Benemérita Universidad Autónoma de Puebla

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
POSGRADO EN CIENCIAS MATEMÁTICAS

Procesos de Decisión de Markov bajo el Criterio de Entropía Relativa

TESIS
PARA OBTENER EL GRADO DE
Doctor en Ciencias (Matemáticas)

PRESENTA
Gladys Denisse Salgado Suárez

DIRECTORES DE TESIS
Dr. Hugo Adán Cruz Suárez
Dr. José Dionicio Zacarías Flores
Dr. Fernando Velasco Luna

PUEBLA, PUE. NOVIEMBRE 2019

Introducción

El tema principal de este trabajo de tesis es Procesos de Decisión de Markov (PDMs). Un PDM es aquél que modela, mediante un Modelo de Control de Markov, un sistema observado en el tiempo por un controlador o agente decisor que influye en la evolución del sistema. El controlador decide la acción (control) a tomar dependiendo del estado actual con el objetivo de que el sistema se desempeñe eficazmente con respecto a cierto criterio de optimalidad (función objetivo o criterio de rendimiento). La acción genera un costo (o recompensa) a pagarse y repercute en el nuevo estado del sistema de acuerdo con una distribución de probabilidad preestablecida. Este procedimiento de selección se repite de manera periódica hasta cierto momento dado llamado horizonte del problema, a la sucesión de acciones determinada se le denomina política. La mejor política será aquella que optimice el criterio de optimalidad, lo cual da origen al problema de control óptimo [18], [30].

El propósito principal de esta tesis es abordar los siguientes dos problemas relacionados con PDMs:

a) El Problema de la Ruta más Corta.

b) El Análisis Asintótico de un Sistema de Control Determinista.

Desde sus inicios, el estudio de problemas relacionados con PDMs ha sido relevante por la diversidad de aplicaciones a las que da solución, por ejemplo en Inteligencia Artificial [42], Finanzas [39], Comunicaciones [1], Administración de Recursos Naturales [27], Ingeniería [21], etc. Las primeras investigaciones se centraron en el desarrollo de los planteamientos y métodos de solución, uno de estos métodos es llamado Programación Dinámica [13], por lo tanto, dicho método será la herramienta principal para resolver los problemas de esta tesis.

El problema clásico de la Ruta más corta estocástica ([2], [4], [5]) comprende un grafo dirigido con un número finito de nodos, donde cada arco tiene una longitud (o recompensa) asignada. El problema implica seleccionar una distribución de probabilidad para cada nodo sobre todos los nodos sucesores posibles. De esta manera, los arcos se forman de un nodo a otro hasta alcanzar un nodo específico, llamado terminal. El objetivo es llegar al nodo terminal con la longitud mínima esperada (o la mayor recompensa). Por lo tanto, de acuerdo a sus características, el problema antes mencionado puede ser identificado como un PDM.

Dentro del estudio de estos problemas a través de un PDM se ha incluido una restricción sobre los costos de información, definido a través de la entropía relativa, concepto dado por la Teoría de la información. La entropía relativa se ha trabajado, en espacios de estados finitos, como una medida del flujo de información generado entre el sistema y el ambiente ([16], [17], [28], [33], [40]), o bien, como elemento importante dentro del control óptimo inverso (o aprendizaje de refuerzo inverso, IRL) en el marco de PDMs linealmente solubles [26].

En el contexto de la teoría de información, los costos de información son generados debido a que están inmersos dos sujetos, el decisor (controlador) y el actuador. El decisor controla y envía el mensaje, mientras que el actuador recibe el mensaje y lleva a cabo la acción. Si ahora, se consideran escenarios donde el controlador y el actuador están separados por algún canal de comunicación, el envío de información a través del canal no es gratuito pues la información enviada por el agente puede llegar incompleta en la etapa siguiente. Por lo tanto, los agentes al decidir la acción a tomar deben considerar no solo la recompensa (o costo) que se genere sino también los costos de comunicación, para ello, es importante incluir en el sistema un segundo criterio que los contemple.

Con base en los estudios anteriores, la primera aportación de este trabajo es, proponer condiciones para obtener una solución al problema de la ruta más corta estocástica con las siguientes características: El espacio de estados es infinito numerable, el criterio de rendimiento es la recompensa total esperada y se incluye una restricción en los costos de información.

Para ello, se consideran recompensas negativas (costos), de manera que cuando se combinan con los costos de información, los arcos se generan con una longitud positiva. Además, es necesaria una condición que garantice la conexión de cada nodo con el nodo terminal. Esta condición es contemplada cuando la solución se busca dentro del conjunto de las llamadas políticas propias [2], las cuales tienen la característica de garantizar que se alcance el estado terminal casi seguramente después de un número finito de pasos, independientemente del estado inicial. Posteriormente, proponemos un método basado en un PDM con el costo total esperado como criterio de rendimiento, donde la función de costo refleja el equilibrio entre las recompensas esperadas (valor) y los costos de información. Para obtener el valor y la política óptimos, demostramos que la función de costo óptimo es la única solución de la ecuación de programación dinámica y que el método de aproximaciones sucesivas converge a la función de valor óptimo.

Presentamos también, ejemplos numéricos del problema de la ruta más corta, determinista y estocástica para el caso de nodos numerables e incluyendo costos de comunicación, además, desarrollamos el ejemplo del mundo de la

rejilla con celdas numerables y subconjuntos de $\mathbb{R}$, como introducción a una posible extensión del espacio de estados [2], [4], [16], [17], [28], [33], [40].

Con respecto al segundo problema, se tiene un Sistema de Control Determinista (SCD), el cual es usado para modelar sistemas dinámicos, en donde el nuevo estado puede ser determinado con base en la acción que se tome y al estado actual, a través de una ecuación en diferencias que no posee un ruido aleatorio. Estas clases de problemas se incluyen en la teoría de los procesos de decisión de Markov ([18], [20]).

Cuando la ecuación en diferencias está perturbada por un ruido aleatorio, llamado problema estocástico, importantes investigaciones han proporcionado una solución a través del sistema determinista asociado ([14], [24], [44]), sin embargo, debe poder garantizarse y caracterizarse el estado de convergencia de su trayectoria óptima. Es por ello, que la convergencia de la trayectoria óptima del SCD resulta de interés.

Al estudiar la convergencia de la trayectoria óptima, se ha encontrado para ejemplos particulares que el punto de equilibrio es el estado límite. El punto de equilibrio del sistema, se entiende como el estado en el que la trayectoria se estabiliza y por ende no existe variación. En términos económicos, el punto de equilibrio sería aquel en el que la economía se vuelve estable y se alcanzarían las mayores utilidades [12]. Una herramienta básica para poder caracterizarlo ha sido la ecuación de Euler.

Es por ello que en este trabajo se buscaron las condiciones necesarias para garantizar la existencia del punto de equilibrio del sistema, así como la convergencia a dicho punto para SCD generales, demostrando que la trayectoria óptima es monótona y acotada, además de caracteriza el punto de equilibrio a través de la ecuación de Euler.

Se plantea además el ejemplo de la Función de Utilidad Logarítmica que describe la producción de una economía mediante una ecuación en diferencias, la cual muestra la teoría determinista desarrollada.

La redacción del presente trabajo se realiza de la forma siguiente: El Capítulo 1 presenta la teoría básica de PDMs, Programación Dinámica y los conceptos principales de Teoría de la Información. En los Capítulos 2 y 3 se abordan el problema de control con el criterio de energía libre y el Análisis asintótico de un Sistema de Control Determinista, respectivamente, para cada uno de ellos, se realiza una revisión de los antecedentes, se plantea formalmente el problema de control, se describe la metodología que se sigue y se expone la solución, así como ejemplos que muestran la aplicación de la teoría trabajada.

Finalmente, se presentan las conclusiones derivadas de este trabajo.

Índice general

Capítulo 1

Preliminares

Para resolver los problemas abordados en esta tesis, la teoría de Procesos de Decisión de Markov es una herramienta adecuada, es por ello que retomamos de [18] y [30] los conceptos principales de esta área así como de la teoría principal de Programación Dinámica. Incluimos también los elementos esenciales de la Teoría de la Información a los cuales se hará referencia en el resto de la tesis.

1.1. Procesos de Decisión de Markov

Un **Proceso de Decisión de Markov** (PDM) modela un sistema observado en el tiempo por un **controlador** o agente decisor que influencía en la evolución del sistema. El controlador decide qué **acción** (control) tomar dependiendo del estado en el que está con el objetivo de que el sistema se desempeñe eficazmente con respecto a ciertos **criterios de optimalidad** (función objetivo). La acción genera un costo (o recompensa) que debe pagarse y repercute en el nuevo estado del sistema de acuerdo a una distribución de probabilidad dada. Para cada estado del sistema debe especificarse una **regla de decisión** que indique qué acción tomar, para ello se establece una política o estrategia. La mejor **política** será la que optimice el criterio de optimalidad lo que da origen al **Problema de Control Óptimo**.

Un Modelo de Control de Markov (MCM) estacionario, consiste de la quíntupla:

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

donde:

- X es un espacio de Borel no vacío, llamado espacio de estados;
- A es un espacio de Borel no vacío, llamado conjunto de acciones o controles;

- $\{A(x)|x \in X\}$ es una familia de subconjuntos medibles, no vacíos $A(x)$ de A, donde $A(x)$ denota el conjunto de controles admisibles cuando el sistema se encuentra en el estado $x \in X$. El conjunto $\mathbb{K}$ de parejas de estados acciones-admisibles, está definido por

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

  y se supone que es un conjunto medible del espacio producto $X \times A$;

- Q es un kérnel estocástico definido en X dado $\mathbb{K}$, llamado la 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 $B \in \mathcal{B}(X)$, donde $\mathcal{B}(X)$ denota la σ-álgebra de Borel de X, $Q(B|\cdot)$ es una función medible;

- $r : \mathbb{K} \longrightarrow \mathbb{R}$ es una función medible y se llama la función de recompensa en un paso.

Observación 1.1.1. *En algunos contextos en lugar de una función de recompensa r, es más conveniente considerar una función de costo $c : \mathbb{K} \to \mathbb{R}$.*

Cuando se trabaja con un MCM a tiempo discreto y estacionario, la interpretación del modelo es la siguiente: en cada tiempo $t \in \mathbb{N}$ se observa el estado del sistema $x_t = x \in X$, se aplica una acción $a_t = a \in A(x)$, y como resultado ocurren dos cosas:

- se obtiene una recompensa $r(x, a)$,

- el sistema se traslada a un nuevo estado x_{t+1}, mediante la distribución de probabilidad $Q(\cdot \mid x, a)$ sobre X, es decir,

$$Q(B \mid x, a) = P(x_{t+1} \in B \mid x_t = x, a_t = a).$$

Una vez hecha esta transición a un nuevo estado, se elige una nueva acción y la dinámica anteriormente descrita se repite.

Una *política* es una sucesión de funciones llamadas reglas de decisión, que seleccionan una acción dada la historia del proceso, las reglas de decisión se clasifican en markovianas o dependientes de la historia dependiendo de cómo se incorpora la información pasada y aleatorias o deterministas según cómo se seleccionan las acciones. Para definir una política se introduce primero la noción de *historia*.

Para cada $t \in \mathbb{N}$, se define $\mathbb{H}_t$ el espacio de historias admisibles al tiempo t como $\mathbb{H}_0 := X$ y $\mathbb{H}_t := X \times \mathbb{H}_{t-1}$ para $t \geq 1$, esto es, todas las posibles historias que se pueden observar hasta el tiempo t. Un elemento genérico de

$\mathbb{H}_t$, llamado $t-$*historia*, es denotado por $h_t = (x_0, a_0, x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$, donde $(x_i, a_i) \in \mathbb{K}$ para $x_i \in X$ y $a_i \in A(x_i)$, $i = 1, 2, \cdots, t-1$.

De esta manera, una política $\pi = \{\pi_t, t = 0, 1, ..\}$ es una sucesión de kérneles estocásticos definidos sobre A dada la historia del proceso $\mathbb{H}_t$ y satisfacen que $\pi_t(A(x_t)|h_t) = 1$ para toda $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$

Al conjunto de todas las políticas se denotará por Π.

Denotemos a la familia de kérneles estocásticos sobre A dado X, como $P(A|X)$ y sea Φ el conjunto de todos los kérneles estocásticos φ en $P(A|X)$ tales que para toda $x \in X$ se tiene $\varphi(A(x)|x) = 1$.

Las políticas se clasifican según el tipo de regla de decisión que la compone y según dependan del tiempo o no, de esta manera una política $\pi \in \Pi$ es:

Markoviana Aleatorizada (Π_{RM}) Si existe una sucesión $\{\varphi_t\}$ de kérneles estocásticos con $\varphi_t \in \Phi$, tales que $\pi_t(\cdot|h_t) = \varphi(\cdot|x_t)$ para toda $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$

Markoviana Aleatorizada Estacionaria (Π_{RS}) Si existe $\varphi \in \Phi$ un kérnel estocástico, tal que: $\pi_t(\cdot|h_t) = \varphi(\cdot|x_t)$ para toda $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$ En este tipo de políticas es usual denotar a π solo por $\pi = \varphi$ y al conjunto Π_{RS} por Φ.

Determinista (Π_D) Si existe una sucesión $\{g_t\}$ de funciones medibles con $g_t : \mathbb{H}_t \to A$, tales que, para cada $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$, se tiene que $g_t(h_t) \in A(x_t)$ y $\pi_t(\cdot|h_t)$ está concentrada en $g_t(h_t)$.

Determinista Markoviana (Π_{DM}) Si existe una sucesión $\{f_t\}$ de funciones medibles $f_t : X \to A$, tales que $f_t(x_t) \in A(x_t)$ para cada $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$

Determinista Markoviana Estacionaria (Π_{DS}) Si existe una función medible $f : X \to A$, tal que $f(x_t) \in A(x_t)$ y $\pi_t(\cdot|h_t)$ está concentrada en $f(x_t)$ para cada $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$

Dado un estado inicial $x_0 = x \in X$, y una política $\pi \in \Pi$ entonces por el Teorema de Ionescu-Tulcea [5], existe una única medida de probabilidad P_x^π, inducida por la pareja (x, π) sobre el espacio $\Omega = (X \times A)^\infty$ con su respectiva σ-álgebra producto $\mathcal{F}$, tal que, para cada $C \in \mathcal{B}(A)$, $B \in \mathcal{B}(X)$, $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, ...$ se tiene que

$$P_x^\pi(a_t \in C \mid h_t) = \pi_t(C \mid h_t), \tag{1.1}$$

$$P_x^\pi(x_{t+1} \in B \mid h_t, a_t) = Q(B \mid x_t, a_t). \tag{1.2}$$

De esta manera, la pareja (x, π) determina un proceso estocástico

$$((\Omega, \mathcal{F}, P_x^\pi), \{x_t\})$$

llamado ***Proceso de Decisión de Markov (PDM)***. La esperanza con respecto a P_x^π será denotada por E_x^π.

Un PDM estará dotado de una función real, llamada **función objetivo** o criterio de rendimiento, que medirá en algún sentido la calidad de cada política. Se definen a continuación los que se utilizarán en esta tesis.

a) **Recompensa Descontada Total Esperada**. Sean $x \in X$ y $\pi \in \Pi$, se define la Recompensa Descontada Total Esperada como

$$V(x, \pi) := E_x^\pi \left[\sum_{t=0}^{N} \alpha^t r(x_t, a_t) \right], \tag{1.3}$$

donde $\alpha \in (0, 1)$ se denomina factor de descuento.

b) **Costo Total Esperado**. Sea $x \in X$ y $\pi \in \Pi$, se define el Costo Total Esperado como

$$V(x, \pi) := E_x^\pi \left[\sum_{t=0}^{N} c(x_t, a_t) \right]. \tag{1.4}$$

b) **Recompensa Total Esperada**. Sea $x \in X$ y $\pi \in \Pi$, se define la Recompensa Total Esperada como

$$V(x, \pi) := E_x^\pi \left[\sum_{t=0}^{N} r(x_t, a_t) \right]. \tag{1.5}$$

Al entero positivo N se le conoce como horizonte del problema, el cual representa el número de etapas en el cual el sistema está operando y puede ser finito o infinito.

El **Problema de Control Óptimo** busca maximizar (o minimizar si $V(x, \pi)$ es como en (1.4)) la función $\pi \to V(x, \pi)$ sobre Π, para toda $x \in X$, esto es, encontrar una política óptima π^* y el valor óptimo V^*, los cuales se definen como sigue:

Definición 1.1.1. *Para cada $x \in X$ se define*

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

V^* se le llama función de valores óptimos o valor óptimo.

Definición 1.1.2. *Una política $\pi^* \in \Pi$, es óptima, si*

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

Observación 1.1.2. *En un MCM dado, una forma equivalente de presentar la dinámica es por medio de una ecuación en diferencias:*

$$x_{t+1} = F(x_t, a_t, \xi_t), \tag{1.6}$$

$t = 0, 1, \ldots$ con $x_0 = x \in X$ fijo, donde la perturbación $\{\xi_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con valores en algún conjunto de Borel no vacío S y una distribución común Λ y $F : \mathbb{K} \times S \to X$ es una función medible. La ley de transición Q está dada por

$$\begin{aligned} Q(B \mid x, a) &= \mathbb{P}(x_{t+1} \in B | x_t = x, a_t = a) \\ &= \mathbb{P}(F(x_t, a_t, \xi_t) \in B | x_t = x, a_t = a) \\ &= \mathbb{P}(F(x, a, \xi_t) \in B) \\ &= \int_S 1_B(F(x, a, s))\Lambda(ds), \end{aligned}$$

para todo $B \in \mathcal{B}(X)$ y $(x, a) \in \mathbb{K}$, donde 1_B es la función indicadora del conjunto B.

Observación 1.1.3. *Algunos ejemplos de este modelo son:*

1. *El modelo aditivo $x_{t+1} = F(x_t, a_t) + \xi_t$.*
2. *El modelo separable $x_{t+1} = H(F(x_t, a_t), \xi_t)$.*

En particular, si la perturbación se concentra en un sólo punto s_0, el sistema se puede representar del modo siguiente:

$$x_{t+1} = F(x_t, a_t),\ t = 0, 1, ... \tag{1.7}$$

Y en este caso, $Q(B \mid x, a) = 1_B(F(x, a))$, $B \in \mathcal{B}(X)$.

A (1.7) lo denominamos *Sistema de Control Determinista*. Aquí la interpretación del modelo es como sigue: en cada tiempo $t \in \{0, 1, 2, ...\}$, se observa el estado del sistema $x_t = x \in X$, se aplica una acción $a_t = a \in A(x)$, y como resultado se obtiene una recompensa $r(x, a)$ y el sistema es transferido a un nuevo estado x_{t+1}, determinado por (1.7) con $x_0 = x \in X$ el estado inicial.

1.2. Programación Dinámica

La técnica de Programación Dinámica proporciona un método para resolver el problema de control óptimo, es decir, bajo condiciones sobre el MCM, esta técnica permite determinar el valor óptimo y la política óptima.

Observación 1.2.1. *La teoría que se presentará a continuación es válida con los cambios necesarios en los supuestos del modelo* $(X, A, \{A(x)|x \in X\}, Q, r)$ *para los problemas que en lugar de una función de recompensa* r*, consideran una función de costo* c*.*

1.2.1. Problemas con Horizonte Finito

Considere un modelo de control de Markov fijo $(X, A, \{A(x)|x \in X\}, Q, r)$ y a la recompensa descontada total esperada como criterio de rendimiento. El siguiente teorema proporciona un algoritmo que permite encontrar una política y valor óptimo para un problema de horizonte finito N.

