Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas
Posgrado en Ciencias Matemáticas

Sobre el Modelo de Bandidos Armados con Recompensa Difusa

Tesis

para obtener el título de
DOCTORA EN CIENCIAS MATEMÁTICAS

Presenta:
Ciria Ruth Briones García

Director de Tesis:
Dr. Víctor Hugo Vázquez Guevara
Dr. Raúl Montes de Oca Machorro (UAM-I)

Puebla, Puebla. Diciembre 2024

Introducción

Este trabajo se centra en la aplicación de la teoría de Bandidos Armados a problemas de decisión de Markov, con recompensas difusas. En particular, se analizará una transformación difusa trapezoidal de la recompensa en un entorno de Bandidos Armados con recompensas esperadas totales descontadas. Este enfoque permitirá explorar la transición entre la teoría de Bandidos Armados nítidos con su contraparte difusa, y proporcionará una herramienta valiosa para la toma de decisiones en entornos inciertos.

En 1933 [40], Thompson introdujo el problema de los Bandidos Armados, que consiste en asignar pacientes a tratamientos de eficacia desconocida. El objetivo es maximizar el número de pacientes asignados al tratamiento más efectivo. Thompson propuso un enfoque adaptativo para resolver este problema. En 1952 [34], Robbins propuso que el tamaño y la composición de las muestras en experimentos de muestreo no debían estar fijos de antemano. El objetivo de Robbins era minimizar el arrepentimiento por la recompensa total esperada en un horizonte finito.

En 1956 [2], Bellman aplicó la teoría de Programación Dinámica para resolver problemas de control óptimo. Su objetivo era maximizar la recompensa total esperada descontada en un horizonte finito. Basándose en el trabajo de Bellman, Gittins y Jones desarrollaron el índice de Gittins, para más información se puede consultar [23]. Este índice es una medida escalar que se asocia con cada proceso en el escenario de Bandidos Armados y nos guía hacia la estrategia óptima en cada etapa. El controlador selecciona el proceso con el mayor índice de Gittins, en cada etapa, lo que define la estrategia óptima.

Por otro lado, la teoría de números difusos se originó en el trabajo pionero de Zadeh, “Fuzzy sets”, publicado en 1965

[42]. En este trabajo, Zadeh introdujo las definiciones fundamentales de la teoría difusa, incluyendo la función de pertenencia, y estableció las operaciones básicas. Además, el trabajo de Rudin [35] proporciona un compendio detallado sobre la teoría de los espacios métricos, ofreciendo un desarrollo sistemático de esta área matemática.

En "Markov decision processes on finite spaces with fuzzy total rewards" [15] se presenta una introducción a las variables aleatorias difusas, donde se analiza el tratamiento de funciones con valores conjuntos y se define el valor esperado de variables aleatorias difusas a través de la integral en sentido de Auman . En los últimos años, se han desarrollado funciones de recompensa difusa en los procesos de decisión de Markov con funciones objetivo distintas al costo total descontado, puede consultar [15] o [19]. Este enfoque amplía el marco clásico de la teoría de la decisión, que se basa en valores precisos y bien definidos, y lo extiende a situaciones con incertidumbre y valores difusos. Luego, en [14], [20], [28], [27], [29] y [36] se han proporcionado resultados de Procesos de Decisión de Markov con diferentes características difusas, pero según nuestro conocimiento, en ninguno de estos trabajos se ha desarrollado específicamente el tema de los Bandidos Armados.

Esta tesis combina la teoría de Bandidos Armados con la teoría difusa para crear un enfoque más robusto y flexible para la toma de decisiones en entornos inciertos. El objetivo es desarrollar un método que pueda manejar la incertidumbre y la vaguedad en la toma de decisiones. Para alcanzar este objetivo, se desarrollan modelos y algoritmos híbridos que combinan la teoría de Bandidos Armados con la teoría difusa. Estos modelos y algoritmos se evalúan en diferentes escenarios para determinar su eficacia y eficiencia.

Se analiza específicamente la aplicación de la teoría difusa en la resolución de problemas de Bandidos Armados con

recompensas difusas. Los resultados numéricos obtenidos demuestran la eficacia del enfoque propuesto en términos de optimización de la recompesa esperada. Se examina la conexión entre la teoría de Bandidos Armados y la teoría de decisión bajo incertidumbre, y se analiza cómo la teoría difusa puede ser utilizada para mejorar la toma de decisiones en entornos inciertos, mediante la incorporación de la incertidumbre y la vaguedad en el proceso de decisión.

En síntesis, esta tesis ofrece una contribución sobre la teoría de los Bandidos Armados y la teoría difusa, al presentar un enfoque innovador para la toma de decisiones en entornos inciertos, que combina la teoría de Bandidos Armados con la teoría difusa. Esta investigación puede ser de interés para investigadores y profesionales en el campo de la toma de decisiones, la inteligencia artificial y la teoría de la decisión bajo incertidumbre, debido a que nos permite modelar situaciones de toma de decisiones o en las que se deben adaptar cambios en el entorno o las condiciones y puede llevar a mejoras significativas en el rendimiento y eficiencia [37].

Índice general

Introducción **IV**

Índice de figuras **IX**

1. Procesos de Decisión de Markov **1**
1.1. Procesos de Decisión de Markov 1
1.1.1. Políticas de control 4
1.1.2. Criterio de Rendimiento 5
1.1.3. Programación Dinámica 6
1.2. Modelo de Bandidos Armados (BA) 7
1.2.1. Modelo 7
1.2.2. Índice de Gittins 10

2. Teoría Difusa **17**
2.1. Números Difusos 17
2.2. Aritmética Difusa 20
2.3. Probabilidad Difusa 23

3. Modelo de Bandidos Armados con Recompensas Difusas Trapezoidales **29**
3.1. Modelo BA con recompesa difusa 30
3.2. Índice de Gittins difuso 33
3.3. Ejemplo 40
3.3.1. Caso con deterioro 41
3.3.2. Resultados numéricos 42

Bibliografía **51**

Índice de figuras

1.1. Diagrama de un Proceso de Decisión de Markov. 3
1.2. Diagrama de un Modelo de Bandidos Armados. 9

2.1. Gráfico de un número difuso Γ. 19
2.2. α-corte de un número difuso trapezoidal Γ. . . 20
2.3. Gráfica de $(\tilde{Y})_\alpha$. 24

3.1. Descripción de los tiempos de paro. 43
3.2. Descripción de los tiempos de paro para el trabajo 1 (frecuencia absoluta). 44
3.3. Descripción de los tiempos de paro para el trabajo 1 (frecuencia relativa). 45
3.4. Descripción de los tiempos de paro para el trabajo 2 (frecuencia absoluta). 45
3.5. Descripción de los tiempos de paro para el trabajo 2 (frecuencia absoluta). 46
3.6. Evolución vs estados. 46
3.7. Evolución de estados. 47

Capítulo 1

Procesos de Decisión de Markov

Este capítulo ofrece los conceptos fundamentales de los Procesos de Decisión de Markov (PDM) y la Programación Dinámica.

