$$\lim_{\alpha\to 1}\sup(1-\alpha)V^\alpha_*(\pi, x) \leq V(\pi, x),$$

de la definición de la función de valor V^* en (2.1.2)

$$\lim_{\alpha\to 1}\sup(1-\alpha)V^\alpha_*(\pi, x) \leq V^*(\pi, x). \qquad (2.3.2)$$

Es decir, si la función V^α_* es multiplicada por $(1-\alpha)$, para un α cercana a uno, entonces se tiene una cota inferior de la función de costo promedio.

Otra relación entre los problemas de costo descontada y costo promedio está dada por la ecuación de costo descontado óptima

$$V_*^\alpha(x) = \min_{a\in A(x)} \left[c(x,a) + \alpha \int_X V_*^\alpha(y) Q(dy|x,a) \right]. \tag{2.3.3}$$

Sea m_α una constante (que se definirá más adelante) que depende de α, (con $\alpha \in (0,1)$), y defínase

$$\begin{aligned} h_\alpha(x) &:= V_*^\alpha(x) - m_\alpha, \\ \rho_\alpha &:= (1-\alpha) m_\alpha. \end{aligned} \tag{2.3.4}$$

Cuando α tiende a uno, la pareja $(\rho_\alpha, h_\alpha(\cdot))$ de (2.3.4) en algunos casos converge al par $(\rho^*, h_\alpha(\cdot))$, el cual satisface la ecuación de costo promedio óptima (2.2.17), en general esto no ocurre, sin embargo, bajo ciertas condiciones es posible hallar la solución de la desigualdad de costo promedio óptima, las cuales se presentan en la siguiente sección.

2.4. Desigualdad de Costo Promedio Óptimo.

En esta sección se darán algunas condiciones para la existencia de una solución a la desigualdad de costo promedio así como para la existencia de políticas estacionarias deterministas óptimas para lo que se supondrá lo siguiente

Suposición 2.4.1. *Existe un estado $z \in X$ y números $\beta \in (0,1)$ y $M \geq 0$ tales que*

(a) $(1-\alpha)V_*^\alpha(z) \leq M$ para cada $\alpha \in [\beta, 1)$. Más aún existe una constante $N \geq 0$ y una función $b(\cdot)$ sobre X tal que $h_\alpha(x) := V_*^\alpha(x) - V^\alpha(z)$.

(b) $-N \leq h_\alpha(x) \leq b(x)$ para todo $x \in X$ y $\alpha \in [\beta, 1)$.

La Suposición 2.4.1 implica que para cada estado x y para cada $\alpha \in (0,1)$, $V_*^\alpha(x)$ es finita.

Lema 2.4.1. *Bajo la Suposición 2.4.1 a) existe una constante ρ^* tal que $0 \leq \rho^* \leq M$ y una sucesión de factores descontados $\alpha(n)$ creciente a uno que satisface, para cada estado x*

$$\lim_{n\to\infty}(1-\alpha(n))V_*^{\alpha(n)}(x) = \rho^* \tag{2.4.1}$$

Demostración. Por hipótesis, existe un número $\rho^* \in [0, M]$ que es un punto límite de $(1-\alpha)V_*^{\alpha}(z)$ cuando α tiende a uno, es decir, para $\alpha(n)$ sucesión creciente a uno se tiene que

$$\lim_{n\to\infty}(1-\alpha(n))V_*^{\alpha(n)}(z) = \rho^*. \tag{2.4.2}$$

Observe que para todo $x \in X$ la siguiente relación se cumple

$$\begin{aligned}
|(1-\alpha(n))V^*_{\alpha(n)}(x) - \rho^*| &= |(1-\alpha(n))(V^*_{\alpha(n)}(x) - V^*_{\alpha(n)}(z) + V^*_{\alpha(n)}(z) - \rho)| \\
&\leq |(1-\alpha(n))(V^*_{\alpha(n)}(x) - V^*_{\alpha(n)}(z))| + \\
&\quad |(1-\alpha(n))(V^*_{\alpha(n)}(z) - \rho)| \\
&= (1-\alpha(n))|h_{\alpha(n)}(x)| + |(1-\alpha(n))V^*_{\alpha(n)}(z) - \rho^*|,
\end{aligned}$$

de la Suposición 2.4.1 *b*) se obtiene

$$\begin{aligned}
&(1-\alpha(n))|h_{\alpha(n)}(x)| + |(1-\alpha(n))V^*_{\alpha(n)}(z) - \rho^*| \\
&\leq (1-\alpha(n))\,\text{máx}\{N, b(x)\} + |(1-\alpha(n))V^*_{\alpha(n)}(z) - \rho^*|,
\end{aligned}$$

para todo $x \in X$ tomando límite cuando n tiende a infinito se tiene que

$$(1-\alpha(n))\,\text{máx}\{N, b(x)\} + |(1-\alpha(n))V^*_{\alpha(n)}(z) - \rho^*| \to 0,$$

por lo tanto

$$|(1-\alpha(n))V^*_{\alpha(n)}(x) - \rho^*| \leq 0,$$

con lo que se tiene la conclusión del Lema. □

Teorema 2.4.1. *Si se cumple la Suposición 1.4.1 entonces*

I) *Existen una constante $\rho^* > 0$, una función medible $h : X \to \mathbb{R}$ con*

$$-N \leq h(x) \leq b, \tag{2.4.3}$$

para cada estado $x \in X$, $h(z) = 0$ y un selector f tales que (ρ^*, h, f) satisfacen la desigualdad de costo promedio óptima (ecuaciones 2.2.18 y 2.2.19), es decir, para cada estado x se cumple

(a)

$$\rho^* + h(x) \geq \min_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy|x,a) \right].$$

(b)

$$\rho^* + h(x) \geq c(x,f) + \int h(y) Q(dy|x,f).$$

(c) $\hat{f}$ es óptima y ρ^* es la función de valor

$$V^*(x) = V(\hat{f}, x) = \rho^*, \tag{2.4.4}$$

para cada estado, es decir, $\rho^* = \text{ínf}_{x \in X} V^*(x) = \text{ínf}_{x \in X} \text{ínf}_{\pi \in \Pi} V(\pi, x)$, de hecho, cualquier selector $f \in \mathbb{F}$ que satisface b) también (2.4.4).

II) *Si $\hat{f}$ es una política determinista estacionaria óptima que satisface (2.4.4) entonces existe una función medible $\hat{h}(x) \geq 0$ con $x \in X$, tal que $(\rho^*, \hat{h}, f)$ satisface b) y por lo tanto a).*

Demostración. I) (a) Sea $\{\alpha(n)\}$ igual que en (2.4.1), se define para cada estado x

$$h(x) := \liminf_{n \to \infty} h_{\alpha(n)}(x) = \lim_{n \to \infty} H_n(x), \tag{2.4.5}$$

donde $H_n(x) := \text{ínf}_{k \geq n} h_{\alpha(k)}(x)$. Ya que se cumple la Suposición 2.4.1 *b*) la función h en (2.4.5) satisface la ecuación (2.4.3).

Por otro lado, H_n crece hacia h, es decir, para cada estado x

$$H_n(x) \uparrow h(x). \tag{2.4.6}$$

Ahora, sea $\rho_\alpha = (1-\alpha)V^\alpha(z)$ y escribiendo la EPD (2.3.3) de la forma

$$\rho_\alpha + h_\alpha(x) = \min_{a \in A(x)} \left[c(x.a) + \alpha \int_X h_\alpha(y) Q(dy|x,a) \right]. \tag{2.4.7}$$

Se usará el lema A.0.2 del apéndice A; en (2.4.7) tomando $\alpha = \alpha(n)$ y tomando límite superior cuando n tiende a infinito y de (2.4.2)-(2.4.6) se obtiene

$$\begin{aligned}
\rho^* + h(x) &= \liminf_{n\to\infty} \min_{a\in A(x)} \left[c(x,a) + \alpha(n) \int_X h_{\alpha(n)}(y) Q(dy|x,a) \right] \\
&\geq \lim_{n\to\infty} \min_{a\in A(x)} \left[c(x,a) + \alpha(n) \int_X H_n(y) Q(dy|x,a) \right] \\
&= \min_{a\in A(x)} \left[c(x,a) + \alpha(n) \int_X h(y) Q(dy|x,a) \right]. \qquad (2.4.8)
\end{aligned}$$

En conclusión, el intercambio entre "lím" y "mín"en (2.4.8) es valido con lo que se completa la prueba a).

(b) Observe que la desigualdad en a) es valida si se reemplaza $h(\cdot)$ por la función no negativa $h(\cdot) + N$ así que existe un selector f, que satisface b) (véase [12]).