Teorema 1.2.1. *Sean* $J_0, J_1, ..., J_N$ *funciones sobre* X *definidas por:*

$$J_N := 0, \tag{1.8}$$

y para $t = N-1, N-2, ..., 0$*,*

$$J_t(x) := \max_{A(x)} \left[\alpha^t r(x,a) + \int_X J_{t+1}(y) Q(dy|x,a) \right], \quad x \in X. \tag{1.9}$$

Suponga que estas funciones son medibles y que, para cada $t = 0, ..., N-1$*, existen selectores* $f_t \in \mathbf{F}$ *tales que* $f_t(x) \in A(x)$ *alcanzan el máximo en (1.9) para todo* $x \in X$*; i.e,* $\forall x \in X$ *y* $t = 0, ..., N-1$*,*

$$J_t(x) = \alpha^t r(x, f_t) + \int_X J_{t+1}(y) Q(dy|x, f_t). \tag{1.10}$$

Entonces la política (determinista Markoviana) $\pi^* = \{f_0, ..., f_{N-1}\}$ *es la óptima y la función de valor* J^* *es igual a* J_0*, es decir,*

$$J^*(x) = J_0(x) = J(\pi^*, x), \ \forall x \in X. \tag{1.11}$$

Notemos que el teorema supone la existencia de maximizadores, en algunas ocasiones, éstos pueden darse explícitamente, sin embargo, desde un punto de vista teórico es conveniente dar condiciones que garanticen este supuesto, llamado *Condición de Selección Medible.*

Condiciones 1.2.1.

1. *Para cada $x \in X$, $r(x,a)$ es semicontinua superiormente (u.s.c) en $a \in A(x)$ (es decir, para cada $x \in X$, $\limsup_{n\to\infty} r(x,a_n) \leq r(x,a)$, para cualquier sucesión $\{a_n\}$ en A que converge hacia a), acotada superiormente e inf-compacta sobre $\mathbb{K}$ (es decir, para toda $x \in X$ y $\lambda \in \mathbb{R}$, el conjunto $\{a \in A(x) | r(x,a) \leq \lambda\}$ es compacto).*

2. *La ley de transición Q es:*

   a) *débilmente continua (es decir, $u'(x,a) := \int u(y)Q(dy \mid x,a)$ es continua y acotada sobre $\mathbb{K}$ para toda función continua y acotada u sobre X), ó*

   b) *fuertemente continua (es decir, $u'(x,a)$ es continua y acotada sobre $\mathbb{K}$ para toda función medible y acotada u sobre X).*

Condiciones 1.2.2.

1. *Para cada $x \in X$, $A(x)$ es compacto.*

2. *$r(x,a)$ es u.s.c. $\forall a \in A(x)$, $x \in X$.*

3. *La función $u'(x,a) := \int u(y)Q(dy \mid x,a)$ sobre $\mathbb{K}$ satisface alguna de las condiciones siguientes:*

   a) *$u'(x,\cdot)$ es u.s.c en $A(x)$ para toda función continua y acotada u sobre X.*

   b) *$u'(x,\cdot)$ es u.s.c en $A(x)$ para toda función medible y acotada u sobre X.*

Condiciones 1.2.3.

1. *Para cada $x \in X$, $A(x)$ es compacto y la multifunción $x \to A(x)$ es u.s.c.*

2. *$r(x,a)$, $a \in A(x)$, $x \in X$ es u.s.c y acotado superiormente.*

3. *La ley de transición Q es:*

   a) *débilmente continua, ó*

   b) *fuertemente continua.*

La Condición de Selección Medible se verifica si se satisface alguna de las Condiciones 1.2.1, 1.2.2 o 1.2.3, la prueba puede verse en [18].

A los modelos que satisfacen la Condición 1.2.1 se les llama *Modelos Semicontinuos - Semicompactos* y a los modelos que satisfacen la Condición 1.2.2 o 1.2.3 se les llama *Modelos Semicontinuos*.

Observación 1.2.2. *Observe que para el modelo determinista, representado por la ecuación en diferencias (1.7), las Condiciones 1.2.2 y 1.2.3 se satisfacen directamente si se cumplen las siguientes:*

Condiciones 1.2.4.

1. *Para cada $x \in X$, $A(x)$ es compacto.*
2. *r y F son funciones continuas sobre $\mathbb{K}$.*

1.2.2. Problemas con Horizonte Infinito

En la sección anterior se consideraron problemas con horizonte finito, sin embargo, existen problemas que no necesariamente tienen un tiempo de detención natural definido, esto es, que poseen horizonte infinito. Se exponen a continuación las características que deben tener este tipo de modelos para que exista una solución.

Definición 1.2.1. *Las funciones de iteración de valores se definen como:*

$$v_n(x) := \max_{a \in A(x)} \left\{ r(x,a) + \alpha \int_X v_{n-1}(y) Q(dy|x,a) \right\}, \tag{1.12}$$

$x \in X$ y $n = 1, 2, ...,$ con $V_0 \equiv 0$.

Definición 1.2.2. *Una función medible $u^* : X \to \mathbb{R}$ se dice ser solución de la* ***Ecuación de Programación Dinámica*** *(EPD) si:*

$$u^*(x) = \max_{a \in A(x)} \left[r(x,a) + \alpha \int_X u^*(y) Q(dy|x,a) \right], \tag{1.13}$$

$x \in X$.

Definición 1.2.3. *Dada una función medible $u : X \to \mathbb{R}$ se define al Operador de Programación dinámica (OPD), denotado por T, como sigue:*
Para cada $x \in X$

$$T(u(x)) := \sup_{A(x)} \left[r(x,a) + \alpha \int_X u(y) Q(dy|x,a) \right]. \tag{1.14}$$

A continuación, se proporcionan las condiciones necesarias para garantizar que V^* satisface la Ecuación de Programación Dinámica y que además, existe una política óptima que genera un máximo en la EPD.

1.2.2.1. Modelos Semicontinuos-Semicompactos

En estos modelos se consideran recompensas acotadas superiormente, y la técnica se basa en utilizar las funciones de iteración de valores v_n, las cuales son las funciones de valor óptimo V_n de la $n-$ésima recompensa descontada con recompensa terminal cero y se prueba que $\lim_{n\to\infty} v_n(x) = V^*(x)$, $x \in X$. Este método es conocido como *Aproximaciones Sucesivas* y requiere de la condición siguiente.

Condiciones 1.2.5. *Existe una política π tal que $V(x,\pi) < \infty$ para cada $x \in X$.*

1.2.2.2. Modelos Semicontinuos

Sea $w : X \to [1,\infty)$ una función medible. Si u es una función medible sobre X, entonces su $w-norma$ es definida como

$$\|u\|_w := \sup_{x\in X} \frac{\mid u(x) \mid}{\mid w(x) \mid},$$

w es llamada función de peso y $\mathbb{B}_w(X)$ denota al espacio de Banach de funciones reales, medibles y w-acotadas definidas en X.

El modelo con horizonte infinito con este enfoque garantiza que la función de valor óptimo V^* es un punto fijo del OPD, probándose inicialmente que el OPD es una contracción, la sucesión de funciones de iteración de valores converge en w-norma a V^* y garantiza también la existencia de una política óptima.

Condiciones 1.2.6.

1. *Existen constantes $\bar{c}$ y β, con $1 \le \beta < 1/\alpha$, ($\alpha \in (0,1)$ el factor de descuento) y una función de peso $w \ge 1$ sobre X tal que para todo estado $x \in X$,*

   a) $\sup_{a\in A(X)} \mid r(x,a) \mid \le \bar{c}w(x)$ *y*

   b) $\sup_{a\in A(X)} \int w(y)Q(dy \mid x,a) \le \beta w(x)$.

2. *Para cada $x \in X$, la función $w'(x,a) = \int w(y)Q(dy \mid x,a)$ es continua en $a \in A(x)$.*

Condiciones 1.2.7.

1. *La misma suposición que 1.2.6 1, excepto que la función w se requiere continua.*

2. Para cada $x \in X$, la función $w'(x,a) = \int w(y)Q(dy \mid x,a)$ es continua sobre $\mathbb{K}$.

Teorema 1.2.2. *Bajo cualquiera de las Condiciones 1.2.5, 1.2.6 o 1.2.7 se tiene que V^* satisface la Ecuación de Programación Dinámica (1.13) y existe $f \in \mathbf{F}$ tal que alcanza el mínimo en (1.13) para todo $x \in X$; es decir, $\forall x \in X$,*

$$V^*(x) = r(x,f) + \alpha \int_X V^*(y)Q(dy|x,f). \tag{1.15}$$

Además,

a) Bajo las Condiciones 1.2.5, $v_n \uparrow V^*$.

b) Bajo las Condiciones 1.2.6 o 1.2.7, $\|v_n - V^*\|_w \leq \bar{c}\gamma^n/(1-\gamma)$, $n = 1,2,\ldots$, donde $\gamma := \alpha\beta$.

La demostración del teorema anterior puede verse en [18] y [19] donde podemos notar además, que bajo las Condiciones 1.2.5, V^* es una solución máxima y bajo las Condiciones 1.2.6 o 1.2.7, V^* es la única solución en el espacio $\mathbb{B}_w(X)$.

Observación 1.2.3. *Observe que para un modelo determinista, representado por una ecuación en diferencias como la dada en (1.7), la función de valor óptimo V^* satisface:*

$$V^*(x) = \max_{a \in A(x)} [r(x,a) + \alpha V^*(F(x,a)).] \tag{1.16}$$

Por otro lado, como el interés es tomar en cuenta los costos de información dentro del problema de control óptimo, presentamos en la sección siguiente los conceptos más importantes de la Teoría de la Información.

1.3. Teoría de la Información

La Teoría de la Información nació del estudio de la comunicación eléctrica y la forma de procesamiento y transmisión de la información, proporciona una medida universal "el bit" de la cantidad de información en términos de elección o inseguridad. El objetivo principal de la Teoría de la Información es la transmisión óptima de mensajes en base a la cantidad de información para aumentar la velocidad, de esta manera la Teoría de la Información tuvo que dar respuesta a dos preguntas fundamentales: ¿Cuál es la compresión máxima de los datos?, es decir, ¿Cuál es el mínimo de información que se necesita para transmitir un mensaje sin pérdida de datos?, cuya respuesta es la **entropía**, la

cual es una medida de la incertidumbre de una variable aleatoria, y la segunda pregunta es ¿cuál es la velocidad de transmisión máxima de la comunicación?, que tiene como respuesta la **capacidad del canal**. Para ello se busca un sistema de codificación que utilice la menor información necesaria de tal manera que al transmitirla no se pierdan partes del mensaje [7].

Los conceptos fundamentales definidos en la Teoría de la Información - entropía, la entropía relativa y mutua información - se definen como los funcionales de las distribuciones de probabilidad, además, caracterizan el comportamiento de secuencias largas de variables aleatorias, permiten estimar las probabilidades de eventos raros (Teoría de las Grandes Desviaciones) y ayudan a encontrar el mejor exponente de error en las pruebas de hipótesis. Como ya se comentó en la sección anterior, un proceso se dice que es de Markov cuando las variables aleatorias forman un proceso estocástico en el que cada variable aleatoria depende sólo de la que le precede y es condicionalmente independiente de todas las otras variables aleatorias precedentes. Se muestra, igual que en el caso independiente e idénticamente distribuido, que la entropía crece (asintóticamente) linealmente con el tamaño de la muestra a una velocidad $H(Y)$, que llamaremos la tasa de entropía del proceso, que es una medida de incertidumbre de la variable aleatoria Y. La velocidad $H(Y)$ se interpreta como la mejor compresión de los datos [7].

Definición 1.3.1. *Para una variable aleatoria discreta Y con valores en $\mathcal{Y} \subseteq \mathbb{R}$ y función de masa de probabilidad p, la **entropía** se define como:*

$$\begin{aligned} H(Y) &:= -\sum_{y\in\mathcal{Y}} p(y)\log_2 p(y) \\ &= E_p\left[\log_2 \frac{1}{p(Y)}\right]. \end{aligned}$$

Donde E_p indica el valor esperado con respecto a la distribución de probabilidad p, esto es, la entropía puede interpretarse también como el valor esperado de la variable aleatoria $\log_2(1/p(Y))$. Es fácil probar que la distribución uniforme es la que genera la mayor entropía $H(Y) = \log_2 |\mathcal{Y}|$ debido a que es la que posee mayor incertidumbre, contraria a esta situación, la distribución que genera la entropía mínima $H(Y) = 0$ es cualquier función que ponga toda la masa en un solo estado. Tal distribución no tiene incertidumbre.

Lema 1.3.1. $H(Y) \geq 0$.

Demostración. Sea $y \in \mathcal{Y}$, como $0 \leq p(y) \leq 1$ implica que $\log_2(1/p(y)) \geq 0$, entonces $E_p\left[\log_2 \frac{1}{p(Y)}\right] \geq 0$, esto es, $H(Y) \geq 0$. □

Cuando se busca comparar dos distribuciones se usa la *entropía relativa* también llamada *divergencia de Kullback-Leibler* (divergencia KL), la cual pro-

porciona una medida de disimilaridad de dos distribuciones de probabilidad, y se define como sigue:

Definición 1.3.2. *Sean p y q dos distribuciones de probabilidad, la* ***entropía relativa*** *se define como:*

$$\begin{aligned} KL(p||q) &:= \sum_{y\in\mathcal{Y}} p(y)\log_2\frac{p(y)}{q(y)}, \\ &= E_p\left[\log_2\frac{p(Y)}{q(Y)}\right], \end{aligned} \tag{1.17}$$

aquí, la suma puede ser reemplazada por integrales para el caso de variables aleatorias continuas.

Se usa la convención de que $0\log\frac{0}{0}=0$, $0\log\frac{0}{q}=0$ y $p\log\frac{p}{0}=\infty$, así, si existe $y'\in\mathcal{Y}$ tal que $p(y')>0$ y $q(y')=0$, entonces $KL(p||q)=0$. Esta medida nos proporciona el número promedio de bits extra que se necesitan para codificar un mensaje que se decodificó usando la distribución q en lugar de usar la distribución real p. De esta manera, si p y q son iguales, no hay diferencia y no hay bits extra porque se estará haciendo la decodificación con la distribución correcta y no habrá pérdida de información, si no fueran iguales, entonces la entropía relativa será positiva, esto se resume en el Teorema 1.3.1 que se presenta a continuación:

Teorema 1.3.1. *Sean p y q dos distribuciones de probabilidad. Entonces*

$$KL(p||q)\geq 0,$$

la igualdad ocurre si $p=q$.

Demostración. Sea $\bar{\mathcal{Y}}=\{y : p(y)>0\}$ el soporte de $p(y)$, entonces

$$\begin{aligned} -KL(p||q) &= -\sum_{y\in\bar{\mathcal{Y}}} p(y)\log_2\frac{p(y)}{q(y)} \\ &= \sum_{y\in\bar{\mathcal{Y}}} p(y)\log_2\frac{q(y)}{p(y)} \\ &\leq \log_2\sum_{y\in\bar{\mathcal{Y}}} p(y)\frac{q(y)}{p(y)} \\ &= \log_2\sum_{y\in\bar{\mathcal{Y}}} q(y) \\ &\leq \log_2\sum_{y\in\mathcal{Y}} q(y) \\ &= \log_2 1 \\ &= 0, \end{aligned} \tag{1.18}$$

la primera desigualdad en (1.18) se sigue por la desigualdad de Jensen [7] y como $\log z$ es una función estrictamente cóncava de z, se da la igualdad sí y sólo si $q(y)/p(y)$ es constante casi seguramente, es decir, $q(y) = cp(y)$ para toda $y \in \mathcal{Y}$. Por consiguiente, $\sum_{y\in\bar{\mathcal{Y}}} q(y) = c\sum_{y\in\bar{\mathcal{Y}}} p(y) = c$. La segunda desigualdad en (1.18) se vuelve igualdad sólo si $\sum_{y\in\bar{\mathcal{Y}}} q(y) = \sum_{y\in\mathcal{Y}} q(y) = 1$, lo cual implica que $c = 1$. Por lo tanto, tenemos que $KL(p||q) = 0$ sí solo si $p(y) = q(y)$ para toda $y \in \mathcal{Y}$. □

Capítulo 2

Problema de la Ruta más Corta con el Criterio de Energía Libre

En este capítulo se presentan algunos antecedentes del estudio de PDMs relacionados con la Teoría de la Información y el problema de la ruta más corta, así también se plantea y resuelve el primer problema principal de ésta tesis.

2.1. Antecedentes

Estamos interesados en "El problema de la Ruta más Corta". Este ha sido estudiado con la teoría de Procesos de Decisión de Markov en [3], [4], para espacios de estados finitos, y en [2] para espacios numerables pero en ambos casos sin considerar costos de información.

Recientemente, se ha estudiado la relación entre los PDMs y la Teoría de la Información, buscando políticas que optimicen el criterio de rendimiento con el menor costo de información el cual es definido a través de la entropía relativa. El equilibrio de estas dos medidas se establece en la llamada función de Energía libre que combina ambos criterios en espacios de estados finitos y con diferentes criterios de rendimiento, por ejemplo, en [28] se pone interés en el costo que se va acumulando a lo largo del tiempo pero que es afectado por un factor de descuento que modifica el costo según el momento del tiempo en que se genere, en [16] el costo total que se espera obtener a lo largo del tiempo en el que se introduce una subtarea, en [41] se analiza la recompensa total esperada a través del tiempo en la que también influye el estado posterior, o en [17] que introduce en el PDM la intuición del agente con respecto a lo que sucederá y principalmente en [33] que trabaja con la recompensa total esperada que sirvió como base para la investigación de los criterios ya mencionados. Sin embargo, todos se desarrollan bajo el supuesto de que el espacio de estados es finito, una extensión a espacios más generales, empezando en espacios

numerables, proporcionaría solución a problemas importantes, por ejemplo, en Finanzas donde el problema central es determinar una estrategia de inversión que optimice el valor de un portafolio. Estos modelos económicos incorporan ventajas para los vendedores y los compradores en diversos mercados, aunque, desde el punto de vista de la Teoría de la Información, es notable la falta de información, como por ejemplo: Calidad del producto, falta de información de los acreedores ante el riesgo, etc. [31]. Por tal motivo consideramos de interés el desarrollo de la teoría expuesta, ya que permitirá dar respuesta a este tipo de problemas, además de que se presentará una aportación en la teoría de PDMs.

2.2. Planteamiento del Problema

Abordamos el problema de la Ruta más Corta a través de un Proceso de Control de Markov con un criterio de rendimiento sujeto a una restricción sobre los costos de información que es medido a través de la entropía relativa, esta última, funciona como una medida en bits de la ineficiencia de asumir una política distinta a la verdadera, situación que ocurre cuando se construye un código para enviar la descripción de la acción computacionalmente [7], además, se busca alcanzar un estado objetivo.