Se define el modelo de PDM y se analiza la estructura de la toma de decisiones en este contexto. Además, se introduce la Programación Dinámica como herramienta para resolver problemas de control óptimo. Finalmente, se discuten las condiciones suficientes para garantizar la convergencia de la función de valor óptimo. Boucherie y Van Dijk (2017) en [10] presentan una colección de aplicaciones prácticas de procesos de decisión de Markov en diferentes campos, tales como logística y cadena de suministros, sistemas de transporte, sistemas de salud, finanzas y economía, energía y medio ambiente, comunicaciones y redes, manufactura y producción.

1.1. Procesos de Decisión de Markov

Los Procesos de Decisión de Markov se basan en la idea de que un sistema puede ser modelado como una sucesión de

estados y acciones. En [25] se definen los procesos de decisión de Markov como un modelo matemático para la toma de decisiones en sistemas dinámicos. Para formalizar esta idea definimos un Modelo de Control de Markov (MCM) de la siguiente manera.

Definición 1.1.1. Un Modelo de Control de Markov (MCM), estacionario, a tiempo discreto, consiste de la quíntupla:

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

en donde X y A son espacios de Borel no vacíos, llamados espacio de estados y de acciones o controles, respectivamente. Para cada $x \in X$, $A(x)$ es un subconjunto medible no vacío de A conocido como un conjunto de acciones admisibles para el estado $x \in X$. El conjunto $\mathbb{K}$ de parejas de estados-acciones admisibles, está definido por:

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

$Q(\cdot|\cdot)$ es un kérnel estocástico definido en X dado $\mathbb{K}$, llamado ley de transición. Por último, $r : \mathbb{K} \to \mathcal{R}$ es una función medible llamada función de recompensa (o costo) en un paso, que asigna a cada pareja $(x, a) \in \mathbb{K}$ una recompensa $r(x, a) \in \mathbb{R}$.

En la Figura 1.1 podemos observar cómo es la transición del estado x_t al estado x_{t+1}, que se elige una acción, transitamos al siguiente estado mediante Q y recibimos una recompensa $r(x_t, a_t)$.

Un MCM estacionario a tiempo discreto se puede pensar como un sistema estocástico controlado que se observa de manera periódica en los instantes $t = 0, 1, 2, \ldots$ La dinámica que describe a este sistema estocástico funciona de la siguiente forma: si el sistema se encuentra en el estado $x_t = x \in X$ en el tiempo t y se aplica la acción $a_t = a \in A(x)$, entonces ocurren dos cosas:

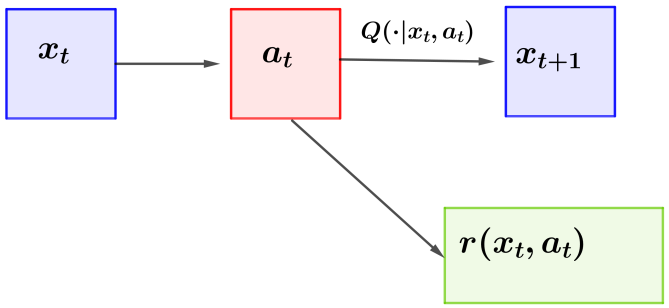

Figura 1.1: Diagrama de un Proceso de Decisión de Markov.

- se recibe una recompensa (o se paga un costo) $r(x, a)$ $(c(x, a))$ y,

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

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

  $B \in \mathcal{B}(X)$, donde $\mathcal{B}(X)$ denota la $\sigma-$álgebra de Borel de X. Es decir, la probabilidad de que el sistema se traslade a un nuevo estado que se encuentre en el conjunto B, dado que se encuentra en el estado $x \in X$ y se aplica la acción $a \in A(x)$, se denota por $Q(B|x, a)$.

Una vez hecha esta transición a un nuevo estado, se elige una nueva acción y el proceso anteriormente descrito se repite.

1.1.1. Políticas de control

Para introducir el concepto de estrategia o política, considerése un MCM y al espacio de las historias observadas del proceso hasta el tiempo t, denotado por $\mathbb{H}_t$, y definido como $\mathbb{H}_0 = X$ y $\mathbb{H}_t = \mathbb{K} \times \mathbb{H}_{t-1}$, para $t = 1, 2, \ldots$ Un elemento de $\mathbb{H}_t$, llamado la historia hasta el tiempo t, es un vector de la forma

$$h_t = (x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$$

donde $(x_i, a_i) \in \mathbb{K}$ para $i = 0, \ldots, t-1$, $x_t \in X$ y $a_i \in A(x_i)$.

Definición 1.1.2. Una política aleatorizada o simplemente política, es una sucesión $\pi = \{\pi_t\}_{t=0}^{\infty}$ de kérneles estocásticos, donde cada π_t está definida sobre A dado $\mathbb{H}_t$ y satisface que: $\pi_t(A(x_t)|h_t) = 1$ para toda $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, \ldots$, el conjunto de todas las políticas se denota por Π.

De acuerdo con esta definición, se puede ver a una política $\pi = \{\pi_t\}$ como una sucesión $\{a_t\}$ de variables aleatorias sobre A, tales que, para cada historia y $t = 0, 1, 2, \ldots$, la distribución de a_t es $\pi_t(\cdot|h_t)$ concentrada en el conjunto de acciones admisibles $A(x_t)$. Es decir, la acción en cualquier tiempo t es una variable aleatoria y depende de todas las historias.

Definición 1.1.3. Una política $\pi \in \Pi$ es **Determinista Markoviana**, si existe una sucesión $\{f_t\} \subset \mathbb{F}$, donde

$$\mathbb{F} = \{f : X \to A | f \text{ es medible y } f(x) \in A(x) \text{ para cada } x \in X\},$$

con la característica de que $\pi_t(\cdot|h_t)$ está concentrada en $f_t(x_t)$ para cada $h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, \ldots$, si existe $f \in \mathbb{F}$ tal que para toda $t = 0, 1, 2, \ldots$, $f_t = f$, entonces a π se le conoce como una política **Determinista Markoviana Estacionaria**.