(c) Sea f un selector que satisface b) entonces de (2.4.3) se cumple la desigualdad (2.2.21) para cada $n \geq 1$

$$n\rho^* + h(x) \geq V_n(\hat{f}, x) - N,$$

entonces $\rho^* \geq V(\hat{f}, x)$, ya que:

$$\begin{aligned}
n\rho^* + h(x) &\geq V_n(\hat{f}, x) - N \\
n\rho^* &\geq V_n(\hat{f}, x) \\
\rho^* &\geq \frac{V_n(\hat{f}, x)}{n} = V(\hat{f}, x).
\end{aligned}$$

Por otro lado, de (2.4.1) y la definición de V^*, se tiene $\rho^* \leq V^*(x) \leq V(\hat{f}, x)$, pues

$$\rho = \lim_{n\to\infty} (1 - \alpha(n)) V^{\alpha}(n)^*(z) = \lim_{\alpha\uparrow 1} (1-\alpha) V_*^{\alpha}(x) \leq V^*(x)$$

es decir, (2.4.4) se cumple para cualquier $f \in \mathbb{F}$, con lo que se satisface b).

II) Defínase

$$h_0 = \rho_0 = 0,$$

y para todo estado x y $n \geq 1$

$$\begin{aligned} \rho_n(x) &:= V_n(\hat{f}, x) - V_{n-1}(\hat{f}, x), \\ M_n &:= \inf_{x \in X} V_n(\hat{f}, x), \\ h_n(x) &:= V_n(\hat{f}, x) - M_n, \\ \hat{h}(x) &:= \liminf_{m \to \infty} \frac{1}{m} \sum_{n=1}^{m+1} h_n(x). \end{aligned} \tag{2.4.9}$$

Note que $h_n(x) \geq 0$ para todo $n \geq 0$ y $\sum_{n=1}^{m} \rho_n(x) = V_m(\hat{f}, x)$; por otro lado,

$$V_n(\hat{f}, x) = c(x, f) + \int V_{n-1}(\hat{f}, y) Q(dy|x, f),$$

reescribiendo esta expresión

$$\rho_n(x) + h_{n-1}(x) = c(x, f) + \int h_{n-1}(y) Q(dy|x, f),$$

sumando desde $n = 1$ hasta m

$$\begin{aligned} \sum_{n=1}^{m} \rho_n(x) + \sum_{n=1}^{m} h_{n-1}(x) &= \sum_{n=1}^{m} c(x, f) + \sum_{n=1}^{m} \int h_{n-1}(y) Q(dy|x, f), \\ V_m(\hat{f}, x) + \sum_{n=1}^{m-1} h_n(x) &= mc(x, f) + \int \sum_{n=1}^{m} h_{n-1}(y) Q(dy|x, f), \\ V_m(\hat{f}, x) + \sum_{n=1}^{m-1} h_n(x) &= mc(x, f) + \int \sum_{n=1}^{m-1} h_n(y) Q(dy|x, f), \end{aligned}$$

multiplicando la última igualdad por $\frac{1}{m}$ se obtiene

$$\begin{aligned}\frac{V_m(\hat{f},x)}{m}+\frac{1}{m}\sum_{n=1}^{m-1}h_n(x) &= \frac{mc(x,f)}{m}+\frac{1}{m}\int\sum_{n=1}^{m-1}h_n(y)Q(dy|x,f),\\ V(\hat{f},x)+\frac{1}{m}\sum_{n=1}^{m-1}h_n(x) &= c(x,f)+\int\frac{1}{m}\sum_{n=1}^{m-1}h_n(y)Q(dy|x,f),\end{aligned}$$

tomando límite inferior cuando m tiende a infinito en la última igualdad se obtiene

$$V(\hat{f},x)+\liminf_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m-1}h_n(x)=c(x,f)+\liminf_{m\to\infty}\int\frac{1}{m}\sum_{n=1}^{m-1}h_n(y)Q(dy|x,f),$$

$$V(\hat{f},x)+\hat{h}(x)=c(x,f)+\liminf_{m\to\infty}\int\frac{1}{m}\sum_{n=1}^{m-1}h_n(y)Q(dy|x,f),$$

finalmente aplicando el Lema de Fatou (véase [3]),

$$\begin{aligned}\rho^*+\hat{h}(x) &\geq c(x,f)+\int\limsup_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m-1}h_n(y)Q(dy|x,f),\\ &= c(x,f)+\int\hat{h}(y)Q(dy|x,f),\\ \rho^*+\hat{h}(x) &\geq \min\left[c(x,a)+\int\hat{h}(y)Q(dy|x,a)\right].\end{aligned}$$

Esto es, $f\in\mathbb{F}$ y $\hat{h}(\cdot)\geq 0$ definido en (2.4.9) satisfacen la la desigualdad de costo promedio óptima.

□

Para concluir esta sección se presentará una condición que implica la Suposición 2.4.1 y también se presentará una condición equivalente.

Suposición 2.4.2. *a)* $V_*^\alpha(x)<\infty$ *para cada estado* x *y* $\alpha\in(0,1)$ *más aún existe un estado* z*, contantes* $N'\geq 0$ *y* $\beta'\in(0,1)$*, una función medible no negativa* $b'(\cdot)$ *y un selector* f' *tal que* $h_\alpha(x)=V_*^\alpha(x)-V_*^\alpha(z)$*.*

b) $-N'(x) \leq h_\alpha(x) \leq b'(x)$ *para cada* $x \in X$ *y* $\alpha \in [\beta', 1]$.

c) $\int_X b'(y)Q(dy|x, f') < \infty$.

Suposición 2.4.3. *a) Existe una política* $\hat{\pi}$ *y un estado inicial* $\hat{x}$ *tal que* $V(\hat{\pi}, \hat{x}) < +\infty$.

b) Existe $\hat{\beta} \in [0, 1)$ *tal que para cada estado* x, $\sup_{\alpha \in (\hat{\beta}, 1)} g_\alpha(x) < +\infty$ *donde*

$$g_\alpha(x) := V_*^\alpha(x) - m_\alpha \text{ con } m_\alpha := \inf_{x \in X} V_*^\alpha(x).$$

Teorema 2.4.2. *Bajo la suposición 1.4.1*

a) La suposición 2.4.2 implica la suposición 2.4.1.

b) La suposición 2.4.3 es equivalente a la suposición 2.4.1.

Demostración. (a) Si la Suposición 2.4.2 se cumple entonces también se cumple *b*) de la Suposición 2.4.1 tomando $\beta = \beta'$, $N = N'$ y $b(\cdot) = b'(\cdot)$. Observe que de la Suposición 2.4.2 *a*) y de la Suposición 1.4.1 la ecuación de programación dinámica está bien definida.

Por lo tanto para todo $\alpha \in (0, 1)$ se cumple (2.3.3) es decir

$$\begin{aligned}
(1-\alpha)V_*^\alpha(z) &= V_*^\alpha(z) - \alpha V_*^\alpha(z), \\
&\leq c(z, f') + \alpha \int_X V_*^\alpha(y)Q(dy|z, f') - \alpha V_*^\alpha(z), \\
&= c(z, f') + \alpha \int_X \big(V_*^\alpha(y) - V_*^\alpha(z)\big)Q(dy|z, f'), \\
&= c(z, f') + \alpha \int_X h_\alpha(y)Q(dy|z, f'),
\end{aligned}$$

de la Suposición 2.4.2 *b*) se cumple la desigualdad

$$c(z, f') + \alpha \int_X h_\alpha(y)Q(dy|z, f') \leq c(z, f') + \alpha \int_X b(y)Q(dy|z, f').$$

Así que, tomando a M como $c(z, f') + \alpha \int_X b(y)Q(dy|z, f')$ se obtiene que para todo $\alpha \in (0, 1)$ se cumple *a*) de la Suposición 2.4.1 .

(b) Por demostrar que la Suposición 2.4.1 implica la Suposición 2.4.3. Ya que se cumplen las Suposiciones 1.4.1 y 2.4.1 se tiene, por el Teorema 2.4.1, que existe una política óptima $\hat{f}$, entonces, tomando $\hat{\pi} := \hat{f}$ se obtiene a) de la Suposición 2.4.3.

Por otro lado, de la Suposición 2.4.1 b), para cada $\alpha \in [\beta, 1)$ se cumple

$$\inf_{x \in X} h_\alpha(x) = V_*^\alpha(z) - m_\alpha = -g_\alpha(z) \geq -N,$$

entonces, para cada estado x y $\alpha \in [\beta, 1)$

$$g_\alpha(x) = V_*^\alpha(x) - m_\alpha = h_\alpha(x) + g_\alpha(z) \leq b(x) + N < \infty,$$

si se toma $\hat{\beta} = \beta$ se tiene la conclusión de la Suposición 2.4.3 b).

Ahora se mostrará que la Suposición 2.4.3 implica la Suposición 2.4.1: defínase $z = \hat{x}$, entonces, en particular se cumple a) de la Suposición 2.4.3 y la desigualdad (2.3.2) también se cumplen,

$$\limsup_{\alpha \to 1}(1-\alpha)m_\alpha \leq \limsup_{\alpha \to 1}(1-\alpha)V_*^\alpha(z) < V^*(z) < \infty, \qquad (2.4.10)$$

más aún

$$\limsup_{\alpha \to 1}(1-\alpha)m_\alpha \leq \inf_{x \in X} V^*(x) < \infty,$$

observe que las funciones $\alpha \to m_\alpha$ y $\alpha \to V_*^\alpha(z)$ son decrecientes en $\alpha \in (0, 1)$, además m_α y $V_*^\alpha(z)$ son finitas para cualquier $\alpha \in (0, 1)$.