Formalmente, se tiene un MCM, consistente de la quíntupla

$$(X, A, \{A(x)|x \in X\}, \hat{Q}, r),$$

donde ahora $X \subset \mathbb{R}$ es un subconjunto infinito numerable, $A(x) = A$ es finito con cardinalidad n, $n \in \mathbb{N}$, $\hat{Q}$ es la ley de transición y además es definida la distribución condicional de la variable aleatoria $X_{t+1}|X_t, a_t$, $t \geq 1$ como sigue: $Q(X_{t+1} = y|X_t = x, a_t = a) := \hat{Q}(y|x, a)$ y r es la función de recompensa de un solo paso, como se definió en el Capítulo 1, la cual se considera negativa.

Estamos interesados en un sistema que se desarrolle como sigue: El agente elige la acción a una vez observado el estado x de acuerdo con la ley $\varphi(a|x)$, la cual es desconocida para el agente y como resultado, el sistema obtiene una recompensa $r(x, a)$ y paga un costo dado por $\log \frac{\varphi(a_t|x_t)}{\rho(a_t|x_t)}$, donde ρ es alguna política por defecto, la cual representa una política alternativa usada por el actuador, en ausencia de la información del controlador. La política generará una recompensa total esperada $V(x, \varphi)$ y un costo de información $J(x, \varphi)$ dados a continuación por (2.1) y (2.2) respectivamente:

$$V(x, \varphi) = \lim_{k \to \infty} E_x^{\varphi} \left[\sum_{t=0}^{k-1} r(x_t, a_t) \right], \tag{2.1}$$

$$J(x,\varphi) = \lim_{k\to\infty} E_x^{\varphi}\left[\sum_{t=0}^{k-1} \log \frac{\varphi(a_t|x_t)}{\rho(a_t|x_t)}\right], \tag{2.2}$$

$x \in X$, $\varphi \in \Phi$.

El objetivo es obtener la mayor recompensa con la menor pérdida de información, entonces, a lo largo del tiempo queremos maximizar la recompensa sujeta a (s.a) un costo de información acotado por $l > 0$, esto es:

$$\sup_{\varphi\in\Phi} V(x,\varphi) \; s.a. \; J(x,\varphi) \leq l. \tag{2.3}$$

El problema de optimización restringido de encontrar el valor máximo dada una cota en la información se puede convertir en uno sin restricciones usando el método de Lagrange como sigue:

$$\inf_{\substack{\varphi\in\Phi \\ \lambda>0}} L(\varphi,\lambda) = \inf_{\substack{\varphi\in\Phi \\ \lambda>0}} \{-V(x,\varphi) + \lambda[J(x,\varphi) - l]\}, \tag{2.4}$$

donde λ es el multiplicador de Lagrange. Sea $\beta = 1/\lambda > 0$ llamado parámetro de compensación, el cual controla el intercambio entre la información y la recompensa, el problema (2.4) es equivalente a

$$\inf_{\varphi\in\Phi,\beta>0} \{J(x,\varphi) - \beta V(x,\varphi)\}. \tag{2.5}$$

Sea $\varphi \in \Phi$, definimos

$$C_\varphi(x,a) := \log n\varphi(a|x) - \beta r(x,a), \; x \in X, \; a \in A. \tag{2.6}$$

Definición 2.2.1. *Sea $\varphi \in \Phi$ y $x \in X$ un estado inicial, definimos la* ***Energía Libre*** *como el Costo Total Esperado (1.4), esto es:*

$$v(\varphi,x,\beta) = \lim_{k\to\infty} E_x^{\varphi}\left[\sum_{t=0}^{k-1} C_\varphi(x_t,a_t)\right] = J(x,\varphi) - \beta V(x,\varphi). \tag{2.7}$$

En consecuencia, consideramos un MCM $(X, A, \{A(x)|x \in X\}, \hat{Q}, C_\varphi)$, este sistema opera como sigue: Si el sistema está en el estado $x_t = x \in X$ al tiempo t y el controlador elige una acción $a_t = a \in A(x_t)$ con probabilidad $\varphi(a|x)$, como una consecuencia, un costo $C_\varphi(x,a)$ es pagado y el sistema se mueve al siguiente estado de acuerdo a la ley $\hat{Q}(\cdot|x,a)$.

Sean $x \in X$, $\beta > 0$ y ρ fijos, el ***Primer Problema de Control Óptimo (PPCO)*** de interés en esta tesis es minimizar el costo total esperado $v(\varphi,x,\beta)$, $\varphi \in \Phi$.

Con lo anterior nos planteamos las siguientes preguntas de investigación:

- ¿Es posible resolver el Problema de Control Óptimo a través de la técnica de Programación Dinámica?
- ¿Son válidos para espacios de estados infinitos algunos resultados ya probados para espacios finitos?

Objetivo General

Resolver el Problema de la Ruta más Corta a través de un Problema de Control Óptimo buscando políticas óptimas en el sentido que permitan alcanzar el máximo valor, dada una restricción sobre el control de la información en espacios de estados infinito numerables.

Objetivos particulares

- Proponer condiciones en las componentes del modelo de control, las cuales permitan llevar a cabo un análisis del problema de optimización en el contexto de programación dinámica para espacios de estados numerables.
- Validar la ecuación de programación dinámica bajo el criterio de entropía relativa.
- Caracterizar la política óptima y el valor óptimo.

2.3. Metodología

Siguiendo las ideas de trabajos anteriores ([2], [3], [4], [16], [17], [28], [41]) se propone abordar el problema de control óptimo a través de la técnica de programación dinámica y proponer una solución, de esta forma, la metodología consiste en buscar políticas óptimas en el sentido que reflejen un equilibrio entre la maximización de las recompensas esperadas (valor) y la minimización de los costos involucrados en el envío de la información, introduciendo en la solución un tipo de políticas especiales llamadas políticas propias [2]. Utilizando como herramienta básica, la aplicación de la técnica de programación dinámica. De este modo es necesario validar el procedimiento de iteración de valores óptimos y la convergencia de éste a la función de valor óptimo. La metodología se puede apreciar en la Figura 2.1.

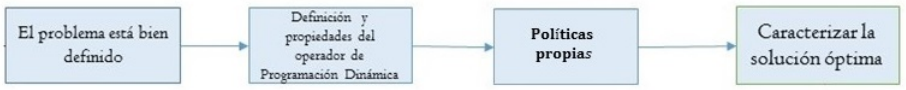

Figura 2.1: Diagrama de solución del **PPCO**.

2.4. Problema de Control Óptimo con el Criterio de Energía Libre

Recordemos que dado el MCM $(X, A, \{A(x)|x \in X\}, \hat{Q}, C_\varphi)$, el problema de control óptimo es minimizar el costo total esperado $v(\varphi, x, \beta)$, $\varphi \in \Phi$ definido en (2.7) con $x \in X$, $\beta > 0$ y ρ fijos.

Introducimos al problema la siguiente condición:

Condiciones 2.4.1. *Existe $g \in X$ tal que $r(g, a) = 0$ y $Q(g|g, a) = 1$ para toda $a \in A(x)$.*

Observación 2.4.1. *El estado g en la Condición 2.4.1 será llamado destino o (estado terminal) ([2], [33]).*

Con base en la Condición 2.4.1, podemos re definir $J(g, \varphi) := 0$ y $C_\varphi(g, a) := 0$, $a \in A$.

Definición 2.4.1. *Para un estado $x \in X$ dado, una política estacionaria $\varphi \in \Phi$ es llamada **política propia** de x [2], si*

$$\sum_{t=0}^{\infty} P_x^\varphi(x_t \neq g) = \sum_{t=0}^{\infty} P_x^\varphi(\tau > t) < \infty,$$

donde $\tau := \text{ínf}\{t \geq 0 : x_t = g\}$. De otra manera, decimos que la política es impropia. Se denota por Π^P al conjunto de políticas propias.

Lema 2.4.1. *Para cada $x \in X$, $\varphi \in \Pi^P$ y $F : X \times A \to \mathbb{R}$, una función medible, no-negativa y acotada, se cumplen las siguientes afirmaciones:*

a) $P_x^\varphi(\tau < \infty) = 1$.

b) $\lim_{k\to\infty} E_x^\varphi \left[\sum_{t=0}^{k-1} F(x_t, a_t)\right] < \infty$.

Demostración. a) Considere $x \in X$ y $\varphi \in \Pi^P$, fijos. Entonces, por la propiedad de continuidad de la medida de probabilidad P_x^φ, se tiene que

$$P_x^\varphi(\tau = \infty) = \lim_{t\to\infty} P_x^\varphi(\tau > t) = 0,$$

la última igualdad es consecuencia de la Definición 2.4.1.

b) Considere F una función medible no-negativa y suponga que existe $K \in \mathbb{R}$ tal que $0 \leq F(x,a) \leq K$, para todo $(x,a) \in \mathbb{K}$. Sea $x \in X$ y $\varphi \in \Pi^P$, fijos, entonces,

$$\begin{aligned}\lim_{k\to\infty} E_x^\varphi\left[\sum_{t=0}^{k-1} F(x_t,a_t)\right] &= E_x^\varphi\left[\sum_{t=0}^{\infty} F(x_t,a_t)I[t>\tau]\right] + E_x^\varphi\left[\sum_{t=0}^{\infty} F(x_t,a_t)I[t\leq\tau]\right] \\ &< KP_x^\varphi(\tau<\infty) + K\sum_{t=0}^{\infty} P_x^\varphi(\tau>t).\end{aligned}$$

En consecuencia, por la parte a) y la Definición 2.4.1 se sigue que

$$\lim_{k\to\infty} E_x^\varphi\left[\sum_{t=0}^{k-1} F(x_t,a_t)\right] < \infty.$$

□

Condiciones 2.4.2. *El conjunto de políticas propias es no vacío, es decir,* $\Pi^P \neq \phi$.

Sea $x \in X$, $\varphi \in \Phi$, denotamos por:

$$\begin{aligned}\hat{Q}(\cdot|x,\varphi) &:= \sum_{a\in A(x)} \hat{Q}(\cdot|x,a)\varphi(a|x), \\ \hat{Q}^k(\cdot|x,\varphi) &:= \sum_{y\in X} \hat{Q}^{k-1}(\cdot|y,\varphi)\hat{Q}(y|x,\varphi), \qquad (2.8) \\ C_\varphi(x,\varphi) &:= \sum_{a\in A(x)} C_\varphi(x,a)\varphi(a|x),\end{aligned}$$

donde $k = 1,2,...,\hat{Q}^0(x|x,\varphi) := 1$, donde (2.8) es la probabilidad de transición en k pasos.

Observación 2.4.2. *Observe que para $x \in X$, $\varphi \in \Phi$, $J(x,\varphi)$ es una función no-negativa, debido a que $J(x,\varphi)$ es el valor esperado de la entropía relativa, la cual es una función no-negativa (Teorema 1.18). Además, observe que $V(x,\varphi)$ es una función no-positiva. Por lo tanto, se garantiza que la Energía Libre $v(x,\varphi,\beta)$ (2.7) es una función no-negativa.*

Lema 2.4.2. *Sean $x \in X$, $\varphi \in \Pi^P$ y $\beta > 0$ fijos, entonces $v(\varphi,x,\beta) < \infty$.*

Demostración. Sean $x \in X$, $\varphi \in \Pi^P$ y $\beta > 0$ fijos. En primer lugar observar que la función de costo, C_φ es acotada:

$$C_\varphi(x,\varphi) = \sum_{a\in A(x)} \varphi(a|x)\log n\varphi(a|x) - \beta \sum_{a\in A(x)} r(x,a)\varphi(a|x) \leq \hat{K},$$

donde $\hat{K} := \log n - H(x,\varphi) - \text{mín}_{a\in A(x)}\, r(x,a)$ y

$$H(x,\varphi) := -\sum_{a\in A(x)} \varphi(a|x)\log\varphi(a|x),$$

es *la entropía de* φ, la cual es una función no negativa (Lema 1.3). En consecuencia, por el Lema 2.4.1 inciso b),

$$v(\varphi,x,\beta) = \lim_{k\to\infty} E_x^\varphi\left[\sum_{t=0}^{k-1} C_\varphi(x_t,\varphi)\right] < \infty.$$

Como $x \in X$, $\varphi \in \Phi$ y $\beta > 0$ son arbitrarios, el resultado ha sido probado. □

Lema 2.4.3. *Sean $x \in X$, $\varphi \in \Phi$ y $\beta > 0$, entonces (2.7) es equivalente a*

$$v(\varphi,x,\beta) = \lim_{k\to\infty} \sum_{t=0}^{k-1}\sum_{y\in X} C_\varphi(y,\varphi)Q^t(y|x,\varphi). \tag{2.9}$$

Observación 2.4.3. *Las siguientes afirmaciones son consecuencia del Lema 2.4.1:*

a) El límite en b) del Lema 2.4.1 es finito para $U \in M^+(X)$, donde $M^+(X)$ es el espacio de funciones medibles, no negativas y acotadas. También,

$$\lim_{k\to\infty} E_x^\varphi\left[\sum_{t=0}^{k-1} U(x_t)\right] = \lim_{k\to\infty}\sum_{t=0}^{k-1}\sum_{y\in X} U(y)Q^t(y|x,\varphi). \tag{2.10}$$

Se sigue que

$$\lim_{k\to\infty}\sum_{y\in X} U(y)Q^k(y|x,\varphi) = 0.$$

b) Para cada política estacionaria impropia μ, se cumple la siguiente relación

$$\lim_{k\to\infty}\sum_{t=0}^{k-1}\sum_{y\in X} C_\mu(y,\mu)Q^t(y|x,\mu) = \infty,$$

para al menos un estado x.

Sea $(X, A, \{A(x)|x \in X\}, \hat{Q}, C_\varphi)$, $x \in X$ y $\beta > 0$ fijo, el ***Problema de Control Óptimo*** de interés es minimizar el costo total esperado $v(\varphi, x, \beta)$, $\varphi \in \Pi^P$.

Definición 2.4.2. *Sea $x \in X$ y $\varphi \in \Phi$, definimos el operador $T_\varphi : M^+(X) \to M^+(X)$, como sigue:*

$$\begin{aligned} T_\varphi v(x) &= \sum_{a \in A(x)} \varphi(a|x) \left[\log n\varphi(a|x) - \beta r(x,a) + E_x^\varphi[v(y)]\right] \\ &= C_\varphi(x,\varphi) + \sum_{y \in X} v(y) Q(y|x,\varphi),\ v \in M^+(X). \end{aligned}$$

Además, podemos reescribir la energía libre de una política $\varphi \in \Phi$, $x \in X$ y $\beta > 0$ como:

$$v(\varphi, x, \beta) = \lim_{k\to\infty} \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y,\varphi) Q^t(y|x,\varphi) = \lim_{k\to\infty} T_\varphi^k v_0(x), \tag{2.11}$$

donde v_0 es la función nula, es decir, $v_0(x) = 0$, $x \in X$.

Lema 2.4.4. *T_φ es un operador monótono.*

Demostración. Sea $x \in X$ fijo. Entonces, para $v, v' \in M^+(X)$, si $v(y) \leq v'(y)$, $\forall y \in X$, note que

$$T_\varphi v(x) - T_\varphi v'(x) = \sum_{y \in X} [v(y) - v'(y)] Q(y|x,\varphi) \leq 0.$$

Así, $T_\varphi v(x) \leq T_\varphi v'(x)$. Debido a que x es arbitrario, T_φ es un operador monótono. □

Lema 2.4.5. *Para toda $v \in M^+(X)$, la k-composición del operador T está dada por:*

$$T_\varphi^k v(x) = \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y,\varphi) Q^t(y|x,\varphi) + \sum_{y \in X} v(y) Q^k(y|x,\varphi),\ k \geq 1,\ x \in X.$$

Demostración. Sea $x \in X$ y $v \in M^+(X)$, haremos la prueba por inducción. El caso $k = 1$ es una consecuencia de la Definición 2.4.2. Supongamos ahora que

$$T_\varphi^k v(x) = \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y,\varphi) Q^t(y|x,\varphi) + \sum_{y \in X} v(y) Q^k(y|x,\varphi),$$

En consecuencia,

$$\begin{aligned} T_\varphi^{k+1}v(x) &= T_\varphi[T_\varphi^k v(x)] \\ &= C_\varphi(x,\varphi) + \sum_{y\in X} T_\varphi^k v(y) Q(y|x,\varphi) \\ &= C_\varphi(x,\varphi) + \sum_{y\in X}\sum_{t=0}^{k-1}\sum_{z\in X} C_\varphi(z,\varphi) Q^t(z|y,\varphi) Q(y|x,\varphi) \\ &\quad + \sum_{y\in X}\sum_{z\in X} v(z) Q^k(z|y,\varphi) Q(y|x,\varphi) \\ &= \sum_{t=0}^{k-1}\sum_{y\in X} C_\varphi(y,\varphi) Q^t(y|x,\varphi) + \sum_{y\in X} v(y) Q(y|x,\varphi). \end{aligned}$$

Por lo tanto, el Lema 2.4.5 se cumple. □

Siguiendo la Definición 1.2.3, el operador de programación dinámica $T : M^+(X) \to M^+(X)$ para este problema, es como sigue

$$Tv(x) := \min_{\varphi\in\Pi^P} \sum_{a\in A(x)} \varphi(a|x) \left[\log n\varphi(a|x) - \beta r(x,a) + \sum_{y\in X} v(y) Q(y|x,a) \right],$$

$x \in X$.

Los siguientes, son resultados generales del enfoque de programación dinámica. Resultados análogos al caso finito pueden ser consultados en [33]. En lo que sigue, se considera la notación: $v_\varphi(x) = v(\varphi, x, \beta)$, $\varphi \in \Phi$, $\beta > 0$.

Proposición 2.4.1. *Las siguientes afirmaciones se cumplen:*

a) Si $\varphi \in \Pi^p$, entonces la función de energía libre asociada v_φ es el único punto fijo de T_φ. Además satisface:

$$\lim_{k\to\infty} T_\varphi^k v(x) = v_\varphi(x),\ x \in X,$$

para toda función $v \in M^+(X)$.

b) Si $\varphi \in \Phi$ satisface

$$v(x) = T_\varphi v(x),\ x \in X,$$

para alguna función $v \in M^+(X)$, entonces φ es propia.

Demostración. (a) Sea $x \in X$ y $v \in M^+(X)$, por el Lema 2.4.5 tenemos,

$$T_\varphi^k v(x) = \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y, \varphi) Q^t(y|x, \varphi) + \sum_{y \in X} v(y) Q^k(y|x, \varphi),$$

$k \geq 1$, entonces, por Observación 2.4.3 b) y (2.11) se cumple que

$$\begin{aligned} \lim_{k\to\infty} T_\varphi^k v(x) &= \lim_{k\to\infty} \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y, \varphi) Q^t(y|x, \varphi) + \lim_{k\to\infty} \sum_{y \in X} v(y) Q^k(y|x, \varphi) \\ &= \lim_{k\to\infty} \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y, \varphi) Q^t(y|x, \varphi), \end{aligned}$$

entonces,

$$\lim_{k\to\infty} T_\varphi^k v(x) = v_\varphi(x). \tag{2.12}$$

Además, por la definición de T_φ,

$$T_\varphi^{k+1} v(x) = T_\varphi[T_\varphi^k v(x)] = C_\varphi(x, \varphi) + \sum_{y \in X} T_\varphi^k v(y) Q^t(y|x, \varphi),$$

por lo tanto, tomando el límite cuando $k \to \infty$, por el Lema 2.4.4, el Teorema de la convergencia monótona [32] y (2.12), se obtiene que

$$v_\varphi(x) = C_\varphi(x, \varphi) + \sum_{y \in X} \lim_{k\to\infty} T_\varphi^k v(y) Q^t(y|x, \varphi) = T_\varphi v_\varphi(x).$$

Esto es, v_φ es un punto fijo de T_φ. En resumen

$$\lim_{k\to\infty} T_\varphi^k v(x) = v_\varphi(x), \text{ y } v_\varphi(x) = T_\varphi v_\varphi(x).$$

Ahora, para demostrar la unicidad, tenga en cuenta que si $v(x) = T_\varphi v(x)$ entonces, $v(x) = T_\varphi^k v(x)$ para todo k, y por lo tanto, $v(x) = \lim_{k\to\infty} T_\varphi^k v(x) = v_\varphi(x)$.