De la definición, se deduce que $\pi(\cdot|h)$ está concentrada en $f(x)$ si, $\pi_t(C|h) = I_C(f(x))$ para cada $C \in \mathcal{B}(A)$. Donde $\mathcal{B}(A)$ denota la σ-álgebra de Borel de A e I_C es la función indicadora sobre C.

Dados $\pi \in \Pi$ y $x_0 = x \in X$, por el Teorema de Ionescu-Tulcea (puede consultar [32]) existe una única medida de probabilidad P_x^π definida en el espacio canónico $(\Omega, \mathcal{F})$, donde $\Omega := (X \times A)^\infty$ y $\mathcal{F}$ es su correspondiente σ-álgebra producto, tal que para cada $C \in \mathcal{B}(A), B \in \mathcal{B}(X), h_t \in \mathbb{H}_t$ y $t = 0, 1, 2, \ldots$, se tiene que

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

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

El proceso estocástico $((\Omega, \mathcal{F}, P_x^\pi), \{x_t\})$ es llamado un *Proceso de Control de Markov a tiempo discreto o Proceso de Decisión de Markov (PDM).*

De acuerdo a (1.2), la distribución del estado x_{t+1} solo depende de la pareja estado-acción (x_t, a_t), dicha condición es una propiedad tipo Markov. La esperanza con respecto a P_x^π será denotada por E_x^π.

1.1.2. Criterio de Rendimiento

Cada PDM estará dotado de una función real, llamada función objetivo o criterio de rendimiento, que medirá en algún sentido la calidad de cada política, a través de la sucesión de recompensas que genera.

Considere un PDM fijo y el conjunto de políticas Π. Se define el criterio de rendimiento **Recompensa (o costo) total descontada** para $\pi \in \Pi$ y $x \in X$, como sigue

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

con $x_0 = x$ y $\alpha \in (0, 1)$. A α se le conoce como factor de descuento. Cuando $\alpha = 1$, el criterio es conocido como **Recompensa (o costo) Total Acumulada**. Al entero positivo

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

Definición 1.1.4. Una política $\pi^* \in \Pi$, es óptima, si para cada $x \in X$

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

La función V definida para $x \in X$ como

$$V(x) := \sup_{\pi \in \Pi} \upsilon(\pi, x)$$

se le llama función de valor óptimo.

El problema de control óptimo consiste en determinar a la política óptima.

1.1.3. Programación Dinámica

La programación dinámica es una técnica utilizada para resolver problemas de optimización en sistemas dinámicos. En este contexto, se consideran problemas de control óptimo en los que se busca maximizar una función de recompensa total esperada.

En problemas con horizonte finito, se define la función de valor óptimo como la máxima recompensa total esperada. El teorema de programación dinámica establece que esta función puede ser calculada de manera recursiva mediante la resolución de subproblemas más pequeños, [7].

En problemas con horizonte infinito, se define la función de valor óptimo como la máxima recompensa total descontada esperada. En este caso, se introduce el operador de programación dinámica, que es una contracción en el espacio de funciones medibles y acotadas.

Se demuestra que la función de valor óptimo es un punto fijo del operador de programación dinámica [32], lo que permite aproximarse a ella mediante problemas con horizonte finito. Además, se establece que existe una política óptima, puede consultar [38], que satisface la ecuación de programación dinámica. Para ésto se asume que las recompensas de cada bandido son independientes entre sí, las distribuciones de recompensa de cada bandido son estacionarias en el tiempo y además las recompensas están acotadas en un intervalo finito.

1.2. Modelo de Bandidos Armados (BA)

El modelo de Bandidos Armados es un ejemplo clásico en la teoría de decisión, [2], [24] y [34]. Consiste en un conjunto de brazos (acciones), cada uno tiene una distribución de recompensas desconocida y un agente debe elegir un brazo en cada instante de tiempo. El objetivo es maximizar la recompensa total obtenida. Un modelo de Bandidos Armados es un caso particular de un Proceso de Decisión de Markov, pues tiene la misma estructura y la decisión (a_i^t) que se toma en el instante t no depende de toda la historia sino del estado X_t.

Primero se presenta la estructura del modelo de Bandidos Armados, más adelante hablamos del índice de Gittins, que es una estrategia o algoritmo para dar una solición óptima al problema de Bandidos Armados, propuesta en [23].

1.2.1. Modelo

Sea $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_k(t))$ el estado de una familia simple de procesos Bandidos Armados, donde para cada

$t \in \{0, 1, \ldots\}$, $x_i(t)$, es el estado del i-ésimo Bandido Armado en el tiempo t, que es una variable aleatoria discreta definida en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y que toma valores en un conjunto no vacío X, para $i \in \{1, 2, \ldots, k\}$. X es llamado el espacio de estados y se supone que es un conjunto finito.

Una vez que se observa el estado de un proceso de Bandidos Armados se selecciona una acción $\mathbf{a}(t) = (a_1(t), a_2(t), \ldots, a_k(t))$ por el controlador en el tiempo t. Para cada $t \in \{0, 1, \ldots\}$, $a_i(t)$, es la acción seleccionada para el i-ésimo Bandido Armado en el tiempo t, es una función binaria, que toma el valor 1 ($a_i(t) = 1$) si el controlador elige el $i-$ésimo proceso Bandido, el espacio de acciones se puede definir para cada instante t como $A(t) := \{(a_1(t), a_2(t), \ldots, a_k(t)) : \sum_{i=1}^{k} a_i(t) = 1\}$.

Además, el estado del proceso seleccionado transita a un nuevo estado con cierta probabilidad $P_i = [p_i(x, y)] := P(x_i(t+1) = y | x_i(t) = x), x, y \in X$ y $i \in \{1, 2, \ldots, k\}$, mientras que los procesos no seleccionados no cambian de estado. Así, la dinámica del proceso estocástico se puede describir de la siguiente manera: en cada instante el jugador observa el estado del sistema $\mathbf{x}(t) = (x_1(t) = x_1, x_2(t) = x_2, \ldots, x_i(t) = x, \ldots, x_k(t) = x_k) \in X^k$ y elige la acción $a(t) \in A(t)$, con $a_i(t) = 1$, para $i \in \{1, 2, \ldots, k\}$, es decir, si el $i-$ésimo Bandido Armado es activado, entonces se obtiene una recompensa $r_i(a_i(t))$ y dicho Bandido Armado se mueve del estado x al estado y de acuerdo con la ley de transición $p_i(x, y)$. Y, para los bandidos armados restantes no habrá cambio en sus estados, de esta manera $\mathbf{x}(t+1) = (x_1(t) = x_1, x_2(t) = x_2, \ldots, x_i(t) = y, \ldots, x_k(t) = x_k)$. La Figura 1.2.1 ilustra éste proceso.