De la ecuación (2.4.10), existe una constantes $M \geq 0$ y $\beta_0 \in (0, 1)$ tal que para cada $\alpha \in [\beta_0, 1)$

$$0 \leq (1-\alpha)V_*^\alpha(z) \leq M.$$

Por otro lado, note que $h_\alpha(x) = V^\alpha(x) - V^\alpha(z) = g_\alpha(x) - g_\alpha(z)$; por lo tanto para toda $\alpha \in (0, 1)$

$$-g_\alpha(z) \leq h_\alpha(x) \leq g_\alpha(x),$$

entonces, definiendo

$$N := \sup_{\hat{\beta} < \alpha < 1} g_\alpha(z), \quad b(x) := \sup_{\hat{\beta} < \alpha < 1} g_\alpha(x), \quad x \in X$$

y $\beta := \text{máx}\{\beta_0, \hat{\beta}\}$, se obtiene la Suposición 2.4.1.

□

En la siguiente sección se presentan dos ejemplos con el criterio de costo promedio esperado.

2.5. Ejemplos.

Ejemplo 2.5.1. *Para un número natural fijo N y un número $p \in [0,1]$ que se supondrá distinto de $1/2$ se tiene:*

- *Espacio de estados del sistema $X = \{0, 1, 2, \dots N\}$.*
- *Espacio de acciones $A = \{0, 1, 2, \dots, [N/2]\}$ donde $[N/2]$ representa la parte entera de $N/2$.*
- *El espacio de acciones admisibles para estado x, $A(x) = \{0, 1, 2, \dots, \text{mín}\{x, N - x\}\}$.*
- *La ley de transición está dada por*

$$\begin{aligned} q_{x,x+a}(a) &= p, \\ q_{x,x-a}(a) &= q = 1 - p, \\ q_{N,0}(a) &= 1, \\ q_{0,0}(a) &= 1. \end{aligned}$$

- *costo $c(x, a) = x$*

Se busca una política π^* tal que, para cada estado x

$$V(\pi^*, x) = \sup_{\pi \in \Pi} V(\pi, x) = V^*(x).$$

Obsérvese que se cumplen las suposiciones

(a)
- *c es semicontinua inferiormente.*
- *c es no negativa.*
- *c es inf-compacta sobre $\mathbb{K}$.*

(b) Existe un estado $z = 0$ tal que la función de valor $V(\pi, 0)$ es finita,

$$V(\pi, 0) = E_0^{\pi}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right] = 0,$$

pues $q_{0,0} = 1$.

Aplicando la técnica de programación dinámica hacia adelante se tiene

$$\begin{aligned} V_0(x) &= 0, \\ V_1(x) &= \min_{a \in A(x)} \{x\} = x, \end{aligned}$$

por lo tanto $f_1(x)$ es cualquier acción admisible. Ahora calculando $V_2(x)$

$$\begin{aligned} V_2(x) &= \min_{a \in A(x)} \Big\{ c(x_t, a_t) + \alpha E[V_1(x_{t+1})] \Big\}, \\ &= \min_{a \in A(x)} \Big\{ x + \alpha(pV_1(x + a) + qV_1(x - a)) \Big\}, \\ &= \min_{a \in A(x)} \Big\{ x + \alpha(x(p + q) + a(p - q)) \Big\}, \\ &= \min_{a \in A(x)} \Big\{ x + \alpha(x(p - q)) \Big\}, \end{aligned}$$

pues $p + q = 1$. Se deben considerar dos casos,

I*) $p > q$*

$$V_2(x) = x + \alpha(x + \min\{x, 3 - x\}(p - q)),$$

así $f_2(x) = \min\{x, 3 - x\} = 1$, por lo tanto

$$\begin{aligned} V_2(x) &= x + \alpha(p - q) \\ V_3(x) &= x(1 + \alpha) + (p - q)(1 + \alpha)\alpha \end{aligned}$$

II) $p < q$

$$V_2(x) = x(1+\alpha),$$

entonces $f_2(x) = 0$.

Para el caso ii)

$$\begin{aligned} V_3(x) &= x(1+\alpha+\alpha^2), \\ V_4(x) &= x(1+\alpha+\alpha^2+\alpha^3) \end{aligned}$$

en general

$$V_n(x) = x\sum_{k=0}^{n-1}\alpha^k,$$

tomando límite cuando n tiende infinito

$$V_n \to \frac{x}{1-\alpha}, \quad n \to \infty.$$

Así que para α *cercano a uno*

$$(1-\alpha)V_\alpha(x) \to x, \quad \textit{cuando} \quad \alpha \to 1,$$

es decir, el valor óptimo para cada estado x *es* x.

Para el caso i), $p > q$*, en general*

$$\begin{aligned} V_n(x) &= \min_{a\in A(x)}\left\{x+\alpha\left[pV_{n-1}(x+a)+qV_{n-1}(x-a)\right]\right\}, \\ &= \min_{a\in A(x)}\left\{x+\alpha\left[p((x+a)Q_{n-2}+R_{n-2})+q((x-a)Q_{n-2}+R_{n-2})\right]\right\}, \\ &= \min_{a\in A(x)}\left\{x+\alpha\left[Q_{n-2}x+Q_{n-2}(p-q)a+R_{n-2}\right]\right\}, \\ &= x(1+Q_{n-2})+\alpha(Q_{n-2}(p-q)+R_{n-2}). \end{aligned}$$

por lo tanto

$$V_n(x) = Q_{n-1}x + R_{n-1}(x),$$

donde

$$\begin{aligned} Q_n &= 1 - Q_{n-1}, \\ R_n &= \alpha(Q_{n-1}(p-q) + R_{n-1}). \end{aligned} \tag{2.5.1}$$

En general

$$Q_n = \sum_{k=0}^{n-1} \alpha^k,$$

tomando límite cuando n tiende infinito se tiene que

$$\lim_{n\to\infty} Q_n = Q = \frac{1-\alpha^n}{1-\alpha}.$$

Ya que $V_n(\cdot)$ converge a V y Q_n es convergente tomando límite en (2.5.1) se obtiene

$$\lim_{n\to\infty} R_n = R = \frac{(p-q)\alpha}{(1-\alpha)}.$$

Así

$$(1-\alpha)V^*(x) = (1-\alpha^n)x + (p-q)\alpha;$$

Para un α cercano a uno se tiene que el valor óptimo es

$$V^*(x) = \alpha(p-q);$$

Ejemplo 2.5.2 (Inventarios)**.** *Considere el Ejemplo de inventarios en la sección 1.5.2 con los valores $K = 5$, $c(x) = 10x$, $g(x) = 0$, $h(x) = 9x$, $M = 3$, $N = 4$, ahora será resuelto el problema de costo promedio.*

Recordando, el costo está dado por

x/a	$c_t(x,a)$			
	0	*1*	*2*	*3*
0	*0*	*47/3*	*89/3*	*47*
1	*2/3*	*32/3*	*32*	$\times$
2	*14/3*	*27*	$\times$	$\times$
3	*12*	$\times$	$\times$	$\times$

La matriz de transición es

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 5/6 & 1/6 & 0 & 0 \\ 1/2 & 1/3 & 1/6 & 0 \\ 1/6 & 1/3 & 1/3 & 1/6 \end{bmatrix}$$

Ya se han verificado las Suposición 1.4.1, y a) de la Suposición 2.4.3 (ver Sección 1.5.2)

entonces, aplicando la técnica de programación dinámica.

$$\begin{aligned} V_0(x) &= 0, \\ V_1(x) &= \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{y \in X} Q(y|x,a) V_0(y) \right\}, \end{aligned}$$

con lo que se obtiene

$$\begin{aligned} V_1(0) &= 0, \\ V_1(1) &= 2/3, \\ V_1(2) &= 14/3, \\ V_1(3) &= 12. \end{aligned}$$

De la misma forma se obtiene $V_2(x)$ utilizando la ecuación

$$V_2(x) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha \sum_{y \in X} Q(y|x,a) V_2(y) \right\},$$

y así sucesivamente hasta que la diferencia entre V_n y V_{n-1} sea menor que ε, tomando $\alpha = 0.5$ y $\varepsilon = 0.005$.