(b) La prueba será por contradicción, entonces supongamos que φ es impropia. Sea $x \in X$, v una función que satisface $v(x) = T_\varphi v(x)$, por Lema 2.4.4, se tiene

$$v(x) = T_\varphi^k v(x) = \sum_{t=0}^{k-1} \sum_{y \in X} C_\varphi(y, \varphi) Q^t(y|x, \varphi) + \sum_{y \in X} v(y) Q^k(y|x, \varphi).$$

Si φ fuera impropia, entonces por la Observación 2.4.3 a) la suma divergerá a ∞ cuando $k \to \infty$, lo cual es una contradicción a que v es acotada. □

Teorema 2.4.1. *La función de Energía Libre óptima v^* es la única función que satisface la ecuación: $v^* = Tv^*$.*

Demostración. En primer lugar, demostraremos que T tiene a lo más un punto fijo. Supongamos que v y v' son dos puntos fijos, luego seleccionamos φ y $\varphi' \in \Phi$ tales que:

$$\varphi(\cdot|x) \in arg\min_{\mu\in\Phi}[T_\mu v](x),$$

$$\varphi'(\cdot|x) \in arg\min_{\mu\in\Phi}[T_\mu v'](x),$$

por lo tanto, $T_\varphi v = v$ y $T_{\varphi'}v' = v'$. Se sabe, por la Proposición 2.4.1 b), que φ y φ' son propias, además $v = v_\varphi$ y $v' = v_{\varphi'}$, así, $v = Tv = T^k v \leq T_\varphi^k v$ para $k \geq 1$.
Tomando $k \to \infty$ y usando la Proposición 2.4.1 a), se obtiene

$$v \leq \lim_{k\to\infty} T_{\varphi'}^k v = v_{\varphi'} = v'.$$

Similarmente, $v' \leq v$. Por lo tanto, $v' = v$ y $v = Tv$ tiene un único punto fijo.

Probaremos ahora que la función óptima de energía libre v^* satisface la ecuación $v^* = Tv^*$. Supongamos que φ^* existe, por lo tanto, el mínimo en la función de valor se alcanza en el conjunto Π^P, es decir, es posible cambiar el ínfimo por mínimo.

Sea $\varphi^* \in \Pi^P$ la política óptima, $\varphi^* = arg\,\text{mín}_\varphi\, v_\varphi$. En consecuencia, para cualquier política $\varphi \in \Phi$, $v^* \leq v_\varphi$.
Aplicando el operador T a v^* se obtiene

$$Tv^* = \min_{\varphi\in\Phi} T_\varphi v^* \leq T_{\varphi^*} v^* = v^*. \tag{2.13}$$

Ahora, seleccionamos una política $\mu \in \Phi$ tal que $T_\mu v^* = Tv^*$, entonces por (2.13)

$$v^* \geq Tv^* = T_\mu v^*.$$

Así, para cualquier $k \geq 1$, $T_\mu v^* \geq T_\mu^k v^*$, y tomando $k \to \infty$ tenemos por Proposición 2.4.1 a)

$$v^* \geq \lim_{k\to\infty} T_\mu^k v^* = v_\mu,$$

además, $v^* \leq v_\varphi$ para cualquier política φ, en particular para μ, por lo tanto, $v^* = v_\mu$, y

$$v^* = v_\mu = T_\mu v_\mu = T_\mu v^* = Tv^*,$$

en la última igualdad usamos que $T_\mu v^* = Tv^*$. En consecuencia v^* es un punto fijo de T. □

Lema 2.4.6. *Sea $v_{k+1} = Tv_k,\ k = 1, 2, ...,$ con $v_0 = 0$. La sucesión de funciones $\{v_k\}$ converge al único punto fijo v^* de T.*

Demostración. Sea $\varphi^* \in \Pi^P$ la política óptima, por lo tanto $v^* = v_{\varphi^*}$, $Tv = T_{\varphi^*}v$ para $v \in M^+(X)$ y

$$T_\varphi^k v_0 = T_\varphi^{k-1}[T_\varphi v_0] = T_\varphi^{k-1} v_1 = ... = Tv_k. \tag{2.14}$$

Tomando $k \to \infty$ en (2.14), se cumple que

$$v_{\varphi^*} = \lim_{k\to\infty} T_\varphi^k v_0 = \lim_{k\to\infty} v_{k+1}.$$

□

El siguiente teorema es consecuencia del Lema que se presenta en [33] haciendo aquí los respectivos cambios al caso infinito.

Teorema 2.4.2. *Sea T el operador de programación dinámica, las siguientes identidades se cumplen:*

a) La política óptima esta dada por

$$\varphi^*(a|x) = \frac{1}{nZ(x,\beta)} \exp\left[\beta r(x,a) - \sum_{y\in X} v(y)Q(y|x,a)\right],$$

$\forall x \in X$, $\forall a \in A(x)$ *donde $Z(x,\beta)$ es la función*

$$Z(x,\beta) = \sum_{a\in A(x)} \frac{1}{n} \exp\left[\beta r(x,a) - \sum_{X} v(y)Q(y|x,a)\right]. \tag{2.15}$$

b) $Tv(x) = -\log Z(x,\beta)$.

Demostración. a) La minimización del operador T sobre el conjunto de distribuciones normalizadas se puede plantear como un problema sin restricciones a través del siguiente Lagrangiano

$$L[\varphi,\lambda_x] = \sum_{A(x)} \varphi(a|x)\left[\log n\varphi(a|x) - \beta r(x,a) + \sum_{y\in X} v(y)Q(y|x,a)\right] + \lambda_x \sum_{A(x)} \varphi(a|x),$$

donde λ_x es el multiplicador de Lagrange. Sean a y x fijos, tomando la derivada de L con respecto a $\varphi(a|x)$ se obtiene

$$\frac{\partial L}{\partial \varphi(a|x)} = \log n\varphi(a|x) - \beta r(x,a) + \sum_{y\in X} v(y)Q(y|x,a) + \lambda_x + 1, \tag{2.16}$$

igualando a cero y despejando a $\varphi(a|x)$ obtenemos

$$\varphi(a|x) = \frac{1}{n} \exp\left[\beta r(x,a) - \sum_{y\in X} v(y)Q(y|x,a) - \lambda_x - 1\right], \tag{2.17}$$

sumando sobre $A(x)$ se tiene

$$\sum_{A(x)} \varphi(a|x) = \sum_{A(x)} \frac{1}{n} \exp\left[\beta r(x,a) - \sum_{y\in X} v(y)Q(y|x,a) - \lambda_x - 1\right], \tag{2.18}$$

como $\sum_{A(x)} \varphi(a|x) = 1$, despejamos λ_x de (2.18)

$$\lambda_x = 1 + \log Z(x, \beta), \tag{2.19}$$

donde $Z(x, \beta)$ es como en (2.15). Por lo tanto, sustituyendo (2.19) en (2.17)

$$\varphi^*(a|x) = \frac{1}{nZ(x,\beta)} \exp\left[\beta r(x,a) - \sum_{y\in X} v(y)Q(y|x,a)\right]. \tag{2.20}$$

b) Observe que en $\varphi^*(a|x)$ se alcanza el mínimo en T y como

$$\log n\varphi^*(a|x) = \beta r(x,a) - \sum_{y\in X} v(y)Q(y|x,a) - \log Z(x,\beta). \tag{2.21}$$

entonces

$$\begin{aligned} Tv(x) &= \sum_{a\in A(x)} \varphi^*(a,x)[-\log Z(x,\beta)] \\ &= -\log Z(x,\beta) \end{aligned}.$$

□

Observación 2.4.4. *En resumen, del Lema 2.4.6 y Teorema 2.4.2, tenemos las relaciones*

$$v_{i+1}(x) = -\log Z_i(x,\beta), \tag{2.22}$$

$$Z_i(x,\beta) = \sum_{a\in A(x)} \frac{1}{n} \exp\left[\beta r(x,a) - \sum_{y\in X} v_i(y)Q(y|x,a)\right], \ \forall x \in X. \tag{2.23}$$

2.5. Ejemplos

Presentamos un ejemplo numérico del problema de la ruta más corta determinista y estocástica utilizando la teoría desarrollada en este capítulo, además incluimos el ejemplo del mundo de la rejilla en espacios numerables y en subconjuntos de $\mathbb{R}$.

2.5.1. El Problema Determinista de la Ruta más Corta

Analizamos aquí la dinámica del problema de la Ruta más Corta con estados infinitos numerable, costos de información y un estado de destino. En cada paso, el agente elige entre dos acciones que corresponden a las dos direcciones posibles que pueden tomar $\{\leftarrow, \rightarrow\}$ con probabilidad $\varphi \in \Phi$. La función de transición de estado indica que el agente se mueve (determinísticamente) al

nodo adyacente que corresponde a la acción elegida hasta que se alcanza el estado de destino. Sean $R_1, R_2 \in \mathbb{R}^-$. El agente paga un costo C_φ (2.6) donde la recompensa en cada paso es R_1 si la acción indica que el agente se aleja del destino, o R_2 si el movimiento lo acerca al estado de destino.

De esta manera, los elementos del modelo son:

- Sea $g \in \mathbb{N}$ fijo. Definimos el espacio de estados (posición del agente) como $X := \{k \in \mathbb{N} | k \geq g\}$. En lo siguiente $g \in X$ y g denotará el estado destino (goal).
- El espacio de acciones es $A = \{-1, 1\}$, $x \in X$, donde la acción -1 indica que el agente toma la dirección $\leftarrow$, y 1 la dirección $\rightarrow$.
- La función de recompensa está dada por

$$r(x,a) = \begin{cases} 0, & \text{si } x = g,\ a \in A, \text{ o} \\ R_1, & \text{si } x \in X - \{g\},\ a = -1, \text{ o} \\ R_2, & \text{si } x \in X - \{g\},\ a = 1. \end{cases}$$

- La función de transición Q está dada de la siguiente manera: $Q(x - 1|x, -1) = 1$, $Q(x + 1|x, 1) = 1$ si $x \neq g$, $Q(g|g, a) = 1$, $a \in A$.

Además, dado que hay dos acciones posibles, $\rho(a|x) = 1/2$ para todos $x \in X$, $a \in A$ y estamos interesados en minimizar $v(\varphi, x, \beta)$, para $\varphi \in \Phi$, $x \in X$, $\beta > 0$.

Observación 2.5.1. *Para este ejemplo, la Condición 2.4.1 se cumple directamente del modelo.*

Lema 2.5.1. *Para el Ejemplo 2.5.1 la Condición 2.4.2 se cumple.*

Demostración. Considere la política propia: $\varphi(-1|x) = 1$, $\forall x \in X$, entonces se verifica que

$$\sum_{t=0}^{\infty} P_x^\varphi(x_t \neq g) = \sum_{t=0}^{x-g-1} P_x^\varphi(x_t = x - t) < \infty.$$

□

Teorema 2.5.1. *Para $\beta \in (0, \infty)$ fijo, se cumple la siguiente relación*

$$Z_i(x,\beta) = \begin{cases} 1, & si\ x = g, \\ \frac{1}{2}\left\{e^{\beta R_1 - v_i(x-1,\beta)} + e^{\beta R_2 - v_i(x+1,\beta)}\right\}, & si\ x \leq i + g + 1, \\ \frac{1}{2}\left\{e^{\beta R_1 - v_i(i+g,\beta)} + e^{\beta R_2 - v_i(i+g+2,\beta)}\right\}, & si\ x > i + g + 1, \end{cases}$$

para $i = 0, 1, 2, ...$, donde $v_0(x, \beta) = 0$, $x \in X$ y v_i es como en (2.22).

Demostración. Probaremos el resultado por inducción. Para $i = 0$ tenemos

$$Z_0(x,\beta) = \frac{1}{2}\{\exp^0 + \exp^0\} = 1,$$

debido a que $r(g,a) = 0$, $a = -1, 1$. Y para $x > g + 1$,

$$Z_0(x,\beta) = \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2}\right\}.$$

Asumimos ahora que

$$Z_{i-1}(x,\beta) = \begin{cases} 1, & \text{si } x = g, \\ \frac{1}{2}\left\{e^{\beta R_1 - v_{i-1}(x-1,\beta)} + e^{\beta R_2 - v_{i-1}(x+1,\beta)}\right\}, & \text{si } x \leq i+g, \\ \frac{1}{2}\left\{e^{\beta R_1 - v_i(i+g-1,\beta)} + e^{\beta R_2 - v_i(i+g+1,\beta)}\right\}, & \text{si } x > i+g. \end{cases} \tag{2.24}$$

Probaremos el resultado para i. En primer lugar, observe que mediante el Lema 2.4.6 y el Teorema 2.4.2 $v_i(x,\beta) = -\log Z_{i-1}(x,\beta)$, entonces

$$v_i(x,\beta) = \begin{cases} 0, & \text{si } x = g, \\ -\log\left(\frac{1}{2}\left\{e^{\beta R_1 - v_{i-1}(x-1,\beta)} + e^{\beta R_2 - v_{i+1}(x-1,\beta)}\right\}\right), \\ \qquad \text{si } x \leq i+g, \\ -\log\left(\frac{1}{2}\left\{e^{\beta R_1 - v_{i-1}(i+g-1,\beta)} + e^{\beta R_2 - v_{i-1}(i+g+1,\beta)}\right\}\right), \\ \qquad \text{si } x > i+g. \end{cases}$$

Usando (2.23) para $x = g$,

$$Z_i(g,\beta) = 1,$$

como

$$\sum_{m=g}^{\infty} v_i(m,\beta)Q(m|g,-1) = v_i(g,\beta) = 0,$$

$$\sum_{m=g}^{\infty} v_i(m,\beta)Q(m|g,1) = v_i(g,\beta) = 0.$$

Para $x \leq i + g + 1$,

$$Z_i(x,\beta) = \frac{1}{2}\left\{e^{\beta R_1 - v_i(x-1,\beta)} + e^{\beta R_2 - v_i(x+1,\beta)}\right\},$$

debido a que

$$\sum_{m=g}^{\infty} v_i(m,\beta)Q(m|x,-1) = v_i(x-1,\beta),$$

$$\sum_{m=g}^{\infty} v_i(m,\beta)Q(m|x,1) = v_i(x+1,\beta),$$

y si $x > i + g + 1$, por (2.24), se sigue que

$$\begin{aligned} Z_i(x,\beta) &= \frac{1}{2}\left\{e^{\beta R_1 - v_i(x-1,\beta)} + e^{\beta R_2 - v_i(x+1,\beta)}\right\} \\ &= \frac{1}{2}\left\{e^{\beta R_1 - v_i(i+g,\beta)} + e^{\beta R_2 - v_i(i+g+2,\beta)}\right\}, \end{aligned}$$

y esto concluye la prueba. □

Finalmente, usando los resultados del Teorema 2.24 proponemos el Algoritmo 1, el cual calcula v^* por medio de iteraciones del operador de programación dinámica $v^* = \lim_{i\to\infty} Tv_i = \lim_{i\to\infty}(-\log Z_i)$ y devuelve la política óptima φ^* para un estado inicial fijo, un modelo de control dado y el parámetro de compensación β.

Algorithm 1 Obtener los v' y φ^* óptimos con x fijo, β y $\varepsilon > 0$ dados.

Require: x, $v'(m,\beta) = 0$, $\beta > 0$, $m \in X$, $i = 1$

repeat
  $v(1,\beta) \leftarrow 0$
  for $m = g + 1 : i + g + 1$ **do**
    $Z(m,\beta) \leftarrow \frac{1}{2}\left\{e^{\beta R_1 - v'(m-1,\beta)} + e^{\beta R_2 - v'(m+1,\beta)}\right\}$
    $v(m,\beta) \leftarrow -\log Z(m,\beta)$
  end for
  $Z(i+g+2,\beta) \leftarrow \frac{1}{2}\left\{e^{\beta R_1 - v'(i+g,\beta)} + e^{\beta R_2 - v'(i+g+2,\beta)}\right\}$
  $v(i+g+2,\beta) \leftarrow -\log Z(i+g+2,\beta)$
  $i++$
until $|v(x,\beta) - v'(x,\beta)| < \varepsilon$
$v' \leftarrow v$
$\varphi^*(-1|x) \leftarrow \frac{1}{2Z(x,\beta)} \exp\{\beta R_1 - v'(x-1,\beta)\}$
$\varphi^*(1|x) \leftarrow \frac{1}{2Z(x,\beta)} \exp\{\beta R_1 - v'(x+1,\beta)\}$
return φ^* y v'

Algorithm 2 Obtener el valor óptimo final v^*

Paso 1. Calcular v' y φ^* con x fijo.

Paso 2. Eligir a con probabilidad φ^*.

Paso 3. Calcular el siguiente x con probabilidad Q.

Paso 4. Estimar $Cost = \log 2 * \varphi^*(a|x) - \beta r(x,a) + Cost$.

Paso 5. Estimar $V = r(x,a) + V$.

Paso 6. Estimar $J = \log 2 * \varphi^*(a|x) + J$.

Paso 7. Repetir hasta alcanzar g.

Paso 8. Obtener v^* como el costo final.

Los Algoritmos 1 y 2 son usados para simular 100 trayectorias del proceso estocástico con $g = 1$, $x_0 = 5$, $R_1 = -10$, $R_2 = -20$ y $\varepsilon = 0.0001$ para diferentes valores de $\beta > 0$. La Tabla 2.1 ilustra los valores de v^* y sus correspondientes recompensas V y costos de información J.

Tabla 2.1: Valores de energía libre óptimos.

β	v^*	V	J
0.1	8.721847	-61.500000	3.998539
1	62.772589	-60.000000	4.00000
10	602.772589	-60.000000	3.972589

Observamos en la Tabla 2.1 que los valores óptimos de los costos de información y de las recompensas, son similares para los diferentes valores de β, siendo solo el costo óptimo v^* el que se ve afectado.