Defina $\Pi := \{\{\mathbf{a}(t) : t = 0, 1, \ldots\} | \mathbf{a}(t) \in A(t)\}$ el conjunto de todas las políticas admisibles (o estrategias de activación). Entonces, para $\pi \in \Pi$ y $\mathbf{x} \in X^k$ definimos la función de valor

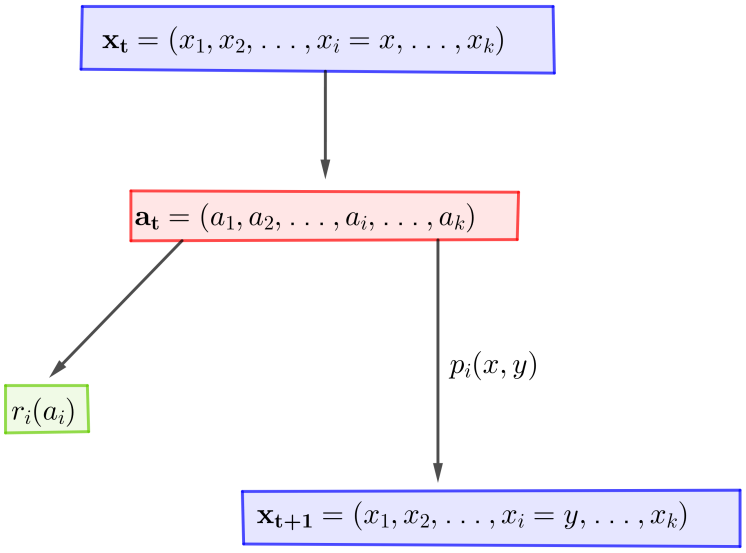

Figura 1.2: Diagrama de un Modelo de Bandidos Armados.

esperado total descontado como sigue:

$$V(\pi, \mathbf{x}) := E_{\mathbf{x}}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t \sum_{i=1}^{k} r_i(x_i(t))a_i(t)\right] \tag{1.3}$$

donde $0 < \beta < 1$ es un factor de descuento. Análogo a la sección anterior, el operador de esperanza $E_{\mathbf{x}}^{\pi}$ está asociado con la medida producto $P_{\mathbf{x}}^{\pi}$ definido en $(\Omega^k, \mathcal{F}^k)$, que se induce mediante el Teorema de Ionescu-Tulcea. Entonces el problema de Bandidos Armados consiste en determinar una estrategia de activación $\pi \in \Pi$ tal que maximice la recompensa total esperada descontada, es decir,

$$V(\pi_0, \mathbf{x}) = \sup_{\pi \in \Pi} V(\pi, \mathbf{x}), \tag{1.4}$$

$\mathbf{x} \in X^k$. La función $V_0(\mathbf{x}) := V(\pi_0, \mathbf{x})$ se llama función de valor óptimo.

Un ejemplo clásico de Bandidos Armados es el de las máquinas tragamonedas, ver [39], tenemos una serie de máquinas tragamonedas cada una con una probabilidad desconocida de pagar una recompensa. Cada vez que jugamos en alguna de las máquinas podemos perder o ganar dinero. El objetivo es maximizar las ganancias totales jugando en las máquinas. En este problema no contamos con las probabilidades de pagar una recompensa para cada máquina y se busca equilibrar la exploración de las máquinas para obtener información sobre sus probabilidades de pagar una recompensa. El modelo de Bandidos Armados proporciona una solución óptima para este problema. La idea de asignar un índice de valor a cada máquina, que refleje el valor esperado de recompensa y su incertidumbre. Luego, se elige la máquina con el índice de valor más alto.

1.2.2. Índice de Gittins

El problema descrito anteriormente puede abordarse mediante la metodología de programación dinámica que puede mostrar la existencia de políticas óptimas estacionarias. Sin embargo, el índice de Gittins resulta una buena opción para resolver este problema, debido a que maximiza la recompensa total obtenida que es el objetivo principal del problema, brinda un equilibrio entre la exploración y explotación de los brazos (o bandidos), se puede aplicar a problemas con horizonte finito o infinito, es simple y flexible.

Gittins y Jones propusieron un enfoque para remediar este problema en su trabajo "A dynamic allocation index for sequential desing of experiments" [23], donde se demuestra que

una política óptima es del tipo índice (o política prioritaria), es decir, para el problema de Bandidos Armados se puede calcular el índice de asignación dinámica (o simplemente índice de Gittins), que depende solo de ese proceso, y luego en cada momento el jugador opera en el proceso Bandido con el índice más alto, para facilitar la escritura se considerará cada proceso Bandido fijo y por lo tanto, se omitirá el subíndice de la etiqueta. Para cada proceso Bandido se define el índice de Gittins de la siguiente manera

$$\mathbb{G}(x) = \sup_{\tau>0} G(x,\tau) = \sup_{\tau>0} \frac{\mathbb{E}_x\left[\sum_{t=0}^{\tau-1} \beta^t r(X(t))\right]}{\mathbb{E}_x\left[\sum_{t=0}^{\tau-1} \beta^t\right]}, \, x \in \mathbb{X} \tag{1.5}$$

donde τ es un tiempo de paro asociado con el proceso estocástico $\{X(t)\}$, es decir, para cada $n \in \mathbb{N}$, $[\tau = n] \in \mathcal{F}_n := \sigma(X(1), X(2), \ldots, X(n))$.

Además, se considera la suposición siguiente sobre la función de recompensa de cada proceso Bandido, para trabajar sobre los estados en los que la recompensa es mayor o igual a cero y tener recompensas comparables lo que nos permitirá poder optimizar.

Suposición 1.2.1. a) $r(\cdot) \geq 0$ y $r(\cdot)$ tiene soporte finito, es decir $F := \{x \in \mathbb{X} : r(x) > 0\}$ es finito .

b) $\mathbb{E}_x\left[\sum_{t=0}^{\infty} r(X(t))\right] < \infty$ para todo $X \in \mathbb{X}$.