Ya que los cálculos se vuelven más complejos en cada periodo se calcularon empleando en software R, los resultados se presentan a continuación

	[V1]	[V2]	[V3]	[V4]	[V5]	[V6]
[0]	0	48.88889	24.44444	12.22222	6.11111	3.05555
[1]	0.66666	33.88889	23.86111	12.84028	6.82928	3.78207
[2]	4.66666	28.88889	24.94444	16.83333	11.26504	8.27141
[3]	12.00000	13.88889	27.69444	24.47917	20.00405	17.19198

	[V7]	[V8]	[V9]	[V10]	[V11]	[V12]
[0]	1.52777	0.76388	0.38194	0.19097	0.09548	0.04774
[1]	2.25498	1.49115	1.10921	0.91824	0.82275	0.77501
[2]	6.75018	5.98695	5.60507	5.41411	5.31862	5.27088
[3]	15.69620	14.93619	14.55469	14.36376	14.26828	14.22054

	[V13]	[V14]	[V15]
[0]	0.02387	0.01193	0.00596
[1]	0.75114	0.73920	0.73324
[2]	5.24701	5.23507	5.22910
[3]	14.19667	14.18473	14.17877

A partir de $V_{14}(x)$ *la diferencia* $|V_{n+1}(x) - V_n(x)|$ *es menor que* ε *para cada estado, así que*

$$\begin{aligned} V_*^\alpha(0) &= 0.00596, \\ V_*^\alpha(1) &= 0.73324, \\ V_*^\alpha(2) &= 5.22910, \\ V_*^\alpha(3) &= 14.17877, \end{aligned}$$

De aquí, se cumple b) de la Suposición 2.4.3, por lo tanto, del Teorema 2.4.2 también se cumple la Suposición 2.4.1.

Así tomando α *cercano a uno, por ejemplo* 0.9 *se tiene que el valor óptimo para cada estado es*

$$\begin{aligned}
(1-\alpha)V_*^\alpha(0) &= (0.1)V_*^\alpha(0) = 0.000596,\\
(1-\alpha)V_*^\alpha(1) &= (0.1)V_*^\alpha(0) = 0.073324,\\
(1-\alpha)V_*^\alpha(2) &= (0.1)V_*^\alpha(0) = 0.522910,\\
(1-\alpha)V_*^\alpha(3) &= (0.1)V_*^\alpha(0) = 1.417877.
\end{aligned}$$

Este ejemplo puede ser resuelto usando un criterio análogo tomando el caso sensible al riesgo, esta teoría será presentada en el siguiente capítulo.

Capítulo 3

Modelos de Control Sensibles al Riesgo.

Los procesos de decisión de Markov se dividen en dos tipos: *neutral* y *sensible al riesgo*, para distinguir ambos tipos se emplea una constante λ conocida como coeficiente de sensibilidad al riesgo; si $\lambda = 0$ se trata del caso neutral al riesgo, si $\lambda \neq 0$ el proceso de control de Markov es sensible al riesgo (ver [1],[13], [17]).

Antes de plantear formalmente el problema es necesario presentar dos conceptos que serán de utilidad en su interpretación.

- *Estado de la naturaleza*: Un estado de la naturaleza es la descripción de un cierto resultado de incertidumbre (en el caso de los PDMs es representado por el espacio de estados).
- *Certeza equivalente*: La certeza equivalente o plan consumo cierto es aquel en que el número de consumo (costos) no varía en los estados de la naturaleza.

En este capítulo será expuesto el problema de costo promedio sensible al riesgo. Se tomará al espacio de estados X finito, el espacio de acciones A como un espacio de Borel y la función de costo c acotada.

Notación: Sea D un espacio de Borel y $L : D \to \mathbb{R}$ una función medible,

se denotará a la norma de L como

$$||L|| := \sup_{x \in D} |L(x)|$$

El planteamiento del problema se presenta a continuación.

3.1. Función de Utilidad

Cuando se considera el caso sensible al riesgo el resultado de un proceso de Markov de costos (recompensas) es evaluado usando la función de utilidad con una constante de sensibilidad al riesgo $\lambda \in \mathbb{R}$ (véase [17]), las preferencias de un consumidor pueden ser representadas por una función real llamada *función de utilidad*, que asigna un número real a cada resultado. Gracias a los trabajos de J. von Neumann y O. Morgenstern en [20] es posible representar preferencias bajo condiciones de incertidumbre, la función de utilidad es denotada por $U_\lambda(Z)$ a través de la cual es posible medir la sensibilidad al riesgo que tiene el controlador.

En el caso de los PDM la función de utilidad modela la preferencia sobre las políticas; teniendo dos política π y π' es posible compararlas en función del valor esperado de su utilidad, si $E[U(\pi)] \geq E[U(\pi')]$ entonces se prefiere la política π'.

Otro concepto fundamental es la función llamada *certeza equivalente* definida a continuación

Definición 3.1.1. *Sea Z una variable aleatoria y suponga que el valor esperado de $U_\lambda(Z)$ está bien definido. La certeza equivalente $\mathbb{E}(\lambda, Z)$ de Z con respecto a U_λ está dada por*

$$\mathbb{E}(\lambda, Z) = \begin{cases} \frac{1}{\lambda}\ln(E[e^{\lambda Z}]) & , \quad si\ \lambda \neq 0, \\ E[Z] & , \quad si\ \lambda = 0. \end{cases}$$

Existe una relación entre la función de certeza equivalente y la función de utilidad: $U_\lambda(\mathbb{E}(\lambda, Z)) = \mathbb{E}[U_\lambda(Z)]$, es decir, el controlador tiene la opción de intercambiar entre la oportunidad de obtener una recompensa aleatoria Z por la certeza equivalente, cuando se le presenta la oportunidad de pagar un monto fijo c para evitar un costo aleatorio Z; este aceptará si $c \leq \mathbb{E}[\lambda, Z]$

y será rechazada si $c > \mathbb{E}[\lambda, Z]$ (véase [1]).

La noción de neutralidad al riesgo corresponde al caso en que la función de utilidad es la función identidad mientras que el controlador es averso al riesgo cuando $E[U(Z)] > U(E[Y])$, si el costo aleatorio Y no es constante, de la desigualdad de Jensen (véase [16]), es equivalente a que la función de utilidad sea convexa, en el caso en que sea amante al riesgo, la función de utilidad es concava.

Se define $\Delta(Y) = \mathbb{E}(\lambda, Y) - E[Y]$ a la *prima de riesgo* correspondiente a una variable aleatoria Y (véase [2]).

Sea Z una variable aleatoria acotada con media cero y varianza uno, defínase otra variable aleatoria $Y(\sigma) = y + \sigma Z$ para cada $\sigma > 0$; observe que, cuando σ tiende a cero, $Y(\sigma)$ converge en probabilidad a y, además $E[Y(\sigma)] = y$ y $Var(Y(\sigma)) = \sigma^2$, lo que implica que $\Delta(Y(0)) = 0$.

Proposición 3.1.1. *Suponga que la función de utilidad U tiene derivada continua de orden dos sobre los reales y con primera derivada no negativa entonces, cuando σ tiende a cero*

$$\frac{\Delta(Y(\sigma))}{\sigma^2} \to \frac{U''(y)}{2U'(y)}.$$

Para la demostración de este resultado véase [2].

La Proposición 3.1.1 , muestra que para un costo aleatorio Y con valores en una vecindad pequeña de y se tiene que $\Delta(Y)$ es proporcional a la varianza de Y y además que la constante de proporcionalidad está dada por $\frac{U''(y)}{U'(y)}$, con lo que se obtiene que $U(y) = ae^{\lambda y} + b$ para cada número real y con a y b constantes y $a > 0$. Dado que la función de utilidad es determinada por una transformación se supondrá que el controlador valora un costo aleatorio Y utilizando la función de utilidad exponencial $e^{\lambda y}$. Por lo tanto, la función de utilidad queda definida de la siguiente forma

Definición 3.1.2. *Se define la función de utilidad para todo $x \in \mathbb{R}$ como*

$$U_\lambda(x) = \begin{cases} sign(\lambda)e^{\lambda x} & , \quad si\ \lambda \neq 0, \\ x & , \quad si\ \lambda = 0, \end{cases}$$

donde $sign(\lambda) = 1$ si $\lambda > 0$ y $sign(\lambda) = -1$ si $\lambda < 0$

Observe que $U_\lambda(x)$ es una función continua, estrictamente creciente y convexa para $\lambda > 0$ (amante al riesgo) y concava cuando $\lambda < 0$ (averso al riesgo), (véase [17]).

3.2. Problemas de Decisión de Markov Sensibles al Riesgo.

La función objetivo empleada será el costo promedio esperado sensible al riesgo; dado un entero positivo n, considere el costo total $\sum_{t=0}^{n-1} c(x_t, a_t)$ pagado después de aplicar las primeras n acciones $a_0, a_1, \ldots, a_{n-1}$. Dada la política π y estado inicial x la certeza equivalente de un costo aleatorio es

$$V_n(\pi, x) = \frac{1}{\lambda} \log \left(E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} c(x_t, a_t)} \right] \right), \tag{3.2.1}$$

así que el costo promedio en el estado inicial x correspondiente a la política $\pi \in \Pi$ está dado por

$$V(x, \pi) = \limsup_{n \to \infty} \frac{1}{n} V_n(\pi, x). \tag{3.2.2}$$

El objetivo es hallar una política π^* que minimice el criterio, es decir,

$$V^*(x) = \inf_{\pi \in \Pi} V(x, \pi), \tag{3.2.3}$$

una política $\pi^* \in \Pi$ es llamada *óptima* si $V(x, \pi^*) = V^*$ para cada $x \in X$. Para resolver este problema son necesarias algunas condiciones.