2.5.2. El Problema Estocástico de la Ruta más Corta

Ahora, analizamos *el problema de la Ruta más Corta estocástico*. Consideramos el modelo del Ejemplo 2.5.1 pero con $g = 1$ y la ley de transición de la siguiente manera

$$Q(m|x,1) = \begin{cases} \left(\frac{1}{2}\right)^m, & \text{si } m > x+1, \\ 1 - \left(\frac{1}{2}\right)^{x+1}, & \text{si } m = x+1 \\ 0, & \text{cualquier otro caso} \end{cases}$$

$$Q(m|x,-1) = \begin{cases} \left(\frac{1}{2}\right)^m, & \text{si } m < x-1, \\ \left(\frac{1}{2}\right)^{x-2}, & \text{si } m = x-1 \\ 0, & \text{cualquier otro caso} \end{cases}$$

y

$$Q(1|1,-1) = Q(1|1,1) = 1.$$

Observación 2.5.2. *En este ejemplo, al igual que en su contraparte determinista, la Condición 2.4.1 se cumple directamente del modelo.*

Lema 2.5.2. *Para este ejemplo la Condición 2.4.2 se cumple.*

Demostración. Considere la política propia: $\varphi(-1|x) = 1$, $\forall x \in X$, entonces se verifica que

$$\sum_{t=0}^{\infty} P_x^{\varphi}(x_t \neq 1) = \sum_{t=0}^{x-1} \sum_{m=2}^{x-1-t} Q^t(m|x,-1) < \infty.$$

Por lo tanto, la Condición 2.4.2 se cumple. □

Observación 2.5.3. *Sea $\varphi, \psi \in \mathbb{F}$, con $\varphi(x) = 1$, $\forall x \in X$ y $\psi(x) = -1$, $\forall x \in X$ con $1, -1 \in A$. Observe que según el modelo, se obtiene que:*

- $E_x^{\varphi}[v_i(x', \beta)] = \sum_{m=1}^{\infty} v_i(m, \beta) Q(m|x, 1)$

$$= \begin{cases} v_i(1,\beta), & si\ x = 1, \\ \sum_{m=x+2}^{\infty} (\frac{1}{2})^m v_i(m,\beta) + v_i(x+1,\beta) \sum_{m=1}^{x+1} (\frac{1}{2})^m, & si\ x > 1, \end{cases}$$

$$= \begin{cases} 0, & si\ x = 1, \\ \sum_{m=x+2}^{\infty} (\frac{1}{2})^m v_i(m,\beta) + v_i(x+1,\beta)[1 - (\frac{1}{2})^{x+1}], & si\ x > 1. \end{cases}$$

- $E_x^{\psi}[v_i(x', \beta)] = \sum_{m=1}^{\infty} v_i(m, \beta) Q(m|x, -1)$

$$= \begin{cases} v_i(1,\beta), & si\ x = 1, 2, \\ \sum_{m=1}^{x-2} (\frac{1}{2})^m v_i(m,\beta) + v_i(x-1,\beta) \sum_{m=x-1}^{\infty} (\frac{1}{2})^m, & si\ x > 2, \end{cases}$$

$$= \begin{cases} 0, & si\ x = 1, 2, \\ \sum_{m=1}^{x-2} (\frac{1}{2})^m v_i(m,\beta) + v_i(x-1,\beta)(\frac{1}{2})^{x-2}, & si\ x > 2. \end{cases}$$

Teorema 2.5.2. *Para $\beta > 0$ fijo y Z_i como (2.23), los siguientes resultados se cumplen*

$$Z_i(x,\beta) = \begin{cases} 1, & si\ x = 1, \\ \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2 - f(\varphi,i,2)}\right\}, & si\ x = 2, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - f(\varphi,i,x)}\right\}, & si\ 2 < x \leq i+2, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - v_i(i+2,\beta)}\right\}, & si\ x > i+2\ . \end{cases}$$

para $i = 0, 1, 2, ...,$ donde

$$f(\varphi,i,x) := \sum_{m=x+2}^{i+4} \left(\frac{1}{2}\right)^m v_i(m,\beta) + v_i(i+2,\beta)\left(\frac{1}{2}\right)^{i+4} + v_i(x+1,\beta)\left[1 - \left(\frac{1}{2}\right)^{x+1}\right],$$

$$f(\psi,i,x) := \sum_{m=1}^{x-2} \left(\frac{1}{2}\right)^m v_i(m,\beta) + v_i(x-1,\beta)\left(\frac{1}{2}\right)^{x-2}.$$

Demostración. Probaremos el resultado por inducción. De esta forma por (2.23), se cumple que,

$$Z_0(x,\beta) = \begin{cases} 1, & \text{si } x = 1, \\ \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2}\right\}, & \text{si } x \geq 2. \end{cases}$$

como $r(1,a) = 0$, $a = -1, 1$ y $v_0(x,\beta) = 0$, $x \in X$, $\beta > 0$.
Ahora, supongamos que

$$Z_{i-1}(x,\beta) = \begin{cases} 1, & \text{si } x = 1, \\ \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2 - f(\varphi,i-1,2)}\right\}, & \text{si } x = 2, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i-1,x)} + e^{\beta R_2 - f(\varphi,i-1,x)}\right\}, & \text{si } 2 < x \leq i+1, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i-1,x)} + e^{\beta R_2 - v_{i-1}(i+1,\beta)}\right\}, & \text{si } x > i+1. \end{cases}$$

Entonces, probaremos el resultado para i. En primer lugar, observe que por (2.22), $v_i(x,\beta) = -\log Z_{i-1}(x,\beta)$.
Usando (2.23), para $x = 1$,

$$Z_i(1,\beta) = 1,$$

y, para $x = 2$,

$$Z_i(2,\beta) = \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2 - f(\varphi,i,2)}\right\}.$$

Entonces, para $2 < x \leq i+2$

$$\begin{aligned} E_x^{\varphi}[v_i(x',\beta)] &= \sum_{m=x+2}^{\infty}\left(\frac{1}{2}\right)^m v_i(m,\beta) + v_i(x+1,\beta)\sum_{m=1}^{x+1}\left(\frac{1}{2}\right)^m \\ &= \sum_{m=x+2}^{i+4}\left(\frac{1}{2}\right)^m v_i(m,\beta) + \sum_{m=i+5}^{\infty} v_i(i+2,\beta)\left(\frac{1}{2}\right)^m + v_i(x+1,\beta)\sum_{m=1}^{x+1}\left(\frac{1}{2}\right)^m \\ &= \sum_{m=x+2}^{i+4}\left(\frac{1}{2}\right)^m v_i(m,\beta) + v_i(i+2,\beta)\left(\frac{1}{2}\right)^{i+4} + v_i(x+1,\beta)\left[1 - \left(\frac{1}{2}\right)^{x+1}\right]. \end{aligned}$$

Ahora, para $x > i+2$, ($x+2 > i+1$ y $x+1 > i+1$),

$$\begin{aligned} E_x^{\varphi}[v_i(x',\beta)] &= \sum_{m=x+2}^{\infty}\left(\frac{1}{2}\right)^m v_i(m,\beta) + v_i(x+1,\beta)\left[1 - \left(\frac{1}{2}\right)^{x+1}\right] \\ &= \sum_{m=x+2}^{\infty}\left(\frac{1}{2}\right)^m v_i(i+2,\beta) + v_i(i+2,\beta)\left[1 - \left(\frac{1}{2}\right)^{x+1}\right] \\ &= v_i(i+2,\beta), \end{aligned}$$

en consecuencia,

$$Z_i(x,\beta) = \begin{cases} 1, & \text{si } x = 1, \\ \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2 - v_i(i,\beta)}\right\} & \text{si } x = 2, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - f(\varphi,i,x)}\right\}, & \text{si } 2 < x \leq i+2, \\ \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - v_i(i+2,\beta)}\right\}, & \text{si } x > i+2\ . \end{cases}$$

□

Algorithm 3 Obtaining optimal v' and φ^* with x fixed, β and $\varepsilon > 0$ given.

Require: x, $v'(m,\beta) = 0$, $m \in X$, $\beta > 0$, $i = 1$

repeat
  $Z(1,\beta) = 1$
  $v(1,\beta) = 0$
  $Z(2,\beta) = \frac{1}{2}\left\{e^{\beta R_1} + e^{\beta R_2 - f(\varphi,i,2)}\right\}$
  $v(2,\beta) - \log Z(2,\beta)$
  for $m = 3 : i+2$ **do**
    $Z(m,\beta) \leftarrow \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - f(\varphi,i,x)}\right\}$
    $v(m,\beta) \leftarrow -\log Z(m,\beta)$
  end for
  for $m = i+3 : i+x+4$ **do**
    $Z(m,\beta) \leftarrow \frac{1}{2}\left\{e^{\beta R_1 - f(\psi,i,x)} + e^{\beta R_2 - v_i(i+2,\beta)}\right\}$
    $v(m,\beta) \leftarrow -\log Z(m,\beta)$
  end for
  $i++$
until $|v(x,\beta) - v'(x,\beta)| < \varepsilon$
$v' \leftarrow v$
$\varphi^*(-1|x) \leftarrow \frac{1}{2Z(x,\beta)} \exp\{\beta R_1 - f(\psi,i,x)\}\}$
$\varphi^*(1|x) \leftarrow \frac{1}{2Z(x,\beta)} \exp\{\beta R_2 - f(\varphi,i,x)\}\}$
return φ^* and v'

La Tabla 2.2 ilustra el valor de la energía libre resultante de la simulación de una trayectoria con $x_0 = 5$, $R_1 = -10$, $R_2 = -20$ y $\varepsilon = 0.0001$ para diferentes valores de $\beta > 0$ usando los Algoritmos 2 y 3.

Tabla 2.2: Valores de energía libre óptimos.

β	v^*	V	J
0.1	13.723763	-106.20	3.103763
1	67.947178	-60.00	7.947178
10	611.18259	-60.00	11.18259
50	3013.49655	-60.00	13.49655

De la Tabla 2.2 podemos concluir que para valores pequeños de β se obtienen menores recompensas óptimas pero menores costos de información, por otro lado, si el valor de β se va tomando más grande, el valor óptimo de las recompensas se estabiliza pero los costos de información crecen.

2.5.3. El Mundo de la Rejilla en Espacios Numerables

El siguiente ejemplo está motivado por el problema clásico llamado *El mundo de la rejilla*, que se presenta generalmente para mostrar la solución del problema de control óptimo en espacios de estados finitos y ha sido trabajado en [33] y [40].

El mundo de la rejilla es un sistema donde un agente se coloca en una rejilla y quiere alcanzar una determinada posición final; en cada paso, el agente elige una dirección entre las posibles con cierta probabilidad, inmediatamente, el agente se mueve (determinísticamente) a la celda que corresponde de acuerdo a la acción elegida o permanece en el mismo lugar si la acción indica ir hacia una pared. El objetivo es elegir una ruta que genere las mayores recompensas.

Analizamos ahora la dinámica de un sistema donde la ubicación del agente en un punto dado es importante, digamos en una cuadrícula de $\{\frac{1}{u}\}_{u\in\mathbb{N}} \times \{\frac{1}{v}\}_{v\in\mathbb{N}}$, con un estado de destino $g = (1, 1)$. En cada paso, el agente elige entre cuatro acciones que corresponden a las cuatro direcciones posibles que pueden tomar $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$ con probabilidad $\varphi \in \Phi$, que es desconocida por el agente. La ley de transición indica que el agente se mueve (de manera determinista) a la celda adyacente que corresponde a la acción elegida, a menos que la acción indique ir a la pared, entonces permanecerá en el mismo lugar. El agente paga un costo C_φ (2.6) donde la recompensa es $R_1 \in \mathbb{R}^-$ si la acción indica que el agente se aleja del destino o va hacia una pared, o $R_2 \in \mathbb{R}^-$ si el movimiento lo acerca a g.

Ahora, los componentes del modelo son:

- $X = \left\{ \left(\frac{1}{x_1}, \frac{1}{x_2}\right) | x_1, x_2 \in \mathbb{N} \right\}$.
- El espacio de acciones $A = A\left(\frac{1}{x_1}, \frac{1}{x_2}\right) = \{1, 2, 3, 4\}$, donde $1, 2, 3, 4$ son las direcciones que puede tomar el agente, asignando a la dirección $\leftarrow$ la acción 1, $\rightarrow$ la acción 2 y $\downarrow, \uparrow$ las acciones 3 y 4 respectivamente.
- La función de recompensa está dada por

$$r\left(\left(\frac{1}{x_1}, \frac{1}{x_2}\right), 1\right) = \begin{cases} 0, & \text{si } x_1 = 1 \text{ y } x_2 = 1, \\ R_1, & \text{si } x_1 > 1 \text{ o } x_2 > 1. \end{cases}$$

$$r\left(\left(\frac{1}{x_1}, \frac{1}{x_2}\right), 2\right) = \begin{cases} 0, & \text{si } x_1 = 1 \text{ y } x_2 = 1, \\ R_1, & \text{si } x_1 = 1 \text{ y } x_2 > 1, \\ R_2, & \text{si } x_1 > 1 \text{ y } x_2 \geq 1. \end{cases}$$

$$r\left(\left(\frac{1}{x_1}, \frac{1}{x_2}\right), 3\right) = \begin{cases} 0, & \text{si } x_1 = 1 \text{ y } x_2 = 1, \\ R_1, & \text{si } x_1 > 1 \text{ o } x_2 > 1. \end{cases}$$

$$r\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),4\right)=\begin{cases}0, & \text{si } x_1=1 \text{ y } x_2=1,\\ R_1, & \text{si } x_1>1 \text{ y } x_2=1,\\ R_2, & \text{si } x_1\geq 1 \text{ y } x_2>1.\end{cases}$$

- La función de transición Q es
  $Q\left(\left(\frac{1}{x_1+1},\frac{1}{x_2}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),1\right)=Q\left(\left(\frac{1}{x_1},\frac{1}{x_2+1}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),3\right)=1$ si $x_1>1$ y $x_2>1$ o $Q((1,1)|(1,1),1)=Q((1,1)|(1,1),3)=1$.
  $Q\left(\left(\frac{1}{x_1-1},\frac{1}{x_2}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),2\right)=1$ si $x_1>1$ y $x_2\in\mathbb{N}$.
  $Q\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),2\right)=1$ si $x_1=1$ y $x_2\in\mathbb{N}$.
  $Q\left(\left(\frac{1}{x_1},\frac{1}{x_2-1}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),4\right)=1$ si $x_1>1$ y $x_2\in\mathbb{N}$.
  $Q\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right)|\left(\frac{1}{x_1},\frac{1}{x_2}\right),4\right)=1$ si $x_1=1$ y $x_2=1$.

Además, debido a que hay cuatro acciones posibles, $\rho\left(a|\left(\frac{1}{x_1},\frac{1}{x_2}\right)\right)=1/4$ para todo $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X$ y estamos interesados en minimizar $v\left(\varphi,\left(\frac{1}{x_1},\frac{1}{x_2}\right),\beta\right)$, para $\varphi\in\Phi$, $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X$, $\beta>0$.

Para calcular Z_i y usar el algoritmo 2, estudiamos el comportamiento de Z dada una partición de X: $X_I\cup X_{II}\cup X_{III}\cup X_{IV}$ donde:

$$\begin{aligned}X_I&=\{(1,1)\},\\ X_{II}&=\left\{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X|x_1>1,\,x_2=1\right\},\\ X_{III}&=\left\{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X|x_1=1,\,x_2>1\right\},\\ X_{IV}&=\left\{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X|x_1>1,\,x_2>1\right\}.\end{aligned}\tag{2.25}$$

Teorema 2.5.3. *Para $\beta\in(0,\infty)$ fijo, se cumple la siguiente relación*

$$z_i\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),\beta\right)=\frac{1}{2}\sum_a\exp\left[\beta r\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),a\right)-E^a_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right]\right],$$

para $i=0,1,2,...,$ donde

- $E^1_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right]=v_i\left(\left(\frac{1}{x_1+1},\frac{1}{x_2}\right),\beta\right),\quad\forall\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X.$
- $E^3_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right]=v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2+1}\right),\beta\right),\quad\forall\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X.$

- $E^2_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right]$
$$= \begin{cases} v_i\left(\left(\frac{1}{x_1-1},\frac{1}{x_2}\right),\beta\right), & si\ \left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_{II}\cup X_{IV}, \\ v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),\beta\right), & si\ \left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_I\cup X_{III}. \end{cases}$$

- $E^4_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right]$
$$= \begin{cases} v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2-1}\right),\beta\right), & si\ \left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_{III}\cup X_{IV}, \\ v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),\beta\right), & si\ \left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_I\cup X_{II}. \end{cases}$$

Demostración. Observe que

$$\begin{aligned} E^1_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right] &= \sum_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X} v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right) Q\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right)\mid\left(\frac{1}{x_1},\frac{1}{x_2}\right),1\right) \\ &= v_i\left(\left(\frac{1}{x_1+1},\frac{1}{x_2}\right),\beta\right), \quad \forall\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X, \end{aligned}$$

análogamente,

$$E^3_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right] = v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2+1}\right),\beta\right), \quad \forall\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X.$$

Para $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_{II}\cup X_{IV}$,

$$\begin{aligned} E^2_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right] &= \sum_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X} v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right) Q\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right)\mid\left(\frac{1}{x_1},\frac{1}{x_2}\right),2\right) \\ &= v_i\left(\left(\frac{1}{x_1-1},\frac{1}{x_2}\right),\beta\right), \end{aligned}$$

análogamente para $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_{III}\cup X_{IV}$,

$$\begin{aligned} E^4_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right] &= \sum_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X} v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right) Q\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right)\mid\left(\frac{1}{x_1},\frac{1}{x_2}\right),4\right) \\ &= v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2-1}\right),\beta\right). \end{aligned}$$

Finalmente, para $a=2$ y $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_I\cup X_{III}$, o $a=4$ y $\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X_I\cup X_{II}$,

$$\begin{aligned} E^a_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)}\left[v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right)\right] &= \sum_{\left(\frac{1}{x_1},\frac{1}{x_2}\right)\in X} v_i\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right),\beta\right) Q\left(\left(\frac{1}{x_1'},\frac{1}{x_2'}\right)\mid\left(\frac{1}{x_1},\frac{1}{x_2}\right),a\right) \\ &= v_i\left(\left(\frac{1}{x_1},\frac{1}{x_2}\right),\beta\right), \end{aligned}$$

Por lo tanto, el resultado sigue directamente de (2.23). □

En consecuencia, según el Teorema 2.5.3, podemos usar el Algoritmo 2 propuesto, para varios valores de $\beta > 0$, para calcular el valor de energía libre óptimo estimado que será el promedio de los valores óptimos. En la Tabla 2.3 se ilustran los valores de v^* y sus correspondientes recompensas V y costos de información J obtenidos de 100 trayectorias simuladas con $x_0 = (4, 4)$, $R_1 = -1$, $R_2 = -100$ y $\varepsilon = 0.0001$ para diferentes valores de $\beta > 0$.

Tabla 2.3: Valores de energía libre óptimos.

β	v^*	V	J
0.1	286.9392	-80.07	22.12098
1	2934.287	-86.85	24.27227
10	22943.51	-69.83	19.13358

En la Tabla 2.3 podemos observar que valores pequeños de β tienen recompensas óptimas y costos óptimos similares, pero para $\beta = 10$ ambos valores disminuyen considerablemente.

2.5.4. El Mundo de la Rejilla en $\mathbb{R}$

Consideremos aquí, el ejemplo *El mundo de la rejilla* en el cual, exploramos la relación entre valor e información en un sistema donde nos importa la ubicación exacta del agente en un área determinada, por ejemplo, en una tabla de 5×5, con un evento objetivo, al que llamamos G. Para una mejor visualización, cuadriculamos el área y el evento objetivo se representa como el área de la celda descrita por $[4, 5] \times [4, 5]$. En cada paso, el agente elige entre 8 acciones que corresponden a las 8 direcciones posibles que pueden tomar $\{\leftarrow, \rightarrow, \uparrow, \downarrow, \nearrow, \searrow, \nwarrow, \swarrow\}$ con probabilidad φ. La función de transición de estado indica que el agente se mueve de manera uniforme a la celda adyacente que corresponde a la acción elegida, a menos que la acción indique ir a la pared, entonces permanecerá en el mismo lugar. Tenga en cuenta que es importante la posición que el agente toma dentro del área de cada celda. El agente paga un costo C_φ como se definió en (2.6) donde $R = -1$ en cada paso, o $R = -100$ si la acción indica moverse hacia una pared.