Observación 1.2.1. a) Se cumple la Suposición 1.2.1, por ejemplo, bajo una condición transitoria de Markov ([13], [30]) , es decir, existe una partición de $\mathbb{X}$, que consta de los subconjuntos $\hat{X}$ y M, tal que las siguientes condiciones se satisfacen:

i) $\sum_{t=0}^{\infty} \mathbb{P}_x(X(t) = y) < \infty$ para toda $y \in \hat{X}$,

ii) $p(x,y) = 0$ para $x \in M$ y $y \in \hat{X}$, y

iii) $r(x) = 0, x \in M$.

b) Además, observe que la Suposición 1.2.1 b) implica que

$$0 \leq \sum_{t=0}^{\infty} \beta^t r(X(t)) \leq \sum_{t=0}^{\infty} r(X(t)) < \infty, \mathbb{P}_x - \text{c.s.} \quad (1.6)$$

Si el lector desea estudiar más sobre el cálculo del índice de Gittins, Chakraborty y Mahajan nos presentan diferentes algoritmos en su trabajo "Multi-Armed Bandits, Gittins Index, and Its Calculation" [17].

Prueba del Teorema del índice de Gittins

El teorema de Gitttins es la herramienta que nos da una política de decisión óptima comom se enuncia a continuación.

Teorema 1.2.1. En cualquier instante de decisión $t \geq 0$, el Proceso Bandido Armado que debe seleccionarse para ser activado es el que tiene el índice de Gittins más alto.

En [22] se muestran 4 pruebas sobre la regla de prioridad del índice de Gittins que es óptima para procesos de Bandidos Armados Multiarmados, aunque no son las únicas que existen. A continuación damos una breve descripción de 3 y presentamos por completo una de ellas (la tercer prueba), porque es la más directa ya que la finalidad de esta sección es la presentación del resultado y no de sus distintas pruebas.

Primer prueba: Argumentos de intercambio. Podemos describir esta prueba en 4 pasos, en el primero se define la política del índice de Gittins y se establecen las políticas $\pi^{(s)}$ que convergen a π^*. En el segundo paso, se demuestra que la política $\pi^{(s)}$ coincide con π^* durante al menos las primeras s unidades de tiempo. Para el tercer paso, se utiliza un argumento de intercambio por pares para demostrar que $V_{\pi^{(1)}}(i) \leq V_{\pi^{(0)}}(i)$. Por último, se concluye que la política de Gittins es óptima.

Segunda prueba: Entrelazamiento de cargos prevalentes: Se define la política del índice de Gittins y se establece la

igualdad entre la recompensa total esperada descontada y la suma de los cargos justos. Se usa la desigualdad de Hardy Littlewood Polya para demostrar que la recompensa total esperada descontada es mayor o igual a la suma de los cargos prevalentes. Después se muestra que, bajo la política del índice de Gittins, la desigualdad anteior se cumple como igualdad. Y se concluye que la política del índice de Gittins es óptima.

Cuarta prueba: El enfoque de la región alcanzable. Esta prueba consite en formular un programa lineal que describe la región alcanzable, demostrar que cualquier política de prioridad óptima es un punto extremo de la región alcanzable, utilizar el algoritmo de Klimov para encontrar una política de prioridad óptima que corresponde a cada punto extremo de la región alcanzable y concluir que la región alcanzable coincide con la región factible del programa lineal.

Tercera prueba: Opción de retiro. Consideramos el problema de los Bandidos Multiarmados con opción de retiro. Se consideran N brazos en el estado inicial $\mathbf{Z}(0) = \mathbf{i}$ y una recompensa por retiro $\mathbf{M}$. Para el brazo n, definimos:

$$M_n(i_n) = \text{ínf}\,\{M : V_n(i_n, M) = M\} \tag{1.7}$$

.

Teorema 1.2.2 (Wittle)**.** Para el problema multiarmado con opción de retiro, la política óptima es:

(a) Si $M \geq M_n(i_n)$ para toda $n = 1, \ldots, N$ retirarse.

(b) De lo contrario, activa el brazo n^* para el cual M_{n^*} para el cual $M_{n^*}(i_{n^*}) = \text{máx}_{n=1,\ldots,N}\, M_n(i_n)$.

Las ecuaciones de optimalidad para el problema de los Bandidos Multiarmados con opción de retiro son:

$$V(i, M) = \max_{n=1\ldots,N} \left\{ M, R(i_n) + \alpha \sum_{j \in E} p(i_n, j) V(i_1, \ldots, j, \ldots, i_N, M) \right\}. \tag{1.8}$$

Si se siguen (a) y (b), se puede especular sobre la forma de $V(i, M)$. Sea $\tau_n(i_n, M)$ el tiempo de retiro para el bandido individual n con recompensa total M. Denotemos por $T(M)$ el tiempo de retiro para todo el sistema de Bandidos Multiarmados. Entonces, (a) y (b) implican:

$$T(M) = \sum_{n=1}^{N} \tau_n(i_n, M) \tag{1.9}$$

Ahora,

$$\begin{aligned} \frac{\partial}{\partial M} V(i, M) &= \mathbb{E}\left[\alpha^{T(M)}\right] = \mathbb{E}\left[\alpha^{\sum_{n=1}^{N} \tau_n(i_n, M)}\right] \\ &= \Pi_{n=}^{N} \mathbb{E}\left[\alpha^{\tau_n(i_n, M)}\right] = \Pi_{n=1}^{N} \frac{\partial}{\partial M} V_n(i_n, M), \end{aligned} \tag{1.10}$$

donde las igualdades se siguen facilmente por el Lema 4 presentado en el artículo, por (1.9) y por la independencia de las variables aleatorias $\tau_n(i_n, M)$.