Suposición 3.2.1. I) *Para cada estado $x \in X$, el espacio de acciones admisibles, $A(x)$ es compacto.*

II) *La función de costo es no negativa y continua.*

III) *Para todo $x, y \in X$, la función $a \mapsto Q(y|x, a)$ es continua sobre $A(x)$.*

Observe que de las ecuaciones (3.2.1), (3.2.2) y (3.2.3) y de la Suposición 3.2.1 se tiene

$$\min_{(x,a)} c(x, a) \leq V(\pi, \cdot) \leq \max_{(x,a)} c(x, a). \tag{3.2.4}$$

Una herramienta empleada para resolver el problema de control óptimo es la técnica de programación dinámica para lo que se requiere el Teorema de Verificación (véase [10]), antes de ser enunciado se probará el Lema 3.2.1 que será empleando en la demostración del teorema.

Lema 3.2.1. *La igualdad*

$$E_x^{\pi}\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}\sum_{y\in X}e^{h(y)}p(y|x_{T-1},a_{T-1})\right]=e^{h(x)}, \qquad (3.2.5)$$

se cumple.

Demostración. la demostración se sigue por inducción sobre T

Para $T=1$

$$E_x^{\pi}\left[\frac{e^{h(x_0)}}{\sum_{y\in X}e^{h(y)}p(y|x_0,a_0)}\sum_{y\in X}e^{h(y)}p(y|x_0,a_0)\right]=e^{h(x_0)}=e^{h(x)}.$$

Suponga que 3.2.5 se cumple para $T-1$, se demostrará para T

$$E_x^{\pi}\left[\prod_{t=0}^{T}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}\sum_{y\in X}e^{h(y)}p(y|x_T,a_T)\right]$$

$$\begin{aligned}
&= E_x^{\pi}\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}\frac{e^{h(x_T)}}{\sum_{y\in X}e^{h(y)}p(y|x_T,a_T)}\sum_{y\in X}e^{h(y)}p(y|x_T,a_T)\right]\\
&= E_x^{\pi}\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}e^{h(x_T)}\right]
\end{aligned}$$

tomando esperanza

$$
\begin{aligned}
& E_x^\pi\left[E_x^\pi\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}e^{h(x_T)}\Big|h_{T-1},a_T\right]\right] \\
= \; & E_x^\pi\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}E_x^\pi\left[e^{h(x_T)}\big|h_{T-1},a_T\right]\right] \\
= \; & E_x^\pi\left[\prod_{t=0}^{T-1}\frac{e^{h(x_t)}}{\sum_{y\in X}e^{h(y)}p(y|x_t,a_t)}\sum_{y\in X}e^{h(y)}p(y|x_{T-1},a_{T-1})\right] \\
= \; & e^{h(x)}.
\end{aligned}
$$

□

Teorema 3.2.1. *(Teorema de Verificación)*
Suponga que existe un número g y una función acotada $h : X \to \mathbb{R}$ tal que para cada estado x,

$$e^{g+h(x)} = \min_{a\in A(x)}\left\{e^{\lambda c(x,a)}\sum_{y\in X}e^{h(y)}p(y|x,a)\right\}, \tag{3.2.6}$$

entonces, para cada estado $x \in X$ y para toda política $\pi \in \Pi$

$$g\frac{1}{\lambda} \leq V(x,\pi),$$

más aún, si f^* es una política estacionaria, tal que $f^*(x)$ alcanza el mínimo en el lado derecho de la ecuación (3.2.6) para cada $x \in X$ entonces f^* es óptima y

$$g = \lim_{T\to\infty}\frac{1}{T}\log E_x^{f^*}\left[e^{\lambda\sum_{t=0}^{T-1}c(x_t,a_t)}\right].$$

Demostración. La demostración se hará por inducción sobre T; Sea $x \in X$ y $\pi \in \Pi$ y $T \geq 1$.

Para $T = 1$ observe que de (3.2.6) se cumple

$$
\begin{aligned}
e^{g+h(x_0)} &\leq e^{\lambda c(x,a)}\sum_{y\in X}e^{h(y)}p(y|x,a), \\
\frac{e^{g+h(x_0)}}{\sum_{y\in X}e^{h(y)}p(y|x_0,a)} &\leq e^{\lambda c(x,a)},
\end{aligned}
$$

ahora, suponga que se cumple

$$\begin{aligned} e^{\lambda \sum_{t=0}^{T-1} c(x,a)} &\geq \prod_{t=0}^{T-1} e^{g} \frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x_t,a_t)}, \\ &= e^{gT} \prod_{t=0}^{T-1} \frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x_t,a_t)}. \end{aligned}$$

Por lo tanto se obtiene

$$\begin{aligned} e^{\lambda \sum_{t=0}^{T} c(x,a)} &= e^{\lambda \sum_{t=0}^{T-1} c(x,a)} e^{\lambda c(x_T,a_T)} \\ &\geq \prod_{t=0}^{T-1} e^{g} \frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x_t,a_t)} e^{g} \frac{e^{h(x_T)}}{\sum_{y\in X} e^{h(y)} p(y|x_T,a_T)}, \end{aligned}$$

tomando esperanza en ambos lados de la desigualdad se tiene

$$\begin{aligned} E_x^{\pi} \exp\left[\lambda \sum_{t=0}^{T-1} c(x_t,a_t)\right] &\geq E_x^{\pi} \prod_{t=0}^{T-1} \left[e^{g} \frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x,a)} \right], \quad (3.2.7) \\ &= e^{g} E_x^{\pi} \prod_{t=0}^{T-1} \left[\frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x,a)} \right], \end{aligned}$$

del Lema 3.2.1 se cumple la igualdad

$$E_x^{\pi}\left[\prod_{t=0}^{T-1} \frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x_t,a_t)} \sum_{y\in X} e^{h(y)} p(y|x_{T-1},a_{T-1}) \right] = e^{h(x)}, \quad (3.2.8)$$

y ya que h es acotada, existen constantes M_1 y M_2 tales que

$$\begin{aligned} M_1 &\leq \frac{e^{h}}{\sup_{x\in X, a\in A(x)} \sum_{y\in X} e^{h(y)} p(y|x,a)} \leq E_x^{\pi} \prod_{t=0}^{T-1} \left[\frac{e^{h(x_t)}}{\sum_{y\in X} e^{h(y)} p(y|x,a)} \right] \\ &\leq \frac{e^{h}}{\text{ínf}_{x\in X, a\in A(x)} \sum_{y\in X} e^{h(y)} p(y|x,a)} \leq M_2, \end{aligned}$$

y con (3.2.7) se obtiene

$$g\frac{1}{\lambda} \leq V(\pi, x).$$

La segunda parte de la prueba se obtiene cambiando la desigualdad en la ecuación (3.2.7) por igualdad con argumentos análogos los anteriores. □

Por lo tanto la ecuación de costo promedio óptimo sensible al riesgo están dado por la ecuación (3.2.6), escrito en términos de la función de utilidad es

$$U_\lambda(g + h(x)) = \min_{a \in A(x)} \left[\sum_{x \in X} p(y|x, a) U_\lambda(c(x, a) + h(y)) \right]. \tag{3.2.9}$$

Otra condición necesaria para garantizar la existencia de una solución acotada de la ecuación de costo óptimo es que se cumple la condición de Doeblin (véase [2], [10], [6]).

Suposición 3.2.2 (Condición de Doeblin)**.** *Existe un estado $z \in X$ y una constante K tal que para cualquier estado $x \in X$ y para toda política estacionaria $f \in \mathbb{F}$ se cumple*

$$E_x^f[T_z] \leq K,$$

donde, para cada $y \in X$

$$T_y := \min\{n \geq 1 | X_n = y\}.$$

Se ha mostrado que bajo las Suposiciones 3.2.1 y 3.2.2 implica la existencia de de una solución acotada de la ecuación de costo promedio óptimo para λ cercano a cero (véase [5]); mientras que, para resolver esta ecuación para cualquier λ distinto de cero es necesario tener una condición de comunicación (véase [6]).

Suposición 3.2.3 (Comunicación)**.** *Bajo cualquier política estacionaria cada pareja de estados es comunicante, es decir, dada $f \in \mathbb{F}$, $x, y \in X$ existe un $n = n(x, y, f) \in \mathbb{N}$ tal que $P_x^f(X_n = y) > 0$.*

Para resolver el problema de control óptimo será usada la técnica de programación dinámica.

3.3. Caracterización por Programación Dinámica.

En esta sección se presentará el teorema fundamental de este capítulo el cual establece que cuando el espacio de estados es una clase comunicante bajo la acción de cada política estacionaria entonces existe una solución de la ecuación de costo promedio óptimo para cada $\lambda > 0$. Antes de enunciar el teorema se presentarán dos resultados auxiliares.