De esta manera, los elementos del modelo son:

- El espacio de estados (posición del agente) $X = [0, 5] \times [0, 5]$.
- El espacio de acciones igual a las acciones admisibles (las direcciones que puede tomar) $A = A(x) = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

Tabla 2.4: Valores de α y γ.

a	1	2	3	4	5	6	7	8
α_a	-1	0	1	-1	1	-1	0	1
γ_a	1	1	1	0	0	-1	-1	-1

- La función de recompensa está dada por

$$R(x,a) = \begin{cases} 0, & \text{si } x \in G, \forall a, \\ -100, & \text{si } x \in [0,1]\times[4,5] \text{ y } a = 1, \\ & \text{si } x \in [0,5]\times[4,5] \text{ y } a = 2, \\ & \text{si } x \in [4,5]\times[4,5] \text{ y } a = 3, \\ & \text{si } x \in [0,1]\times[0,5] \text{ y } a = 4, \\ & \text{si } x \in [4,5]\times[0,5] \text{ y } a = 5, \\ & \text{si } x \in [0,1]\times[0,1] \text{ y } a = 6, \\ & \text{si } x \in [0,10]\times[0,1] \text{ y } a = 7, \\ & \text{si } x \in [4,5]\times[0,1] \text{ y } a = 8, \\ -1 & \text{en cualquier otro caso.} \end{cases}$$

- Sea $x = (x_1, x_2)$, $[x_1]$ representa la parte entera de x_1 y α y γ como se muestran en la Tabla 2.4. La función de transición Q se asume uniforme en el cuadrado $[[x_1]+\alpha_a, [x_1]+\alpha_a+1]\times[[x_2]+\gamma_a, [x_2]+\gamma_a+1]$ o bien $Q(x|x,a) = 1$ en los casos donde la recompensa es igual a -100, es decir, cuando la acción indica ir hacia una pared.

Además, como hay ocho acciones posibles, $\rho(a|x) = 1/8$ para todo $x \in X$ y nos interesa minimizar $F(\pi, x, \beta)$.

Estudiamos el comportamiento de Z_i conforme se hacen las iteraciones y observamos similitudes en algunos elementos de su dominio, dividiendo de esta manera al espacio X en $X = X_I \cup X_{II} \cup X_{III} \cup X_{IV}$ donde:

$$\begin{aligned} X_I &= [1,4]\times[0,1] \cup [1,4]\times[4,5] \cup [4,5]\times[1,9] \cup [0,1]\times[1,4], \\ X_{II} &= [0,1]\times[4,5] \cup [0,1]\times[0,1] \cup [4,5]\times[0,1], \\ X_{III} &= [1,4]\times[1,4], \\ X_{IV} &= [4,5]\times[4,5], \end{aligned} \tag{2.26}$$

y establecemos el teorema siguiente

Teorema 2.5.4. *Para $\beta > 0$ fijo y Z_i como en (2.23), se cumple*

$$Z_i(x,\beta) = h^{i+1}(x, 1, \beta),$$

para $i = 0, 1, ...$, donde el super índice en h indica composición de funciones, y la función h está dada por:

$$h(x, y, \beta) = \begin{cases} \frac{1}{8}\left\{5e^{-\beta+\log y} + 3e^{-100\beta+\log y}\right\} & si\ x \in X_I, \\ \frac{1}{8}\left\{3e^{-\beta+\log y} + 5e^{-100\beta+\log y}\right\} & si\ x \in X_{II}, \\ e^{-\beta+\log y} & si\ x \in X_{III}, \\ 1 & si\ x \in X_{IV}. \end{cases}$$

Demostración. Probaremos por inducción para cada conjunto que conforma a X como se vió en (2.26), que Z_i se puede expresar a través de la función h como lo indica el teorema.
Así, para $x \in X_I$ y $k = 0$, como $v_0(x, \beta) = 0$, $\forall\beta$, para $x \in X$, por lo tanto,

$$\begin{aligned} Z_0(x, \beta) &= \sum_a \exp\left[\beta r(x, a) - \int v_0(x', \beta) Q(dx'|x, a)\right], \\ &= \frac{1}{8}\left\{5e^{-\beta} + 3e^{-100\beta}\right\}, \\ &= h(x, 1, \beta). \end{aligned} \tag{2.27}$$

Suponemos ahora para $i - 1$ que

$$Z_{i-1}(x, \beta) = h^i(x, 1, \beta)$$

y vemos para i que

$$v_i(x, \beta) = -\log Z_{i-1}(x, \beta) = -\log h^i(x, 1, \beta),$$

además,

$$\begin{aligned} Z_i(x, \beta) &= \sum_a \exp\left[\beta r(x, a) - \int v_i(x', \beta) Q(dx'|x, a)\right], \\ &= \frac{1}{8}\left\{5e^{-\beta - v_i(x,\beta)} + 3e^{-100\beta - v_i(x,\beta)}\right\}, \\ &= \frac{1}{8}\left\{5e^{-\beta+\log h^i(x,1,\beta)} + 3e^{-100\beta+\log h^i(x,1,\beta)}\right\}, \\ &= h(x, h^i(x, 1, \beta), \beta), \\ &= h^{i+1}(x, 1, \beta). \end{aligned} \tag{2.28}$$

Para $x \in X_{II}$ la demostración es análoga.
Para $x \in X_{III}$ y $k = 0$, nuevamente $v_0(x, \beta) = 0$, $\forall x$, por lo tanto,

$$\begin{aligned} Z_0(x, \beta) &= \sum_a \exp\left[\beta r(x, a) - \int v_0(x', \beta) Q(dx'|x, a)\right], \\ &= e^{-\beta} \\ &= h(x, 1, \beta). \end{aligned} \tag{2.29}$$

Para i se tiene

$$\begin{aligned} Z_i(x,\beta) &= \sum_a \exp\left[\beta r(x,a) - \int v_i(x',\beta)Q(dx'|x,a)\right], \\ &= e^{-\beta - v_i(x,\beta)} \\ &= e^{-\beta + \log h^i(x,1,\beta)} \\ &= h(x, h^i(x,1,\beta),\beta), \\ &= h^{i+1}(x,1,\beta). \end{aligned} \tag{2.30}$$

Finalmente, para $x \in X_{IV}$. Observemos que este conjunto es en realidad el evento objetivo por lo que al llegar a él, el proceso termina y permanece aquí con recompenza cero, así $v_0(x,\beta) = 0$, $\forall x$ y $Z_0(x,\beta) = e^0 = 1$, en general,

$$\begin{aligned} Z_i(x,\beta) &= \sum_a \exp\left[\beta r(x,a) - \int v_i(x',\beta)Q(dx'|x,a)\right], \\ &= e^{-v_i(x,\beta)} \\ &= e^{\log h^i(x,1,\beta)} \\ &= e^{\log 1} \\ &= h(x, h^i(x,1,\beta),\beta), \\ &= h^{i+1}(x,1,\beta). \end{aligned} \tag{2.31}$$

Por lo tanto, para cada $x \in X$, Z_i se puede expresar a través de la función h como lo establece el teorema.

□

Por consiguiente, de acuerdo con el Teorema 2.5.4, proponemos el Algoritmo 4 que seguido por el Algoritmo 2, calcula para diversos valores de β, el valor óptimo de una trayectoria simulada para un estado inicial dado en base a la función h. El valor de energía libre óptimo estimado será el promedio de los valores óptimos de las trayectorias simuladas.

Algorithm 4 Obtención de v' óptima y π^* con x fijo y β y ε dados.

Require: $x = x_0$, $v'(x,\beta) = 0$, $\forall\beta$, $Z(x,\beta) = 1$

1: **repeat**
2: $v(x,\beta) \leftarrow v'(x,\beta)$
3: $Z(x,\beta) \leftarrow h(x,Z,\beta)$
4: $v(x,\beta) \leftarrow -\log Z(x,\beta)$
5: **until** v converge ($|v' - v| < \varepsilon$)
6: para cada $a \in A$,
7: $\pi^*(a|x) \leftarrow \frac{\rho(a|x)}{Z(x,\beta)} \exp\{\beta r(x,a) - v'(x,\beta)\}$
8: **return** π^* y v'

En la Tabla 2.5 podemos ver el valor de la energía libre resultante de la simulación de 100 trayectorias simuladas con $x_0 = (0, 0)$, $R_1 = -1$, $R_2 = -100$ y $\varepsilon = 0.0001$ para diferentes valores de $\beta > 0$.

Tabla 2.5: Valores de energía libre óptimos.

β	v^*	Número promedio de pasos	V	J
0	1	88	-2629.47	0
1	16921.27	365.11	-365.11	1043.539
10	193969	419.88	-419.88	9510.932

En la Tabla 2.5 podemos observar el número promedio de pasos que fueron necesarios para llegar al estado objetivo, siendo $\beta = 0$ el que da la ruta más rápida y los menores costos de información (J) pero se sacrifican las recompensas (V), siendo éstas muy bajas. Con forme β aumenta, las recompensas óptimas se van estabilizando pero los costos de información crecen considerablemente.

Capítulo 3

Análisis Asintótico de un Sistema de Control Determinista

En este capítulo se aborda el segundo problema principal de esta tesis, el cual es, proporcionar las condiciones necesarias para garantizar la existencia del punto de equilibrio del sistema y la convergencia de la trayectoria óptima determinista a dicho punto.

3.1. Antecedentes

Consideramos un Sistema de Control Determinista. Los problemas de control óptimos deterministas, como ya se mencionó en el Capítulo 1, consisten en determinar una política óptima, es decir, una política que maximice (o minimice) la función objetivo [18], [20]. En este contexto, un problema de interés es analizar el comportamiento asintótico de la trayectoria óptima del sistema.

Investigaciones importantes basan sus resultados en la existencia y caracterización del punto de equilibrio del sistema determinista asociado con el sistema estocástico [24], por ejemplo, en [44] se obtiene la solución del problema estocástico a partir del determinista en tiempo discreto, y en [14] en tiempo continuo. El trabajo está motivado por el estudio del modelo de crecimiento económico tal como se expone en [12], donde se garantiza la estabilidad del sistema caracterizando el punto de convergencia de la trayectoria óptima por medio de la ecuación de Euler. En este trabajo buscamos condiciones para la estabilidad de sistemas más generales.

3.2. Planteamiento del Problema

Teniendo en cuenta los antecedentes de la relación entre los MCM ya mencionados, se plantea un Modelo de Control Determinista como se definió en (1.7), conformado por los siguientes componentes:

$$(X, A, \{A(x) \mid x \in X\}, Q, r) \tag{3.1}$$

donde $X = [0, \infty)$, $A = [0, \infty)$, y Q dada a través de $F : \mathbb{K} \to X$ por $Q(B|x, a) = 1_B(F(x, a))$, $B \in \mathcal{B}(X)$.

Consideremos a la recompensa total descontada como criterio de rendimiento, esta es, para $x \in X$, $\pi \in \Pi_{DM}$ y $\alpha \in (0, 1)$,

$$v(x, \pi) := \sum_{t=0}^{\infty} \alpha^t r(x_t, a_t). \tag{3.2}$$

El objetivo del problema de control es maximizar $\pi \to v(x, \pi)$ en Π_{DM}, para todo $x \in X$, es decir, determinar la política óptima π^* y la función de valor óptimo V.

El ***Segundo Problema*** de interés en esta tesis es realizar un análisis asintótico de la trayectoria óptima del SCD.

El planteamiento anterior da origen a las siguientes preguntas de investigación:

- ¿Cómo se puede garantizar la existencia del punto de equilibrio del sistema?
- ¿Qué condiciones serán necesarias para garantizar la convergencia de la trayectoria óptima del sistema determinista?
- ¿En caso de que exista dicha convergencia, será posible caracterizar el punto de convergencia a través de la ecuación de Euler?

Objetivo General

Proporcionar las condiciones necesarias para garantizar la convergencia de la trayectoria óptima y caracterizar su punto de convergencia.

3.3. Metodología

El método que se propone es demostrar que la trayectoria óptima del sistema de control determinista es monótona y acotada, lo que demuestra que es

convergente; posteriormente, garantizar la existencia del punto de equilibrio, el cual se define más adelante y caracterizarlo a través de la ecuación de Euler. Finalmente, demostrar que se cumple la convergencia al punto de equilibrio. En la Figura 3.1 se puede apreciar la metodología propuesta.

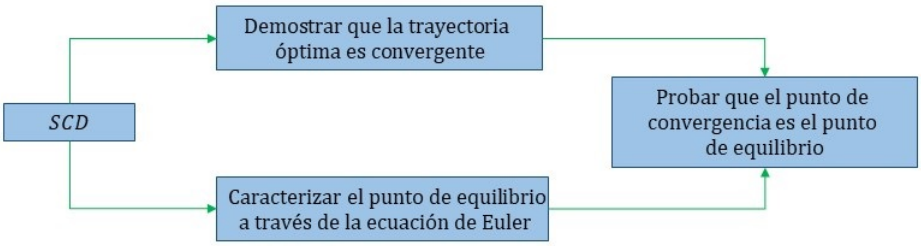

Figura 3.1: Diagrama de solución del segundo problema.

3.4. Análisis Asintótico de un Sistema de Control Determinista Mediante el Enfoque de la Ecuación de Euler

Iniciamos presentando una versión de la Ecuación de Euler para PDMs, de manera particular para el modelo determinista. Consecuentemente se caracteriza el punto de equilibrio del sistema vía dicha ecuación y establecemos algunas propiedades de la trayectoria óptima, que nos ayudarán a garantizar su convergencia. En lo que sigue, la derivada de una función multivariada M con respeto a una variable y, se denotará con M_y, el gradiente de la función M se denota por $DM(x, y)$ y la derivada de una función de una variable h se denotará por h'. Diremos que una función g es de clase C^m sobre U, lo que denotaremos por $g \in C^m(U)$, si para cada punto de u de U existen y son continuas todas las derivadas parciales de g hasta el orden m inclusive.

Tomemos el modelo $(X, A, \{A(x) \mid x \in X\}, Q, r)$ definido en (3.1) con la recompensa total descontada como criterio de rendimiento (ver (3.2)).

Sea v_n la función de iteración de valores (ver Definición 1.2.1) y suponga que se cumple el Teorema de Programación Dinámica 1.2.2, establecemos además

las siguientes condiciones:

Condiciones 3.4.1.

a) Existe una función $H : X \to X$ *tal que* $\sup_{a\in A(X)} F(x,a) \leq H(x)$*, para cada* $x \in X$*. Además, para la función* H*, existe* $x_M \in X$ *tal que* $H(x_M) = x_M$*.*

b) Existe $(y,b) \in \mathbb{K}$ *tal que* $\lim_{(x,a)\to(y,b)} F_x(x,a) = +\infty$*.*

c) $F_x(x,a) \to 0$ *cuando* $(x,a) \to +\infty$*.*

d) $V \in C^2$ *y* $V'' \leq 0$*.*

e) $f \in C^1$ *con* $f' \geq 0$*.*

Observación 3.4.1.

a) Para garantizar la diferenciabilidad de la solución óptima (ver Condición 3.4.1 d) y e)), puede consultar las siguientes referencias: [8], [12]y [18].

b) En [10] puede ver la prueba de la concavidad de la función de valor óptimo bajo la Condición 3.4.1 d).

c) La monotonicidad de la política óptima se puede consultar en [15].

Observación 3.4.2. *Observe que si se cumple el Teorema 1.2.2, se puede caracterizar la solución óptima del problema de control óptimo, a través del enfoque de programación dinámica.*

Entonces, por el Teorema 1.2.2, existe $f \in \mathbb{F}$ tal que:

$$x^*_{t+1} = F(x^*_t, f(x^*_t)) := g(x^*_t), \tag{3.3}$$

$t = 1, 2, ...$, donde $\{x^*_t\}$ representa la trayectoria óptima del proceso de control, estamos interesados en estudiar el comportamiento de (3.3) en su punto de equilibrio (ver Definición 3.4.1).

Definición 3.4.1. *Un punto de equilibrio* $\bar{x} \in X$ *asociado con (3.3), es aquel que satisface la identidad siguiente [23]:*

$$\bar{x} = F(\bar{x}, f(\bar{x})) = g(\bar{x}). \tag{3.4}$$

Teorema 3.4.1. *Si r y F son funciones cóncavas y r, $F \in C^2(\mathbb{K})$ entonces la política óptima, f, satisface la siguiente ecuación funcional, conocida en la literatura como Ecuación de Euler:*

$$r_a(x, f(x)) + \alpha\Delta(F(x, f(x)), f(F(x, f(x))))F_a(x, f(x)) = 0,\ x \in (0, \infty) \quad (3.5)$$

donde $\Delta(x, a) := (r_x - r_a F_x/F_a)(x, a)$, $(x, a) \in \mathbb{K}$, $F_a > 0$.
Recíprocamente, si $f \in \mathbb{F}$ es una política que satisface (3.5) para cada $x \in (0, \infty)$ y

$$\lim_{t\to\infty} \alpha^t \Delta(x_t, f(x_t))x_t = 0, \quad (3.6)$$

entonces f es una política óptima.

Demostración. Sea f la política óptima y como V satisface la EPD (1.13), entonces:

$$V(x) = \max_{a\in A(x)} \{r(x, a) + \alpha V(F(x, a))\},$$

en consecuencia, la condición de primer orden está dada por

$$r_a(x, f(x)) + \alpha V'(F(x, f(x)))F_a(x, f(x)) = 0. \quad (3.7)$$

Por otro lado, como V satisface la EPD (1.2.2):

$$V(x) = r(x, f(x)) + \alpha V(F(x, f(x))),$$

entonces

$$V'(x) = r_x(x, f(x)) + r_a(x, f(x))f'(x) + \alpha V'(F(x, f(x)))[F_x(x, f(x)) + F_a(x, f(x))f'(x)], \quad (3.8)$$

sustituyendo (3.7) en (3.8), se obtiene que

$$V'(x) = r_x(x, f(x)) + \alpha V'(F(x, f(x)))F_x(x, f(x)). \quad (3.9)$$

En consecuencia, de (3.7) y (3.9), se deduce que:

$$\begin{aligned} V'(x) &= r_x(x, f(x)) - \alpha \tfrac{r_a(x,f(x))}{\alpha F_a(x,f(x))} F_x(x, f(x)) \\ &= \left[r_x - \tfrac{r_a F_x}{F_a}\right](x, f(x)) \\ &= \Delta(x, f(x)). \end{aligned} \quad (3.10)$$

Finalmente, sustituyendo (3.10) en (3.7), el resultado se sigue.