También tenemos que $V(\mathbf{i}, M) = M$ para $M \geq \frac{c}{1-\alpha}$. Integrando (1.10), obtenemos la siguiente forma conjeturada para la función de valor óptima:

$$\hat{V}(\mathbf{i}, M) = \frac{c}{1-\alpha} - \int_{M}^{\frac{c}{1-\alpha}} \Pi_{n=1}^{N} \frac{\partial}{\partial M} V_{n'}(i_{n'}, M) dm. \tag{1.11}$$

Para cada n, definimos

$$Q_n(\mathbf{i}, M) = \Pi_{n' \neq n} \frac{\partial}{\partial M} V_{n'}(i_{n'}, M). \tag{1.12}$$

Por el Lema 3, que aparece en el artículo, Q_n es no negativo y no decreciente, variando desde 0 cuando $M \leq -\frac{c}{1-\alpha}$ a 1 cuando $M \geq \frac{c}{1-\alpha}$. Sustituyendo (1.12) en (1.11) e integrando por partes, obtenemos para cada n:

$$\hat{V}(\mathbf{i}, M) = V_n(i_n, M) Q_n(i, M) + \int_{M}^{\frac{c}{1-\alpha}} V_n(i_n, m) dQ_n(\mathbf{i}, m). \tag{1.13}$$

Para completar la prueba necesitamos demostrar que $\hat{V} = V$ mostrando que $\hat{V}$ satisface la ecuación de optimalidad. Primero demostremos que $\hat{V}(\mathbf{i}, M) \leq M$. Usando que $V_n(i_n, m)$ es monótona en m y usando (1.13), tenemos

$$\begin{aligned}\hat{V}(\mathbf{i}, M) &\geq V_n(i_n, M)Q_n(\mathbf{i}, M) + V_n(i_n, M)\int_M^{\frac{c}{1-\alpha}} V_n(i_n, M)dQ_n(\mathbf{i}, m) \\ &= Q_n(\mathbf{i}, \frac{c}{1-\alpha})V_n(i_n, M) = V_n(i_n, M) \geq M.\end{aligned} \tag{1.14}$$

Demostramos ahora que $\Delta_n \geq 0$ para cualquier n, donde:

$$\Delta_n = \hat{V}(\mathbf{i}, M) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j)\hat{V}(i_1, \ldots, j, \ldots, i_N, M). \tag{1.15}$$

Observamos que $Q_n(\mathbf{i}, m)$ no depende del valor de i_n, es decir

$$Q_n(\mathbf{i}, m) = Q_n(i_1, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N, m).$$

Sustituyendo (1.13) en (1.15), Δ_n se escribe como:

$$\begin{aligned}\Delta_n &= Q_n(\mathbf{i}, M)(V_n(i_n, M) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j)V_n(j, M)) \\ &+ \int_M^{\frac{c}{1-\alpha}} (V_n(i_n, m) - R(i_n) - \alpha \sum_{i \in E} p(i_n, j)V_n(j, m)dQ_n(i, m)) \\ &\geq 0\end{aligned} \tag{1.16}$$

donde

$$V_n(i_n, m) \geq R(i_n) + \alpha \sum_{j \in E} p(i_n, j) \tag{1.17}$$

por la ecuación de optimalidad para $V_n(i_n, m)$.

La igualdad se cumple en 1.14 y 1.16, exactamente bajo la política de Whittle. Consideramos $M, m > M_{n^*}(i_{n^*})$, entonces $Q_n(\mathbf{i}, M) = 1$ y $dQ_n(\mathbf{i}, m) = 0$. Observando (1.13), vemos

que (1.14) se cumple como igualdad para $M > M_{n^*}(i_{n^*})$.

Consideramos $M \leq m \leq M_n(i_n)$, para el cual (1.17) se cumple como igualdad. Sustituyendo esto en (1.16), tenemos que para tal M:

$$\begin{aligned}\Delta_n = &\int_{M_n(i_n)}^{\frac{c}{1-\alpha}} (V_n(i_n, m) - R(i_n) \\ &- \alpha \sum_{j \in E} p(i_n, j) V_n(j, m)) dQ_{n^*}(\mathbf{i}, m).\end{aligned} \tag{1.18}$$

En particular, si tomamos n^* tenemos

$$\begin{aligned}\Delta_{n^*} = &\int_{M_{n^*}(i_{n^*})}^{\frac{c}{1-\alpha}} (V_{n^*}(i_{n^*}, m) - R(i_{n^*}) \\ &- \alpha \sum_{j \in E} p(i_{n^*}, y) V_{n^*}(j, m)) dQ_{n^*}(\mathbf{i}, m) \\ =&0\end{aligned} \tag{1.19}$$

desde $dQ_{n^*}(\mathbf{i}, m) = 0$ para $m \geq M_{n^*}(i_{n^*})$. Y con esto se concluye la demostración.

Capítulo 2

Teoría Difusa

La teoría difusa es una herramienta matemática que permite modelar y analizar la incertidumbre y la vaguedad en sistemas complejos. En este capítulo, se exploran los conceptos básicos de la teoría difusa. Para una revisión más detallada, véase [42].

Algunos ejemplos de la teoría difusa pueden ser: la evaluación de la probabilidad de que un sistema de transporte público funcione correctamente, los números difusos pueden representar la incertidumbre en la evaluación de la probabilidad; cuando se considera invertir en una empresa que tiene un futuro incierto, los números difusos pueden representar la incertidumbre en la evaluación de la rentabilidad de la inversión.

La primera sección de este capítulo se enfoca en la introducción de conceptos básicos relacionados con la teoría difusa.

2.1. Números Difusos

En la teoría de conjuntos, cada elemento de un universo puede pertener o no a un conjunto. Sin embargo, en la teoría de conjuntos difusos, se introduce un concepto al que le llamamos "grado de pertenencia" definido a partir de una "función

de pertenencia", que se encarga de asignar un valor entre 0 y 1 indicando el grado en el que el elemento pertenece al conjunto.

Por ejemplo, considere un conjunto de personas que son "jóvenes". En la teoría de conjuntos difusos, se puede asignar un grado de pertenencia y decir que una persona de 20 años podría tener un grado de pertenencia de 0.8, mientras que una persona de 60 años podría tener un grado de pertenencia de 0.2.

Intuitivamente un número difuso es una representación matemática de un valor incierto o vago, que se define mediante una función de pertenencia. Sea Λ un conjunto no vacío. Entonces un conjunto difuso Γ sobre Λ es definido en términos de una función de pertenencia Γ', que asigna a cada elemento de Λ un valor real del intervalo $[0, 1]$. El $\alpha-$corte de Γ, denotado por Γ_α, es definido por el conjunto $\Gamma_\alpha := \{x \in \Lambda | \Gamma'(x) \geq \alpha\}$ ($0 \leq \alpha \leq 1$ y Γ_0 es la cerradura de $\{x \in \Lambda | \Gamma'(x) > 0\}$ denotado por $cl\{x \in \Lambda | \Gamma'(x) > 0\}$.

Definición 2.1.1. Un número difuso Γ es un conjunto difuso definido sobre el conjunto de los números reales $\mathbb{R}$, que satisface:

a) Γ' es normal, es decir, que existe $x_0 \in \mathbb{R}$ tal que $\Gamma'(x_0) = 1$;

b) Γ' es convexa, es decir, Γ_α es convexa para toda $\alpha \in [0, 1]$;

c) Γ' es semi-continua superiormente;

d) Γ_0 es compacto.