Lema 3.3.1. *Suponga que se tiene una cadena comunicante (Suposición 3.2.3). Sea $f \in \mathbb{F}$ fijo y suponga que $\mathcal{A}$ es un subconjunto no vacío del espacio de estados que cumple la siguiente propiedad:*

$$x \in \mathcal{A} \Rightarrow y \in \mathcal{A} \quad si \quad p(y|x, f(x)) > 0.$$

entonces $\mathcal{A} = X$.

Demostración. Sea $y \in X$ un estado arbitrario y $z \in \mathcal{A}$; ya que se tiene una clase comunicante existe un estado x_i, $i = 1, 2, \ldots, n-1$ tal que $x_0 = z$ y $x_n = y$, además $p(x_{i+1}|x_i, f(x_i)) > 0$ para $i = 1, 2, \ldots, n-1$.

Como $x_i \in \mathcal{A}$ por hipótesis se cumple que $x_{i+1} \in \mathcal{A}$, por lo tanto, para $x_0 = z \in \mathcal{A}$ se cumple que $x_n = y \in \mathcal{A}$ ya que y es arbitrario se cumple $\mathcal{A} = X$. □

El Teorema 3.3.1 enunciado más adelante se basa en el llamado *operador contracción* en el contexto de sensibilidad al riesgo es usando el enfoque de descuento desvaneciente análogo a la sección 2.3. Para cada $\alpha \in (0, 1)$; $T_\alpha : \mathcal{B}(X) \to \mathcal{B}(X)$, dada una función $W \in \mathcal{B}(X)$, $T_\alpha W$ es definida por

$$U_\lambda(T_\alpha W(x)) = \min_{a \in A(x)} \left[\sum_{y \in S} p(y|x, a) U_\lambda(c(x, a) + \alpha W(y)) \right], \tag{3.3.1}$$

algunas propiedades del operador T_α son presentadas a continuación.

Lema 3.3.2. I) *T_α es un operador contracción.*

II) *Existe una función única $V_\alpha \in \mathcal{B}(X)$ tal que $T_\alpha V_\alpha = V_\alpha$.*

III) $||(1-\alpha)V_\alpha|| \leq ||c||$.

Demostración. I) Sean $V, W \in \mathcal{B}(X)$, la desigualdad

$$c(x,a) + \alpha V(y) \leq c(x,a) + \alpha W(y) + \alpha||V - W||$$

siempre es válida. Aplicando la función de utilidad

$$U_\lambda(c(x,a) + \alpha V(y)) \leq e^{\alpha||V-W||} U_\lambda(c(x,a) + \alpha W(y)),$$

por tanto, (3.2.9) se cumple para cada estado x;

$$\begin{aligned} U_\lambda(T_\alpha W(x)) &= \min_{a\in A(x)} \left[\sum_{y\in S} p(y|x,a) U_\lambda(c(x,a) + \alpha V(y)) \right] \\ &\leq e^{\alpha||V-W||} \min_{a\in A(x)} \left[\sum_{y\in S} p(y|x,a) U_\lambda(c(x,a) + \alpha W(y)) \right] \\ &= e^{\alpha||V-W||} U_\lambda(T_\alpha W(x)), \end{aligned}$$

tomando logaritmo $T_\alpha V(x) \leq T_\alpha W(x)(x) + \alpha||V - W||$, intercambiando V por W se obtiene $|T_\alpha V(x) - T_\alpha W(x)| \leq \alpha||V - W||$ para cada $x \in X$, por lo tanto se tiene I).

II) La existencia de un único punto fijo está dada por el teorema del punto fijo de Banach (véase [16]).

III) Observe que para todo $x, y \in X$ y $a \in A$

$$-||c|| - \alpha||V_\alpha|| \leq c(x,a) + \alpha(y) \leq ||c|| + \alpha||V_\alpha||,$$

así

$$U_\lambda(-||c|| - \alpha||V_\alpha||) \leq \sum_{y\in S} p(y|x,a) U_\lambda(c(x,a) + \alpha V(y)) \leq U_\lambda(||c|| + \alpha||V_\alpha||),$$

tomado mínimo sobre todas las acciones en A y sustituyendo el punto fijo en II) en la ecuación (3.2.9) se obtiene

$$U_\lambda(-||c|| - \alpha||V_\alpha||) \leq U_\lambda(V_\lambda(x)) \leq U_\lambda(||c|| + \alpha||V_\alpha||),$$

así $|V_\alpha(x)| \leq ||c|| + \alpha||V_\alpha||$, como x es arbitrario, se sigue que $||V_\alpha|| \leq ||c|| + \alpha||V_\alpha||$ por lo tanto se cumple III).

□

Teorema 3.3.1. *Suponga que cada política estacionaria induce una cadena de Markov para la que el espacio de estados es una clase comunicante, es decir, se cumple la Suposición 3.2.3. Entonces para cada $\lambda > 0$, existe una pareja $(g, h(\cdot))$, la cual satisface la ecuación de costo promedio óptima (3.2.9).*

Demostración. De la ecuación (3.3.1) y suponiendo que que V_α es el punto fijo de T_α se tiene que

$$U_\lambda(V_\alpha(x)) = \min_{a\in A(x)} \left[\sum_{y\in X} p(y|x,a) U_\lambda(c(x,a) + \alpha V_\alpha(y)) \right], \tag{3.3.2}$$

ya que el espacio de estados X es finito entonces, para cada $\alpha \in (0,1)$ existe un estado $z_\alpha \in X$ tal que

$$V_\alpha(z_\alpha) = \min V_\alpha(x),$$

definiendo

$$\begin{aligned} g_\alpha &:= (1-\alpha) V_\alpha(z_\alpha), & (3.3.3)\\ h_\alpha &:= V_\alpha(x) - V_\alpha(z_\alpha), & (3.3.4) \end{aligned}$$

con esta definición $h_\alpha(\cdot) > 0$, ahora, reescribiendo la ecuación (3.3.2)

$$U_\lambda(g_\alpha + h_\alpha(x)) = \min_{a\in A(x)} \left[\sum_{y\in X} p(y|x,a) U_\lambda(c(x,a) + \alpha h_\alpha(y)) \right], \tag{3.3.5}$$

de la continuidad de p_{xy} (de la Suposición 3.2.1) sobre el espacio de acciones implica que existe una política estacionaria $f_\alpha \in \mathbb{F}$ tal que

$$U_\lambda(g_\alpha + h_\alpha(x)) = \sum_{y\in X} p(y|x,f_\alpha) U_\lambda(c(x,f_\alpha) + \alpha h_\alpha(y)), \tag{3.3.6}$$

del Lema 3.3.2 III) y de las ecuaciones (3.3.3) y (3.3.4) se tiene que $|g| \leq ||c||$.

Ahora, sea $\{\alpha_n\}$ una sucesión en el intervalo $(0,1)$, creciente y convergente a uno, ya que el espacio de estados es finito y por compacidad del espacio

de acciones es posible elegir una subsucesión de $\{\alpha_n\}$, que seguirá siendo denotada por $\{\alpha_n\}$ tal que

$$\begin{aligned} f_{\alpha_n} &:= f \in \mathbb{F}, n \in \mathbb{N} \\ z_{\alpha_n} &:= z \in X, n \in \mathbb{N} \end{aligned} \tag{3.3.7}$$

y existen los límites

$$\lim_{n\to\infty} g_{\alpha_n} = g, \tag{3.3.8}$$

$$\lim_{n\to\infty} h_{\alpha_n}(x) = h(x) \in [0,\infty] \tag{3.3.9}$$

Resta demostrar que la función $h(\cdot)$ así definida es finita y además que la pareja $(g, h(\cdot))$ satisface la ecuación de costo promedio óptimo (3.2.9).