Ahora, se probará el recíproco. Sea f una función que satisfaga (3.5) y (3.6) y considere $x \in (0, \infty)$ fijo. Sea $\hat{f} \in \mathbb{F}$ otra función y para $t = 0, 1, ...$, se denotarán las trayectorias de las políticas f y $\hat{f}$ por x_t y $\hat{x}_t$, respectivamente. De manera similar, $a_t = f(x_t)$ y $\hat{a}_t = \hat{f}(\hat{x}_t)$ denotan sus acciones respectivas,

donde $x_0 = \hat{x}_0 = x$ para ambos. Como r es cóncavo y $r \in C^2$, aplicando el Teorema 2.17 en [12], p. 258, se obtiene que

$$\begin{aligned}\sum_{t=0}^{T-1} \alpha^t[r(x_t,a_t) - r(\hat{x}_t,\hat{a}_t)] &\geq \sum_{t=0}^{T-1} \alpha^t[Dr(x_t,a_t)(x_t-\hat{x}_t, a_t-\hat{a}_t)] \\ &= \sum_{t=0}^{T-1} \alpha^t[r_x(x_t,a_t)(x_t-\hat{x}_t) + r_a(x_t,a_t)(a_t-\hat{a}_t)],\end{aligned}$$

para T entero positivo mayor o igual a 1. Como $x_{t+1} = F(x_t,a_t)$ y $\hat{x}_{t+1} = F(\hat{x}_t,\hat{a}_t)$, entonces

$$x_t - \hat{x}_t = F(x_{t-1},a_{t-1}) - F(\hat{x}_{t-1},\hat{a}_{t-1}).$$

Como F es cóncava y $F \in C^2$, resulta que

$$F(x_{t-1},a_{t-1}) - F(\hat{x}_{t-1},\hat{a}_{t-1}) \geq F_x(x_{t-1},a_{t-1})(x_{t-1}-\hat{x}_{t-1}) + F_a(x_{t-1},a_{t-1})(a_{t-1}-\hat{a}_{t-1}),$$

o equivalentemente

$$\hat{a}_{t-1} - a_{t-1} \geq \frac{F_x(x_{t-1},a_{t-1})(x_{t-1}-\hat{x}_{t-1}) - (x_t-\hat{x}_t)}{F_a(x_{t-1},a_{t-1})},$$

en consecuencia, se obtiene que

$$\begin{aligned}&\textstyle\sum_{t=0}^{T-1} \alpha^t[r_x(x_t,a_t)(x_t-\hat{x}_t) + r_a(x_t,a_t)(a_t-\hat{a}_t)] \\ &\geq \textstyle\sum_{t=0}^{T-1} \alpha^t \left[r_x(x_t,a_t)(x_t-\hat{x}_t) - r_a(x_t,a_t)\left(\frac{F_x(x_t,a_t)(x_t-\hat{x}_t)-(x_{t+1}-\hat{x}_{t+1})}{F_a(x_t,a_t)}\right)\right] \\ &\geq \textstyle\sum_{t=1}^{T-1} \alpha^{t-1}(x_t-\hat{x}_t)\left[\alpha\left(r_x(x_t,a_t) - \frac{r_a(x_t,a_t)F_x(x_t,a_t)}{F_a(x_t,a_t)}\right) - \frac{r_a(x_{t-1},a_{t-1})}{F_a(x_{t-1},a_{t-1})}\right] \\ &-\alpha^{T-1}\tfrac{r_a(x_{T-1},a_{T-1})}{F_a(x_{T-1},a_{T-1})}x_T.\end{aligned} \tag{3.11}$$

Ahora, como $a_t = f(x_t)$ y f satisfacen (3.5), se sigue que

$$r_a(x_{T-1},a_{T-1}) + \alpha\Delta(x_T,a_T)F_a(x_{T-1},a_{T-1}) = 0,$$

y por (3.11), es posible concluir que

$$\alpha^{T-1}\frac{r_a(x_{T-1},a_{T-1})}{F_a(x_{T-1},a_{T-1})}x_T = -\alpha^T x_T\Delta(x_T,a_T). \tag{3.12}$$

Similarmente,

$$\begin{aligned}&(x_t-\hat{x}_t)\left[\alpha\left(r_x(x_t,a_t) - \tfrac{r_a(x_t,a_t)F_x(x_t,a_t)}{F_a(x_t,a_t)}\right) - \tfrac{r_a(x_{t-1},a_{t-1})}{F_a(x_{t-1},a_{t-1})}\right] \\ &= (x_t-\hat{x}_t)[\alpha\Delta(x_t,a_t) - \alpha\Delta(x_t,a_t)]. \\ &= 0.\end{aligned}$$

Así,

$$\sum_{t=0}^{T-1} \alpha^t[r(x_t,a_t) - r(\hat{x}_t,\hat{a}_t)] \geq \alpha^T x_T\Delta(x_T,a_T).$$

Entonces, tomando $T \to \infty$, por (3.6), se sigue que

$$v(f,x)) \geq v(\hat{f},x). \tag{3.13}$$

Por lo tanto, f es una política óptima. □

Teorema 3.4.2. *Supongamos que r y F son funciones estrictamente cóncavas y las Condiciones 1.2.6 y 3.4.1 se cumplen. Entonces, existe un único punto de equilibrio $\bar{x} \in X$ de g.*

Demostración. En primer lugar, observe que el punto de equilibrio $\bar{x}$ satisface el siguiente sistema de ecuaciones:

$$\bar{x} = F(\bar{x}, f(\bar{x}))$$

$$r_a(\bar{x}, f(\bar{x})) + \alpha\Delta(F(\bar{x}, f(\bar{x})), f(F(\bar{x}, f(\bar{x}))))F_a(\bar{x}, f(\bar{x})) = 0.$$

Equivalentemente

$$r_a(\bar{x}, f(\bar{x})) + \alpha\Delta(\bar{x}, f(\bar{x}))F_a(\bar{x}, f(\bar{x})) = 0. \tag{3.14}$$

Se probará ahora que (3.14) tiene una solución única. Para ello, considere la siguiente función

$$W(x) := r_a(x, f(x)) + \alpha\Delta(x, f(x))F_a(x, f(x)),\ x \in X, \tag{3.15}$$

en particular, observe que $W(\bar{x}) = 0$.
Ahora, como r y F son estrictamente cóncavas, se deduce que F_a y r_a son no negativas y estrictamente decrecientes, así también por la Condición 3.4.1 e), f es creciente y debido a la Condición 3.4.1 d), V es estrictamente cóncava, por lo tanto, la función W definida en (3.15) es estrictamente decreciente. Sustituyendo Δ y reescribiendo W, se obtiene que:

$$W(x) = r_a(x, f(x))(1 - \alpha F_x(x, f(x))) + r_x F_a(x, f(x)).$$

Entonces considere los siguientes hechos:

1. Por la Condición 3.4.1 b), podemos garantizar que existe $z \in X$ tal que $F_x(x, f(x)) \to \infty$ cuando $x \to z$.

2. Por la Condición 3.4.1 c) y el punto anterior, podemos garantizar, que la ecuación $F_x(x, f(x)) - 1/\alpha = 0$, tiene una única solución, dicha solución se denota por y.

Por un lado, dado que f es una función creciente (Condición 3.4.1 e)) y F es una función estrictamente cóncava, la función $1-\alpha F_x(x, f(x))$ es estrictamente creciente en X. Además, se cumple que $1 - \alpha F_x(y, f(y)) = 0$ por el punto 2 y $r_x(y, f(y))F_a(y, f(y)) \geq 0$, por lo tanto,

$$W(y) = r_a(y, f(y))(1 - \alpha F_x(y, f(y))) + r_x F_a(y, f(y)) \geq 0.$$

Ahora por el punto 1, existe $z \in X$ tal que $F_x(x, f(x)) \to \infty$, cuando $x \to z$ entonces $1 - \alpha F_x(x, f(x)) \to -\infty$ cuando $x \to z$, así,

$$W(z) = r_a(z, f(z))(1 - \alpha F_x(z, f(z))) + r_x F_a(z, f(z)) \leq 0.$$

En conclusión, encontramos dos elementos y y z tales que $W(y) \geq 0$ y $W(z) \leq 0$. Dado que W es una función estrictamente decreciente y continua, W tiene un único punto $\bar{x} \in X$ tal que $W(\bar{x}) = 0$, como consecuencia del teorema del valor intermedio [32]. En conclusión, encontramos un punto de equilibrio único de g, que es $\bar{x}$. □

Observación 3.4.3. *Note que la Condición 3.4.1 e), contiene las siguientes afirmaciones:*

a. *Si F es una función creciente en $\mathbb{K}$, entonces la función g definida en (3.3) es una función creciente.*

b. *Si F es una función continua en $\mathbb{K}$ entonces g es una función continua en X.*

Proposición 3.4.1. *La trayectoria óptima $\{x_t^*\}$ definida recursivamente por $x_{t+1}^* = g(x_t^*)$ con $x_0 \in X$ dado, es una sucesión monótona.*

Demostración. Como V es estrictamente cóncava (Condición 3.4.1 d)), entonces V' es estrictamente decreciente, es decir, que para y, $z \in X$:

$$\text{Si } y > z \text{ entonces } V'(y) < V'(z).$$

Consideremos dos estados sucesivos de la trayectoria óptima x_t^* y x_{t+1}^*, donde $x_{t+1}^* = g(x_t^*)$. Entonces, si $x_t^* < x_{t+1}^*$ entonces $V'(x_t^*) > V'(x_{t+1}^*)$ (o si $x_t^* > x_{t+1}^*$ entonces $V'(x_t^*) < V'(x_{t+1}^*)$) esto implica que $x_t^* - x_{t+1}^*$ y $V'(x_t^*) - V'(x_{t+1}^*)$ tendrán signos opuestos, es decir

$$(x_t^* - x_{t+1}^*)(V'(x_t^*) - V'(x_{t+1}^*)) \leq 0,$$

además, por las ecuaciones (3.7) y (3.10), se obtiene que

$$V'(x_t^*) = \Delta(x_t^*, f(x_t^*)),$$
$$V'(x_{t+1}^*) = -(r_a/\alpha F_a)(x_t^*, f(x_t^*)),$$

entonces

$$(x_t^* - x_{t+1}^*)[\Delta(x_t^*, f(x_t^*)) + (r_a/\alpha F_a)(x_t^*, f(x_t^*))] \leq 0,$$

multiplicando por $\alpha F_a(x_t^*, f(x_t^*)) > 0$ se tiene

$$(x_t^* - x_{t+1}^*)[\alpha r_x - \alpha r_a F_x + r_a(x_t^*, f(x_t^*))] \leq 0,$$

reescribiendo se obtiene que

$$(x_t^* - x_{t+1}^*)[r_a(x_t^*, f(x_t^*)) + \alpha\Delta(x_t^*, f(x_t^*))F_a(x_t^*, f(x_t^*))] \leq 0. \tag{3.16}$$

Note que, bajo (3.14)

$$W(\bar{x}) = r_a(\bar{x}, f(\bar{x})) + \alpha\Delta(\bar{x}, f(\bar{x}))F_a(\bar{x}, f(\bar{x})) = 0.$$

Entonces, como W es decreciente y continuo las siguientes afirmaciones son válidas:

- Si $x_0 > \bar{x}$ entonces como g es creciente (Observación 3.4.3 a)), $x_t^* > \bar{x}$. En consecuencia, $W(x_t^*) < W(\bar{x}) = 0$ y (3.16) implica que $x_t^* > x_{t+1}^*$, esto es $\{x_t^*\}$ es una sucesión monótona decreciente.
- Análogamente, si $x_0 < \bar{x}$ entonces $\{x_t^*\}$ es una sucesión monótona creciente.
- Si $x_0 = \bar{x}$, entonces $x_t^* = x_{t+1}^* = \bar{x}$.

□

Observación 3.4.4. *Observe que:*

1. *Si $x_0 < \bar{x}$, se tiene que:*
   a) *Según la prueba de la Proposición 3.4.1, $\{x_t^*\}$ es creciente, por lo tanto $x_0 < x_t^*$.*
   b) *Como g es creciente, $x_{t+1}^* = g(x_t^*)$ y $g(\bar{x}) = \bar{x}$, aplicando g, t veces a $x_0 < \bar{x}$, entonces $x_t^* < \bar{x}$.*
   c) *Si la Condición 3.4.1 a) se cumple, entonces $\bar{x} < x_M$.*

   En resumen, $x_0 \leq x_t^* \leq x_M$.
2. *De otra manera, si $x_0 > \bar{x}$, $\{x_t^*\}$ es decreciente y $x_t^* \in X = [0, \infty)$ entonces $0 \leq x_t^* \leq x_0$.*

De esta manera, el compacto en el cual f_n converge a f se puede definir como $S := [0, \text{máx}\{x_M, x_0\}]$.

Teorema 3.4.3. *La trayectoria óptima $\{x_t^*\}$ converge monótonamente en $S \subseteq X$ al punto de equilibrio $\bar{x}$ para cualquier valor inicial $x_0 \in X$.*

Demostración. Dado que $\{x_t^*\}$ es una sucesión monótona y acotada (ver Proposición 3.4.1 y Observación 3.4.4, respectivamente), entonces es convergente. Sea x^* el límite de la sucesión $\{x_t^*\}$. Entonces, por la continuidad de la función g (Observación 3.4.3 b)), x^* debe ser un punto fijo de g, debido a las siguientes identidades:

$$x^* = \lim_{t\to\infty} x_{t+1}^* = \lim_{t\to\infty} g(x_t^*) = g(\lim_{t\to\infty} x_t^*) = g(x^*).$$

Entonces x^* es un punto de equilibrio, pero como $\bar{x}$ es el único punto de equilibrio, $x_t^* \to \bar{x}$. □

3.4.1. Ejemplo: Función Utilidad Logarítmica

Este ejemplo proviene de los modelos de crecimiento económico en los que el agente debe decidir en cada periodo t qué parte de la producción generada por un capital debe ser consumida y qué parte debe ser invertida en el siguiente periodo. El objetivo es encontrar una estrategia que optimice el criterio de rendimiento el cual se define a través de una función de utilidad. Diversos trabajos han estudiado este problema en busca de su solución óptima [22], [25], tanto en su planteamiento determinista [29], como en sus versiones estocásticas donde la dinámica o la inversión se ve afectada por una perturbación aleatoria [6].

Este problema puede ser modelado a través de un PDM y su solución óptima puede ser caracterizada a través de la Ecuación de Euler [11].

Considere una economía en la que en cada tiempo discreto t, $t = 0, 1, ...$, hay L_t consumidores (población o mano de obra), con consumo a_t por persona, cuyo crecimiento se rige por la siguiente ecuación en diferencias:

$$L_{t+1} = L_t\eta. \tag{3.17}$$

Se supone que inicialmente se conoce el número de consumidores, L_0. En este caso, $\eta > 0$ fijo representa una variable exógena que afecta a la población de consumidores. En este contexto η representa el valor esperado [43]. La función de producción para la economía está dada por

$$Y_t = G(K_t, L_t),$$

con K_0 conocida, es decir, la producción Y_t es una función del capital, K_t, y mano de obra, L_t, donde la función de producción, G, es una función homogénea de grado uno, es decir, $G(\lambda x, \lambda y) = \lambda G(x, y)$. La salida debe dividirse entre los consumos $C_t = a_t L_t$ y la inversión bruta I_t, es decir,

$$C_t + I_t = Y_t. \tag{3.18}$$

Sea $\delta \in (0, 1)$ la tasa de depreciación del capital. Entonces la ecuación de evolución para el capital viene dada por:

$$K_{t+1} = (1 - \delta)K_t + I_t. \tag{3.19}$$

Sustituyendo (3.19) en (3.18), se obtiene que

$$C_t - (1 - \delta)K_t + K_{t+1} = Y_t. \tag{3.20}$$

De forma habitual, todas las variables se pueden normalizar en términos del capital, es decir, $y_t := Y_t/L_t$ y $x_t := K_t/L_t$. Entonces (3.20) se puede expresar de la siguiente manera:

$$a_t - (1 - \delta)x_t + K_{t+1}/L_t = G(x_t, 1).$$

Ahora, usando (3.17) en la relación anterior, se obtiene que

$$x_{t+1} = \xi(G(x_t, 1) + (1 - \delta)x_t - a_t),$$

$t = 0, 1, 2, ...$, donde $\xi := \eta^{-1}$. Se define $h(x) := G(x, 1) + (1 - \delta)x$, $x \in X := [0, \infty)$, h de ahora en adelante se identificará como la función de producción.

En particular, tenga en cuenta que h es una función potencia, es decir, $h(x) = x^\gamma$, con $\gamma \in (0, 1)$. Entonces, la ley de transición del sistema está dada por

$$x_{t+1} = \xi(x_t^\gamma - a_t),$$

$t = 0, 1, 2, ...$, y $x_0 = x \in X := [0, \infty)$ es el capital inicial. Observe que aquí, x_t^γ indica la producción que genera x_t, a_t la cantidad que se consume y el resto $x_t^\gamma - a_t$ la parte que se invierte en el siguiente periodo.

Las acciones admisibles son dadas por $A(x) = [0, x^\gamma]$ que suponen que no se permite el endeudamiento. La función de recompensa viene dada por una utilidad de consumo, $r(x, a) = U(a) = \ln a$, si $x \in (0, \infty)$, $a \in (0, x^\gamma]$, y $r(0, 0) = U(0) = \infty$. Además, supongamos que $0 < \alpha\gamma < 1$.

Lema 3.4.1. *El ejemplo de utilidad logarítmica es un modelo semicontinuo, es decir, cumple las Condiciones 1.2.4.*

Demostración. a) $A(x) = [0, x^\gamma]$, por lo tanto, $A(x)$ es compacto para cada $x \in X$.

b) Observe que $r(x, a) = \ln(a)$ si $a \in (0, x^\gamma]$ para cada $x \in X$ y como la función logaritmo natural es continua en $(0, x^\gamma]$, en particular es u.s.c. Para $a = 0$, cualquier sucesión $\{a_n\}$ que converja a 0, se tiene que $\lim\sup_{n\to\infty} r(x, a_n) \leq \infty = r(x, 0)$.

c) Este inciso se verifica directamente porque $F(x, a) = \frac{1}{\eta}(x^\gamma - a)$, $(x, a) \in \mathbb{K}$ es una función continua.

□

Lema 3.4.2. *El ejemplo de utilidad logarítmica cumple el Teorema de programación dinámica 1.2.2.*

Demostración. La prueba de este lema puede consultarse en [11]. □

Lema 3.4.3. *El ejemplo de utilidad logarítmica satisface la Condición 3.4.1.*

Demostración. a) Sea $H(x) = x^\gamma/\eta$, una función continua definida en el conjunto compacto y convexo $X = [0, \infty)$. Luego por el Teorema 3.2 en [12] p. 221, H tiene un punto fijo $x_M = \eta^{\frac{1}{\gamma-1}}$ en X, en consecuencia $H(\eta^{\frac{1}{\gamma-1}}) = \eta^{\frac{1}{\gamma-1}}$. De esta manera, $F(x, a) = \frac{1}{\eta}(x^\gamma - a) \leq x^\gamma = H(x)$ para cada $a \in A(x)$, $x \in [0, \infty)$.

b) $F_x(x,a) = \frac{1}{\eta}\gamma x^{\gamma-1}$ y como $(x,a) \to (0,0)$, $F_x(x,a) \to \infty$ porque $\gamma \in (0,1)$.

c) Cuando $(x,a) \to \infty$, $\frac{1}{\eta}\gamma x^{\gamma-1} \to 0$.

d) En primer lugar, observe que $V \in C^2$, debido a r y $F \in C^2$ (consulte la Observación 3.4.1 y [8]). Además, la segunda derivada de la función de valor es estrictamente positiva, ya que r y F son funciones estrictamente cóncavas y crecientes y la función multifunción, $x \mapsto A(x)$, es creciente y convexa.

e) Como r y $F \in C^2$, por Observación 3.4.1, $f \in C^1$ y debido a que la multifunción $x \to A(x)$ es creciente y convexa, por Observación 3.4.4, f es creciente.