En la Figura 2.1 se ilustra un número trapezoidal Γ donde podemos apreciar el α-corte para el α dado (Γ_α) y el conjunto Γ_0 que fueron definidos con anterioridad.

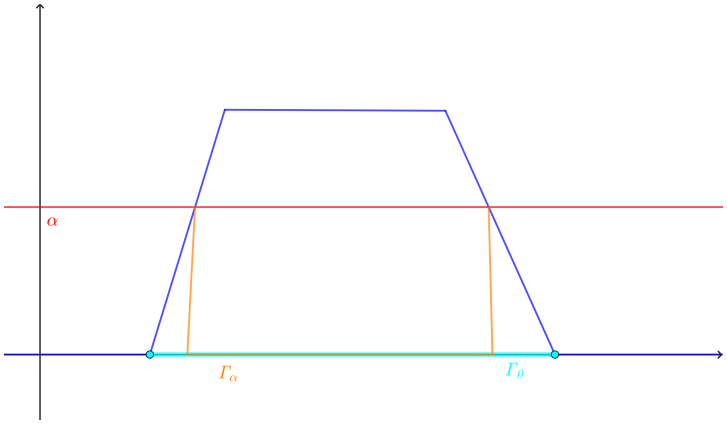

Figura 2.1: Gráfico de un número difuso Γ.

El símbolo $\mathcal{F}(\mathbb{R})$ representa el conjunto de todos los números difusos que se pueden definir en el conjunto de los números reales $\mathbb{R}$.

Definición 2.1.2. Un número difuso Γ se clasifica como trapezoidal si su función de pertenencia tiene la forma siguiente:

$$\Gamma'(x) = \begin{cases} 0 & si \quad x \leq l \\ \frac{x-l}{m-l} & si \quad l < x \leq m \\ 1 & si \quad m < x \leq n \\ \frac{p-x}{p-n} & si \quad n < x \leq p \\ 0 & si \quad p < x. \end{cases} \tag{2.1}$$

donde l, m, n y p son números reales, tales que $l < m \leq n < p$. Un número trapezoidal es denotado por (l, m, n, p).

La Figura 2.1 ilustra la Definición 2.1.2.

Observación 2.1.1. a) El caso en el que $m = n$ en (2.1) se denominará número difuso triangular y se denotará como (l, m, p). Y, considerando el caso degenerado en el

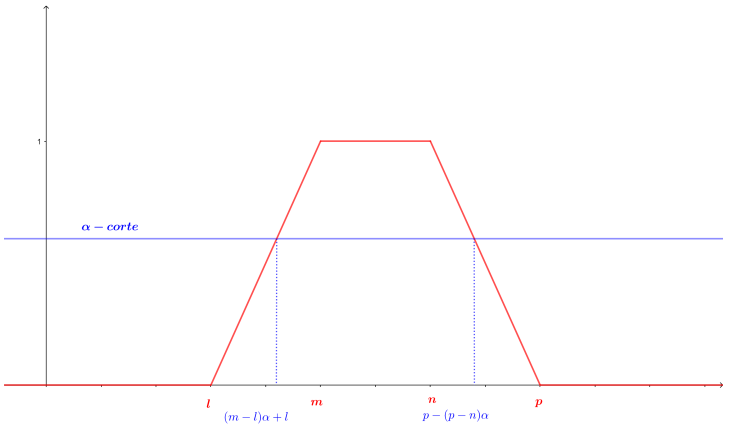

Figura 2.2: α-corte de un número difuso trapezoidal Γ.

que $l = m = p$ en un número triangular, la representación difusa del número real m se obtiene con la función de pertenencia dada por:

$$m'(x) = \begin{cases} 1 & si \quad x = m \\ 0 & si \quad x \neq m \end{cases} \tag{2.2}$$

b) Para un número difuso trapezoidal $\Gamma = (l, m, n, p)$ puede demostrarse que los correspondientes $\alpha-$cortes están dados por

$$\Gamma_\alpha = [(m-l)\alpha + l, p - (p-n)\alpha], \quad \alpha \in [0, 1]$$

2.2. Aritmética Difusa

En esta sección, se presentan las operaciones básicas de la aritmética difusa incluyendo la suma y el producto de números difusos.

Definición 2.2.1. Sean Γ y Υ números difusos. Si "$\star$" denota la adición o multiplicación entre números difusos, entonces

se define el conjunto difuso en $\mathbb{R}$, $\Gamma \star \Upsilon$, por la función de pertenencia:

$$(\Gamma \star \Upsilon)'(u) = \sup_{u=x\star y} \text{mín}\{\Gamma'(x), \Upsilon'(y)\}, \tag{2.3}$$

para todo $u \in \mathbb{R}$ [21].

Como consecuencia de la Definición 2.2.1, es posible obtener el siguiente resultado para números trapezoidales.

Lema 2.2.1. Si $H = (a_l, a_m, a_n, a_p)$ e $I = (b_l, b_m, b_n, b_p)$ son dos números difusos trapezoidales y λ un número real positivo, entonces tenemos lo siguiente:

a) $H +^* I = (a_l + b_l, a_m + b_m, a_n + b_n, a_p + b_p)$,

b) $\sum_{t=0}^{N}{}^*(a_l^t, a_m^t, a_n^t, a_p^t) = \left(\sum_{t=0}^{N} a_l^t, \sum_{t=0}^{N} a_m^t, \sum_{t=0}^{N} a_n^t, \sum_{t=0}^{N} a_p^t\right)$,

c) $\lambda H = (\lambda a_l, \lambda a_m, \lambda a_n, \lambda a_p)$.

Para realizar la prueba de este lema se usa la definición de α-cortes de un número trapezoidal, como se expresa en el inciso b) de la Observación 2.1.1 para encontrar la suma de dos números difusos trapezoidales (H, I). Se muestra que la suma es un número trapezoidal con parámetros $H +^* I = (a_l + b_l, a_m + b_m, a_n + b_n, a_p + b_p)$. Para la parte b) se prueba por inducción que la suma de los N números difusos trapezoidales es un número difuso trapezoidal con parámetros $\left(\sum_{t=0}^{N} a_l^t, \sum_{t=0}^{N} a_m^t, \sum_{t=0}^{N} a_n^t, \sum_{t=0}^{N} a_p^t\right)$. Y para la parte c) se aplica el resultado de la parte b) para probar que el producto de un número real positivo λ y un número difuso trapezoidal H con parámetros $\lambda a_l, \lambda a_m, \lambda a_n, \lambda a_p)$.