Definiendo al conjunto $\mathcal{A} := \{x | h(x) < \infty\}$ note que el estado z en (3.3.7) pertenece a este conjunto, pues

$$\begin{aligned} h(z) &= \lim_{n\to\infty} h_{\alpha_n}(z), \\ &= \lim_{n\to\infty} [V_{\alpha_n}(z) - V_{\alpha_n}(z_\alpha)], \\ &= \lim_{n\to\infty} [V_{\alpha_n}(z) - V_{\alpha_n}(z)] = 0. \end{aligned}$$

sea $x \in X$ tal que $x \in \mathcal{A}$, reemplazando α por α_n en (3.3.6) y haciendo a n tender a infinito

$$\begin{aligned} \lim_{n\to\infty} U_\lambda(g_{\alpha_n} + h_{\alpha_n}(x)) &= \lim_{n\to\infty} \left[\sum_{y\in X} p(y|x, f_{\alpha_n}(x)) U_\lambda(c(x, f_{\alpha_n}(x)) + \alpha_n h_{\alpha_n}(y)) \right], \\ \infty > U_\lambda(g + h(x)) &= \sum_{y\in X} p(y|x, f(x)) U_\lambda(c(x, f(x)) + h(y)) \end{aligned} \tag{3.3.10}$$

así que, $h(y) < \infty$ cuando $p(y|x, f(x)) > 0$. Por lo tanto, $y \in \mathcal{A}$, del Lema 3.3.1 se tiene que $\mathcal{A} = X$, es decir, $h(\cdot)$ es finita. Por otro lado, de la ecuación (3.3.5) se tiene que para todo $x \in X$, $a \in A$, $n \in \mathbb{N}$ y tomando límite se cumple

$$\begin{aligned} \lim_{n\to\infty} U_\lambda(g_{\alpha_n} + h_{\alpha_n}(x)) &\leq \lim_{n\to\infty}\left[\sum_{y\in X} p(y|x,a)U_\lambda(c(x,a)+\alpha_n h_{\alpha_n}(y))\right], \\ U_\lambda(g+h(x)) &\leq \sum_{y\in X} p(y|x,a)U_\lambda(c(x,a)+h(y)). \end{aligned} \tag{3.3.11}$$

De las ecuaciones (3.3.10), (3.3.11) y ya que la pareja $(x,a) \in X \times A$ es arbitraria, se tiene que $(h(\cdot), g)$ definidas en (3.3.8) y (3.3.9) satisface la ecuación de costo promedio óptimo. □

3.4. Ejemplos.

En esta sección se presentan ejemplos aplicando el Teorema 3.3.1.

Ejemplo 3.4.1. *Sea $X = \{s_1, s_2\}$ con $A(s_i) = \{a_1, a_2\}$, $i = 1, 2$ con función de costo $c(s_i, a_1) = c_1$, $c(s_i, a_2) = c_2$, $i = 1, 2$ donde $0 < c_1 < c_2$ y $\alpha \in (0, 1)$.*

La matriz de transición bajo la acción a_1 está dada por

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \tag{3.4.1}$$

mientras que bajo la acción a_2 está dada por

$$\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \tag{3.4.2}$$

Dado que el espacio de estados y el de acciones es finita se cumple la Suposición 3.2.1, para que se cumpla la condición de Doeblin (Suposición 3.2.2) y la Suposición 3.2.3 se supondrá que $p_{ij} \neq 0$ y $q_{ij} \neq 0$, $i, j = 1, 2$. Así que es posible aplicar programación dinámica.

Tomando $V_0(s_i) = 0$, $i = 1, 2$ se obtiene

$$\begin{aligned} e^{\lambda V_1(s_1)} &= \min_{a\in A(s_1)} \left\{ \sum_y p(y|s_1,a) e^{\lambda(c(s_1,a)+\alpha V_0(y))} \right\}, \\ &= \min \left\{ \sum_y p(y|s_1,a_1) e^{\lambda c_1}, \sum_y p(y|s_1,a_1) e^{\lambda c_2} \right\}, \\ &= e^{\lambda c_1}, \end{aligned}$$

así que $V_1(s_1) = c_1$.

Análogamente $V_1(s_2) = c_1$.

Así $V_2(s_1) = V_2(s_2) = c_1(1+\alpha)$

En general se tiene

$$V_n(s_i) = c_1 \sum_{j=0}^{n-1} \alpha^j, \quad i = 1, 2$$

tomando el límite cuando n tiende a infinito

$$V_n(s_i) = \frac{c_1}{1-\alpha},$$

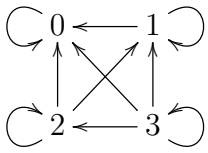
entonces

$$\begin{aligned} V_\alpha(z_\alpha) &= \min_{x\in X} V(x) = \frac{c_1}{1-\alpha} \\ g_\alpha &= (1-\alpha) V_\alpha(z_\alpha) = c_1 \\ h_\alpha &= c_1. \end{aligned}$$

Ejemplo 3.4.2 (Inventarios)**.** *Recuerde el ejemplo de inventarios dado en los Ejemplos 1.5.1 y 2.5.2, los datos dados son* $K = 5$, $c(x) = 10x$, $g(x) = 0$, $h(x) = 9x$, $M = 3$, $N = 4$ *y la matriz de transición está dada por*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 5/6 & 1/6 & 0 & 0 \\ 1/2 & 1/3 & 1/6 & 0 \\ 1/6 & 1/3 & 1/3 & 1/6 \end{bmatrix}$$

En este caso se tiene un estado absorbente, 0 *y tres estados transitorios,* $1, 2, 3$ *por lo tanto no se cumple la Suposición de comunicación 3.2.3 por lo tanto con esta matriz no es posible resolver el problema de costo promedio sensible al riesgo. Bajo las hipótesis estudiadas en este trabajo.*



Sin embargo existe otro modelo de inventarios que se presenta en el Ejemplo 3.4.3 con el que es posible resolver dicho problema.

Ejemplo 3.4.3 (Inventarios (s, S))**.** *Suponga que se tiene en un almacén cierto número de un producto, y una demanda aleatoria* ξ *del producto en el periodo* n*, suponga que* $P(\xi_n = y) = p_y$ *para* $y = 0, 1, \ldots, S$ *tal que* $p_y \geq 0$ *y* $\sum_y p_y = 1$*. El número de bienes almacenados es revisado en cada periodo aplicando la siguiente política de reabastecimiento: Si al final de cada periodo la cantidad del bien es menor o igual a un nivel* s*, entonces la bodega se reabastece hasta un nivel máximo* S*, si al final del periodo el nivel del inventario es mayor a* s*, entonces no hay reabastecimiento (véase [15]).*

Sea x_n *el nivel del inventario en el periodo* n *antes de reabastecer. En este caso no se permitirán valores negativos para* x_n*, los cuales, representan demandas del producto no satisfechas así que el espacio de estados es* $X = \{0, 1, 2, \ldots, S\}$*. El modelo es presentado a continuación*

$$x_{t+1} = \begin{cases} (x_t + \xi_t)^+ & , \quad s < x_t \leq S, \\ (S - \xi_t)^+ & , \quad x_t \leq s. \end{cases}$$

La probabilidad de transición está dada por

$$Q(y|x) = \begin{cases} P((x_t + \xi_t)^+ = y) & , \quad s < x_t \leq S, \\ P((S - \xi_t)^+ = y) & , \quad x_t \leq s. \end{cases}$$

La primera expresión representa el caso en el que no hay reabastecimiento, la probabilidad de pasar de x a y es la probabilidad de que la demanda al final de ese periodo sea de $y - x$ pues el nuevo nivel en el almacén será de $x - (x - y) = y$. La siguiente ecuación representa el caso cuando hay reabastecimiento y el nuevo nivel es de $S - (S - y) = y$, cuando la demanda sea de $S - y$.

A continuación se presenta un ejemplo numérico de este modelo. Se tomará $s = 1$ y el nivel máximo del inventario será $S = 3$, es decir, $X = \{0, 1, 2, 3\}$ la distribución de la demanda $P(\xi_n = y) = 1/4$, calculando la matriz de transición se obtiene

$$\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

y suponga que se tiene un costo constante $c(x) = b$ para cada estado con $b \geq 0$.

Ya que el espacio de estados es finito, la Suposición 3.2.1, también se cumple la condición de Doeblin (Suposición 3.2.2) y además se tiene una cadena comunicante, es decir, se cumple la Suposición 3.2.3. Por lo tanto es posible aplicar el algoritmo de programación dinámica.

Tomando $V_0(x) = 0$, $x \in X$ se obtiene

$$\begin{aligned} e^{\lambda V_1(x)} &= \sum_y p(y|x) e^{\lambda(c(x)+\alpha V_0(y))}, \\ &= \sum_y p(y|x) e^{\lambda b} \\ &= e^{\lambda b}, \end{aligned}$$

así que $V_1(s_1) = b$.

De la misma forma se obtiene $V_2(x) = b(1+\alpha)$.

En general se tiene

$$V_n(x) = b\sum_{i=0}^{n-1}\alpha^i, \quad x \in X.$$

tomando el límite cuando n tiende a infinito

$$V_n(x) = \frac{b}{1-\alpha},$$

entonces

$$\begin{aligned} V_\alpha(z_\alpha) &= \min_{x\in X} V(s) = \frac{b}{1-\alpha} \\ g_\alpha &= (1-\alpha)V_\alpha(z_\alpha) = b \\ h_\alpha &= b. \end{aligned}$$

Conclusiones

En este trabajo de tesis se estudió parte de la teoría de Procesos de Decisión de Markov, en particular fue expuesto el problema de control óptimo con los criterios de costo total, costo promedio y costo promedio sensible al riesgo. Este problema fue resuelto usando la técnica de programación dinámica.

En los casos de costo decontado y costo promedio neutral al riesgo se requieren condiciones referentes a la función de costo como por ejemplo que sea semicontinua inferiormente, no negativa e inf-compacta para hallar una solución al problema de control óptimo con este criterio. En este caso fue presentado un ejemplo aplicado a la teoría de inventarios y el problema de tipo lineal cuadrático.