□

Lema 3.4.4. *El punto de equilibrio es* $\bar{x} = \left(\frac{\eta}{\alpha\gamma}\right)^{\frac{1}{\gamma-1}}$.

Demostración. El punto de equilibrio se caracteriza por las ecuaciones (3.4) y (3.14). Luego sustituyendo los valores respectivos en ellas, resulta que

$$\begin{cases} \frac{1}{f(\bar{x})}(1 - \frac{\alpha\gamma\bar{x}^{\gamma-1}}{\eta}) = 0, \\ \frac{1}{\eta}(\bar{x}^{\gamma} - f(\bar{x})) = \bar{x}. \end{cases}$$

Resolviendo el sistema anterior para $\bar{x}$ y $f(\bar{x})$, se obtiene que:

$$\bar{x} = \left(\frac{\eta}{\alpha\gamma}\right)^{\frac{1}{\gamma-1}}$$

$$f(\bar{x}) = \left(\frac{\eta}{\alpha\gamma}\right)^{\frac{\gamma}{\gamma-1}}(1 - \frac{\alpha\gamma}{\eta}).$$

En conclusión, para este ejemplo, por el Teorema 3.4.3, la trayectoria óptima converge a $\bar{x} = \left(\frac{\eta}{\alpha\gamma}\right)^{\frac{1}{\gamma-1}}$. □

La Figura 3.2 ilustra las 15 primeras realizaciones de la trayectoria óptima para distintos valores de x_0, α, γ y η, así como el punto de equilibrio en la posición 16.

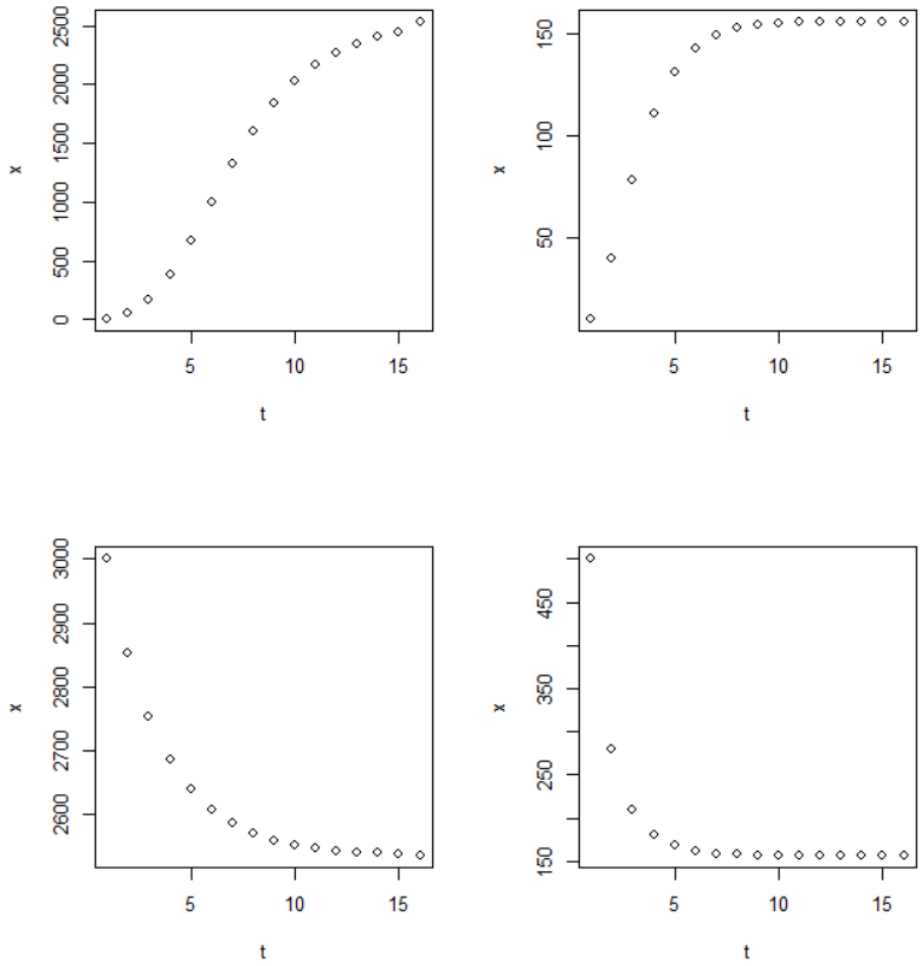

Figura 3.2: Trayectoria Óptima. (a la izquierda $\alpha = 0.3$, $\gamma = 0.7$, $\eta = 0.02$ con $x_0 = 10$ arriba y 3000 abajo; a la derecha $\alpha = 0.5 = \gamma$, $\eta = 0.02$ con $x_0 = 10$ arriba y 500 abajo).

Capítulo 4

Resumen, Conclusiones y Trabajo Futuro

El tema principal de este trabajo de tesis es Procesos de Decisión de Markov con horizonte infinito, el cual se trabajó con los criterios de rendimiento: costo total esperado, recompensa total esperada y recompensa descontada total esperada, se incluyó además, en uno de los problemas, el criterio de entropía relativa, el cual mide los costos que se generan en el envío de la información y restringe el funcionamiento del sistema. Así también, fue necesario estudiar la teoría de Programación Dinámica y los conceptos elementales de Teoría de la Información. En el Capítulo 1 se presentaron los elementos esenciales de dichos temas.

Del área de PDMs se expusieron los Modelos de Control de Markov y los tipos de políticas, los cuales generan el espacio de probabilidad que da lugar al Proceso estocástico de interés (el Proceso de Decisión de Markov). Se expusieron también los criterios que se trabajaron en esta tesis y que dieron lugar a los PDMs estudiados.

En el tema de Programación Dinámica se realizó una recopilación de las condiciones necesarias para resolver los PDMs con ésta técnica, que se basan en garantizar que se cumpla la Condición de Selección Medible y el método de aproximaciones sucesivas si se trabaja en horizonte infinito, tanto para modelos semicontinuos-semicompactos (Condiciones 1.2.1 y 1.2.5) como para modelos semicontinuos (Condiciones 1.2.2, 1.2.3, 1.2.6 y 1.2.7).

De manera adicional, se incluye una introducción a la Teoría de la Información, exponiendo los conceptos principales que se usan durante el desarrollo de la tesis, como son: Entropía y entropía relativa.

Se desarrollaron dos problemas relacionados con PDMs: El problema de

la Ruta más Corta en el Capítulo 2 y el análisis asintótico de un sistema de control determinista en el Capítulo 3.

El problema de la Ruta más Corta se abordó a través de un PDM con el objetivo de optimizar simultáneamente la recompensa total esperada y el costo de información, en espacios de estados infinito numerables, lo cual aporta una extensión a la teoría ya existente. La estrategia fue transformar el problema en uno sin restricciones y utilizar políticas propias, las cuales garantizan que el problema tiene una solución y que existe convergencia hacia la solución óptima. Se demostró que resultados ya probados para el caso finito, los cuales proporcionan características de las políticas propias y las soluciones óptimas (Proposición 2.4.1 y Teoremas 2.4.1 y 2.4.2) se pueden extender al caso numerable. Se presentaron y desarrollaron ejemplos numéricos del problema de la Ruta más Corta en el caso determinista y estocástico, proponiendo una solución basada en la teoría desarrollada y se generó un algoritmo particular que resuelve estos problemas debido a que para espacios numerables el proceso computacional es restringido.

Por otro lado, en el análisis asintótico de problemas de control determinista, se presentó inicialmente la Ecuación de Euler para este modelo y a través de ella se caracterizó la política óptima. Posteriormente se probó la existencia de un único punto de equilibrio del sistema, estableciendo para ello las Condiciones 3.4.1, las cuales permitieron también, garantizar la convergencia de la trayectoria óptima del SCD probando que es monótona (Proposición 3.4.1) y acotada, en concreto, se caracterizó un subconjunto compacto de los números reales donde la trayectoria óptima se concentra (Observación 3.4.4), para finalmente demostrar que converge al punto de equilibrio del sistema (Teorema 3.4.3).

Enlistamos a continuación posibles trabajos futuros en esta dirección.

- Abordar el problema de restricción de la información para espacios de estados que sean subconjuntos de los números reales como se ve en el Ejemplo 2.5.4.
- Considerar los criterios de optimalidad dados en [16], [17], [28], [33], [40] y resolver el problema de restricciones para espacios de estados infinito numerables como se resolvió en esta tesis para el criterio de recompensa total esperada.
- Estudiar problemas de control estocástico inducido por el SDC [9], y garantizar la convergencia de la trayectoria óptima estocástica para modelos generales.
- Aplicar los resultados obtenidos, por ejemplo en el área de Finanzas.

Las siguientes, son publicaciones en las que fueron plasmados los resultados de esta tesis.

- *Asymptotic Analysis of a Deterministic Control System via Euler's Equation Approach.* Journal of Mathematics Research. 2018 (Artículo publicado) [34].
- *Stochastic Shortest Path Problems with Free Energy Criterion.* 2019 (Artículo sometido) [35].
- *Introducción a los Procesos de Decisión de Markov bajo el Criterio de Entropía Relativa.* Compendio de Investigaciones Científicas en México. 2016 (Capítulo de libro) [37].
- *Un Modelo de Inventarios con Demanda Estocástica.* Compendio de Investigaciones Científicas en México. 2016 (Capítulo de libro) [38].
- *Simulación de un Sistema de Inventarios Dinámico bajo la Presencia de Incertidumbre.* 10 Semana Internacional de la Estadística y la Probabilidad. 2017 (Memoria en extenso) [36].

El proyecto de investigación fue premiado por la Sociedad Matemática Mexicana y la Fundación Sofía Kovalevskaia, otorgándole el apoyo que lleva por nombre "Sofía Kovalevskaia" en octubre de 2018. Además, los resultados de la investigación se presentaron en los siguientes eventos:

- Optimización Estocástica de un Proceso de Decisión de Markov con restricción en la información y su aplicación en Finazas. Encuentro Anual de la Sociedad de Matemática de Chile. Universidad de O'Higgins, Rancagua, Chile. 2018.
- El Sistema de Inventarios. Una Aplicación de los Procesos de Decisión de Markov. IV Encuentro sobre didáctica de la estadística, probabilidad y el análisis de datos. Tecnológico de Costa Rica, Cartago, Costa Rica. 2018.
- Convergencia del Modelo Determinista de un Proceso de Decisión de Markov. Congreso Mesoamericano de Investigación UNACH, Tuxtla Gutiérrez, Chiapas, México. 2017.
- Convergencia de un Sistema de Control Determinístico a su Punto de Equilibrio. XX Evento Internacional "La Matemática, la Estadística y la Computación: enseñanza y aplicaciones" MATECOMPU 2018, en la Facultad de Educación de la Universidad de Matanzas, Varadero, Cuba. 2017.

- Simulación de un Sistema de Inventarios Dinámico bajo la Presencia de Incertidumbre. XII encuentro Participación de la Mujer en la Ciencia. Guanajuato. México. 2016.
- Introducción a los Procesos de Decisión de Markov bajo el Criterio de Entropía Relativa. XII encuentro Participación de la Mujer en la Ciencia. Guanajuato. México. 2016.

Bibliografía

[1] Altman, E., "Applications of Markov decision processes in communication networks". In Handbook of Markov decision processes. Springer, Boston, MA. pp. 489-536. 2002.

[2] Bertsekas, D. P., "Proper Policies in Infinite-State Stochastic Shortest Path Problems". IEEE Transactions on Automatic Control, 63(11), pp. 3787-3792. 2018.

[3] Bertsekas, D. P., and Yu, H., "Stochastic shortest path problems under weak conditions."Lab. for Information and Decision Systems Report LIDS-P-2909, MIT. 2013.

[4] Bertsekas, D.P. and Tsitsiklis, J.N., "An analysis of stochastic shortest path problems". Mathematics of Operations Research, 16(3), pp. 580-595. 1991.

[5] Bertsekas, D.P. and Shreve, S.E., "Stochastic Optimal Control: The Discrete Time Case". Academic Press, New York. 1978.

[6] Brock W. and Mirman L., "Optimal economic growth and uncertainty: the discounted case". J. Econ. Th. 4. pp. 479-513. 1972.

[7] Cover, T. M., and Joy A. T., "Elements of information theory". John Wiley and Sons, 2nd ed. ISBN 13 978-0-471-24195-9. 2012.

[8] Cruz-Suárez, H., and Montes-de-Oca, R., "An envelope theorem and some applications to discounted Markov decision processes". Mathematical Methods of Operations Research. 67(2). pp.299-321. 2008.

[9] Cruz-Suárez, H. and Montes-de-Oca, R., "Discounted Markov control processes induced by deterministic systems". Kybernetika. 42(6). pp. 647-664. 2006.

[10] Cruz-Suárez, D., Montes-de-Oca, R., and Salem, F., "Conditions for the uniqueness of optimal policies of discounted Markov decision processes". Mathematical Methods of Operations Research. 60(3). pp. 415-436. 2004.

[11] Cruz-Suárez H., Montes-de-Oca R. and Zacarías G., "A Consumption-Investment problem modelled as a discounted Markov decision process". Kybernetika. 47. pp. 740-760. 2011.

[12] De La Fuente, A., "Mathematical Methods and Models for Economists". Cambridge University Press. 1st ed. ISBN 0-521-58512-0. 2000.

[13] Feinberg, E. A. and Shwartz, A., eds., "Handbook of Markov decision processes: methods and applications". Springer Science and Business Media. Vol. 40. ISBN 978-1-4613-5248-8. 2002.

[14] Flemming, W. H., "Stochastic control for small noise intensities". SIAM J. Control Optim. 9(3). 1971.

[15] Flores-Hernández, R. M., "Monotone optimal policies in discounted Markov decision processes with transition probabilities independent of the current state: existence and approximation". Kybernetika. 49(5). pp. 705-719. 2013.

[16] Fox, R., Moshkovitz, M. and Tishby, N., "Principled option learning in Markov decision processes". arXiv:1609.05524. 2016.

[17] Grau-Moya, J., Leibfried, F., Genewein, T., and Braun, D. A., "Planning with information-processing constraints and model uncertainty in Markov decision processes". Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham. pp. 475-491. 2016.

[18] Hernández-Lerma, O, and Lasserre, J. B., "Discrete-time Markov control processes: basic optimality criteria". Springer Science and Business Media. Vol. 30. 1st ed. ISBN 978-1-4612-6884-0. 1996.

[19] Hernández-Lerma, O, and Jean B. L., "Further Topics on Discrete-time Markov control processes". Springer Science and Business Media. ISBN 978-1-4612-6818-5. 1999.

[20] Hinderer, K., Rieder, U., and Stieglitz, M., "Dynamic Optimization". Springer International Publishing AG. ISBN 978-3-319-48813-4. 2017.

[21] Hu, Q. and Yue, W., "Optimal control for resource allocation in discrete event systems". J. Industr. Manage. Optim. 2(1). pp. 63-80. 2006.

[22] Jaskiewics A. and Nowak A.S., "Discounted dynamic programming with unbounded returns: application to economic models". J. Math. Anal. Appl. 378. pp. 450-462. 2011.

[23] Judd, K. L., "Numerical methods in economics". MIT press. ISBN 0-262-10071-1. 1998.

[24] Judd, K. L. and Guu, S. M., "Perturbation solution methods for economic growth models". In Economic and Financial Modeling with Mathematica, Springer New York. pp. 80-103. 1993.

[25] Kamihigashi T., "Stochastic optimal growth with bounded or unbounded utility and bounded or unbounded shocks". J. Math. Econom. 43. pp. 477-500. 2007.

[26] Krishnamurthy, D. and Todorov, E., "Inverse optimal control with linearly-solvable MDPs". Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 335-342. 2010.

[27] Lamond, B. F., and Boukhtouta, A., "Water reservoir applications of markov decision processes". Handbook of Markov decision processes. Springer, Boston, MA. pp. 537-558. 2002.

[28] Larsson, D.T., Braun, D. and Tsiotras, P., "Hierarchical state abstractions for decision-making problems with computational constraints". Decision and Control (CDC). IEEE 56th Annual Conference. pp. 1138-1143. 2017.

[29] Levhari D. and Srinivasan T. N., "Optimal savings under uncertainty". Rev. Econ. Stud. 36. pp. 153-163. 1969.

[30] Puterman, M. L., "Markov Decision Processes: Discrete Stochastic Dynamic Programming". Wiley, New York. ISBN 0-471-72782-2. 1994.

[31] Rodríguez, M. T., "Premio Nobel de Economía 2001: El libre mercado no funciona". Revista Momento Económico". (118). pp. 47-96. 2001.

[32] Royden, H. L. and Fitzpatrick, P., "Real analysis". Macmillan New York, (32). ISBN 978-0131437470. 1988.

[33] Rubin, J., Shamir, O. and Tishby, N., "Trading value and information in MDPs". Decision Making with Imperfect Decision Makers. Springer Berlin Heidelberg. pp. 57-74. 2012.

[34] Salgado-Suárez G. D., Cruz-Suárez, H. and Zacarías-Flores, J.D., "Asymptotic Analysis of a Deterministic Control System via Euler's Equation Approach". Journal of Mathematics Research, 10(1), pp. 115-123. 2018.

[35] Salgado-Suárez G. D., Cruz-Suárez, H. and Zacarías-Flores, J.D., "Stochastic shortest path problems with free energy criterion". (Sometido). 2019.

[36] Salgado-Suárez G. D., Cruz-Suárez, H., Zacarías-Flores, J.D. and Velasco-Luna, F., "Un Modelo de Inventarios con Demanda Estocástica.". Memorias de la 10ª. Semana Internacional de la Estadística y la Probabilidad. FCFM-BUAP, Puebla, México. 2017.

[37] Salgado-Suárez G. D., Cruz-Suárez, H., Zacarías-Flores, J.D. and Velasco-Luna, F., "Introducción a los Procesos de Decisión de Markov bajo el Criterio de Entropía Relativa". Compendio de Investigaciones Científicas en México. ISBN 978-607-95228-7-2. pp. 1982-1988. 2016.

[38] Salgado-Suárez G. D., Cruz-Suárez, H., Zacarías-Flores, J.D. and Velasco-Luna, F., "Simulación de un Sistema de Inventarios Dinámico bajo la Presencia de Incertidumbre". Compendio de Investigaciones Científicas en México. ISBN 978-607-95228-7-2. pp. 2085-2094. 2016.

[39] Schäl, M., "Markov decision processes in finance and dynamic options". Handbook of Markov decision processes. Springer, Boston, MA. pp. 461-487. 2002.

[40] Tanaka, T., Sandberg, H. and Skoglund, M., "Finite state Markov decision processes with transfer entropy costs". arXiv:1708.09096. pp. 1-12. 2017.

[41] Tishby, N., and Polani, D., "Information theory of decisions and actions". Perception-reason-action cycle: Models, algorithms and systems. Springer. pp. 601-636. 2010.

[42] Van Roy, B., "Neuro-dynamic programming: Overview and recent trends". Handbook of Markov decision processes. Springer, Boston, MA. pp. 431-459. 2002.

[43] Vitoriano, B. De Werra, D. and Parlier, G. H., "Operations Research and Enterprise Systems". 5th International Conference on Operations Research and Enterprise Systems, Rome, Italy. ISBN 978-3-319-53982-9. 2016.

[44] William, N., "Small noise asymptotics for a stochastic growth model". Journal of Economic Theory. 119. pp. 271-298. 2004.