Para resolver el problema de control óptimo con el criterio de costo promedio se empleó la terna (ρ, h, f) conocida como tripleta canónica (véase [11]), la cual proporciona una solución a la n-ésima función de valor. Además se introdujo el concepto de política fuertemente RP-óptima y el de F-fuertemente RP-óptima así como la relación con la tripleta canónica, también, fue presentada la desigualdad de costo promedio óptima. En el caso de costo promedio se enunciaron condiciones adicionales a las dadas con el criterio de costo total con las que es posible garantizar la existencia de políticas óptimas utilizando el factor de descuento desvaneciente basado en el criterio de costo descontada con factor de descuento $\alpha \in (0, 1)$ (véase [11]) y su relación con el costo promedio dada por la ecuación de optimalidad de costo descontada.

Respecto al caso del costo promedio sensible al riesgo un concepto fundamental es el de la función de utilidad y la relación que existe con la certeza equivalente (véase [2]). En este caso un supuesto es que el espacio de ac-

ciones admisibles es compacto, además, se mantiene la suposición de que la función de costo es no negativ y continua. Fue necesario validar la técnica de programación dinámica presentado como el Teorema de verificación 3.2.1 (véase [6]). El procedimiento para garantizar la existencia de una solución fue similar al presentado con el criterio de costo promedio -por medio del costo descontado- para lo que se empleó el operador de contracción. Otra condición importante es la condición de Doeblin (véase [2], [10]) así como la suposición de comunicación de la cadena (véase [6]) la cual resulta ser fuerte pues en el Ejemplo 3.4.2 de inventarios no fue posible trabajar con los mismos datos empleados en el criterio de costo total y costo promedio ya que la matriz de transición no satisface esta característica.

Un posible trabajo a futuro consiste en estudiar alternativas a la suposición de comunicación con la que sea posible garantizar la existencia de una solución acotada de la ecuación de optimalidad. Además es necesario extender las condiciones del Teorema 1.3.1 de Programación Dinámica y el Lema 2.4.1 para permitir funciones de costo con signo variante.

Apéndice A

Resultados Auxiliares.

Sean X, A y Y espacios de Borel.

Definición A.0.1. *Un kérnel estocástico es una función $P(\cdot|\cdot)$ tal que*

(a) $P(\cdot|y)$ es una medida de probabilidad sobre X para cada $y \in Y$.

(b) $P(B|\cdot)$ es una función medible sobre Y para cada $B \in \mathcal{B}$.

Definición A.0.2. *Sea $f : \mathbb{K} \to \mathbb{R}$ una función, se dice que es inf-compacta sobre $\mathbb{K}$ si para toda $x \in X$ y $r \in \mathbb{R}$, el conjunto $\{a \in A(x)|v(x,a) \leq r\}$ es compacto donde $A(x) \subseteq A$.*

Definición A.0.3. *Sea X un espacio métrico y v una función $v : X \to \bar{\mathbb{R}}$ tal que $v(x) < \infty$ para al menos un $x \in X$. Se dice que v es semicontinua superiormente (u.s.c) en x si para cualquier sucesión $\{x_n\}$ en X que converge a x se cumple*

$$\limsup_{n\to\infty} v(x_n) \geq v(x),$$

se dice que que v es semicontinua superiormente si es u.s.c. en cada punto de X.

Definición A.0.4. *Un kérnel estocástico $P \in Q(X|A)$ es fuertemente continuo si la función*

$$v'(a) := \int_X v(x)P(dx|a),$$

es continua y acotada en A para cada función medible y compacta sobre X.

Teorema A.0.1 (Teorema de la Convergencia Monónota)**.** *Sean* $(X, \mathcal{P}(X), \mathbf{P})$ *un espacio de probabilidad y* $\{f_n\}$ *una sucesión de funciones medibles en* X *y suponga que para cada* $x \in X$ *se cumple*

a) $0 \leq f_1(x) \leq f_2(x) \leq \cdots \leq \infty$.

b) $f_n(x) \longrightarrow f(x)$ *cuando* $n \to \infty$.

Entonces f *es medible y*

$$\int_X f_n d\mathbf{P} \longrightarrow \int_X f d\mathbf{P} \quad n \to \infty$$

Lema A.0.1 (Lema de Fatou)**.** *Sean* $(X, \mathcal{P}(X), \mathbf{P})$ *un espacio de probabilidad y suponga que* $f_n : X \to [0, +\infty]$ *es medible para cada entero positivo* n *entonces*

$$\int_X \left(\liminf_{n\to\infty} f_n \right) d\mathbf{P} \leq \liminf_{n\to\infty} \int_X f_n d\mathbf{P}$$

Lema A.0.2. *Sean* u *y* u_n, $n = 1, 2, \ldots$ *funciones semicontinuas superiormente, acotadas superiormente e sup-compactas sobre* $\mathbb{K}$, *si* $u_n \uparrow u$ *entonces*

$$\lim_{n\to\infty} \max_{A(x)} u_n(x, a) = \max_{A(x)} u(x, a).$$

Lema A.0.3. *Sea* $\{s_t | t = 0, 1, \ldots\}$ *una sucesión de números no negativos, entonces*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} s_t \leq \liminf_{\alpha\to 1} (1-\alpha) \sum_{t=0}^{\infty} \alpha^t s_t \leq \limsup_{\alpha\to 1} (1-\alpha) \sum_{t=0}^{\infty} \alpha^t s_t \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} s_t.$$

para la demostración véase [21] y [18].

Bibliografía

[1] Arriaga, M. *Problemas de control de Markov con recompensa total esperada en espacios finitos*, Tesis de Maestría Depto matemáticas. Universidad Autónoma Metropolitana-Iztapalapa (2008).

[2] Alanís-Durán, A., *An Optimality System for the Risk-Sensitive Average Cost Criterion in Markov Decision Chains on a Finite State Space*, Tesis Doctoral, Universidad Autónoma de Nuevo León Facultad de Ciencias Físico-matemáticas (2013).

[3] Ash Robert B. *Probability and Measure Theory*, UK, Harcourt Academic Press (2000).

[4] Bellman R, Dynamic Programming. Dover (2003).

[5] Cavazos-Cadena, R., Fernández-Gaucherand. *Controlled Markov Chains with Risk-Sensitive Criteria: Average Cost, Optimality Equations and Optimal Solutions*, Mathematics Methods of Operations Research, Vol. 49: 299-324 (1998).

[6] Cavazos-Cadena, R., Hernández-Hernández, D. *Solution to the risk-sensitive average optimality equation in communicating Markov decision chains with finite state space: An alternatve approach.*, Mathematics Methods of Operations Research, Vol. 56: 4473-479 (2002).

[7] Flynn, J. *On optimality criteria for dynamic programming with long finite horizon* J. Math. Anal. Appl. 76, 202-208 (1980).

[8] Bäuerle, N., Rieder U. *More risk-sensitive Markov decision processes*, Mathematics Methods of Operations Research, vol 39 no.1, Pp. 105-120 (2013).

[9] Ephremides, A. and Verdu, S. *Control and optimization methods in communication networks.* IEEE Trans. Autom. Control 34, 930-942 (1989).

[10] Hernández-Hernández, D., Steven M. *Risk sensitive control of Markov processes in contable state space* Systems & Control letters 29 Pp.147-155 (1996).

[11] Hernandez-Lerma, O., Muñoz de Ozak, M. *Discrete-time MCPs with discounted unbounded costs: Optimality criteria.* Kybernetika, Prague 28, 191-212 (1992).

[12] Hernández-Lerma, O., Lasserre, J. B., *Discrete-Time Markov Control Processes Basic Optimality Criteria*, Springer (1996).

[13] Howard, R., Mathenson J., *Risk-sensitive Markov decision processes*, Management science, vol 18, no. 7, 356-369.

[14] Puterman, M. L. *Markov decision processes, discrete stochastic dynamic programming* (1994).

[15] Rincón, L. *Introducción a los Procesos Estocásticos* Departamento de Matemáticas Facultad de Ciencias UNAM, México (2011).

[16] Rudin, W. *Real and Complex Analysis* McGraw Hill (1970).

[17] Sladky, K., *Growth rates and average optimality in risk-sensitive Markov decision chains*, Kybernetica, vol 4, no 2, 205-226 (2008).

[18] Sznajder, R and Filar, J.A. *comments on a theorem of Hardy and Littlewood. J. Optim. Theory Appl* 75, 201-208 (1992).

[19] López E., Barea R., Escudero M. S., *Navegación topológica mediante POMDPs incorporando información visual. Departamento de Electrónica*, Universidad de Alcalá, (2003).

[20] Von Neumann, J., Morgenstern, O. *Theory of games economic behavior*, Princeton University Press, Princeton, NJ (1947).

[21] Widder, D.V. *The Laplace Transform* Princeton University Press, Princeton,NJ (1941).

[22] Zacarías, G. *Procesos de Decisión de Markov Descontados* Tesis de Licenciatura Facultad de Cs. Fisico-Matemáticas. BUAP (2007).