Es posible dotar a los números difusos de una estructura de espacio métrico a partir de una métrica definida en el

conjunto de intervalos compactos. Sea $\mathbb{D}$ el conjunto de todos los intervalos acotados y cerrados en la recta real $\mathbb{R}$. Para $\Psi = [a_l, a_u]$, $\Phi = [b_l, b_u] \in \mathbb{D}$ definimos

$$d(\Psi, \Phi) = \text{máx}(|a_l - b_l|, |a_u - b_u|). \tag{2.4}$$

Es posible verificar que d define una métrica sobre $\mathbb{D}$ y que $(\mathbb{D}, d)$ es un espacio métrico completo [1]. Ahora, si $\tilde{\eta} \in \mathfrak{F}(\mathbb{R})$, entonces $\tilde{\eta}_\alpha$ es un conjunto compacto porque su función de pertenencia es semi-continua superiormente y tiene soporte compacto. Por lo tanto, se define la función $\hat{d} : \mathfrak{F}(\mathbb{R}) \times \mathfrak{F}(\mathbb{R}) \to \mathbb{R}$ por

$$\hat{d}(\tilde{\eta}, \tilde{\mu}) = \sup_{\alpha \in [0,1]} d(\tilde{\eta}_\alpha, \tilde{\mu}_\alpha), \text{para} \tilde{\eta}, \tilde{\mu} \in \mathfrak{F}(\mathbb{R}). \tag{2.5}$$

Se puede ver que $\hat{d}$ es una métrica en $\mathfrak{F}(\mathbb{R})$.

Lema 2.2.2. [33] El espacio métrico $(\mathfrak{F}(\mathbb{R}), \hat{d})$ es completo.

Una vez definida una métrica en el conjunto de números difusos, es posible introducir la noción de convergencia.

Definición 2.2.2. Una sucesión $\{\tilde{\eta}_t\}_{t=0}^{\infty}$ de números difusos se dice que es convergente al número difuso $\tilde{\mu}$, escribimos como $\lim^*_{t\to\infty} \tilde{\eta}_t = \tilde{\mu}$ si y solo si $\hat{d}(\tilde{\eta}_t, \tilde{\mu}) \to 0$ cuando $t \to \infty$.

Del Lema 2.2.1 b) y de la métrica (2.5), se puede verirficar el siguiente lema.

Lema 2.2.3. Si $\{y_k = (l_k, m_k, n_k, p_k) : k \geq 1\}$ es una sucesión de números difusos tr´pezoidales tales que $\sum_{k=1}^{\infty} l_k$, $\sum_{k=1}^{\infty} m_k$, $\sum_{k=1}^{\infty} n_k$, y $\sum_{k=1}^{\infty} p_k$ convergen, entonces $\sum_{k=1}^{t}{}^* y_k$ converge cuando $t \to \infty$ al número trapezoidal:

$$\left(\sum_{k=1}^{\infty} l_k, \sum_{k=1}^{\infty} m_k, \sum_{k=1}^{\infty} n_k, \sum_{k=1}^{\infty} p_k \right).$$

Ahora, para poder establecer una comparación entre números difusos trapezoidales, consideremos $\tilde{\eta}, \tilde{\mu} \in \mathfrak{F}(\mathbb{R})$, con $\alpha-$cortes $\tilde{\eta}_\alpha = [a_\alpha, b_\alpha]$ y $\tilde{\mu}_\alpha = [c_\alpha, d_\alpha]$, $\alpha \in [0,1]$, respectivamente y definimos el orden parcial sobre $\mathfrak{F}(\mathbb{R})$ "$(\leq^*)$" por

$$\tilde{\eta} \leq^* \tilde{\mu} \text{ si y solo si } a_\alpha \leq c_\alpha \text{ y } b_\alpha \leq d_\alpha \text{ para todo } \alpha \in [0,1]. \tag{2.6}$$

Observación 2.2.1. Sean $w, z \in \mathbb{R}$, y $\tilde{w}$ y $\tilde{z}$ números difusos con función de pertenencia dada por $(\tilde{w})'(x) = 1$, $x = w$ y $(\tilde{w})'(x) = 0$, $x \neq w$ y $(\tilde{z})'(x) = 1$, $x = z$ y $(\tilde{z})'(x) = 0$ $x \neq z$. Entonces, se puede ver que $\tilde{w} \leq^* \tilde{z}$ es equivalente a $w \leq z$.

También es posible comparar números difusos mediante la denominada "ranking promedio", dado para el número difuso $\tilde{\eta} = (a_l, a_m, a_n, a_p)$ por

$$R(\tilde{\eta}) := \frac{1}{4}(a_l + a_m + a_n + a_p).$$

Por tanto, se dirá que los números difusos trapezoidales $\tilde{\eta} = (a_l, a_m, a_n, a_p)$ y $\tilde{\mu} = (a_l^1, a_m^1, a_n^1, a_p^1)$ son tales que $\tilde{\eta} \leq^{**} \tilde{\mu}$ si y sólo si $R(\tilde{\eta}) \leq R(\tilde{\mu})$.

2.3. Probabilidad Difusa

Una probabilidad difusa es una función que asigna a cada evento de un espacio de probabilidad un conjunto difuso de números reales, en lugar de un número real.

En este capítulo se establecen definiciones necesarias para formalizar los conceptos concernientes a la probabilida difusa, siendo particularmente relevantes las variables aleatorias difusas y sus correspondientes esperanzas difusas. Para esto,

$\mathbb{C}(\mathbb{R})$ denota la clase de subconjuntos compactos no vacíos de $\mathbb{R}$, y si $(\Omega_1, \mathcal{A}_1)$ y $(\Omega_2, \mathcal{A}_2)$ son espacios medibles, entonces $\mathcal{A}_1 \bigotimes \mathcal{A}_2$ denota la correspondiente $\sigma-$álgebra producto asociada al espacio producto $\Omega_1 \times \Omega_2$.

Definición 2.3.1. Sean $(\Omega, \mathcal{F})$ un espacio medible y $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ el espacio medible del conjunto de los números reales. Una función $\tilde{Y} : \Omega \to \mathfrak{F}(\mathbb{R})$ se dice que es variable aleatoria difusa si $Gr(\tilde{Y}_\alpha) = \{(\omega, x) \in \Omega \times \mathbb{R} | x \in (\tilde{Y}(\omega))_\alpha\} \in \mathcal{F} \bigotimes \mathcal{B}(\mathbb{R})$, para todo $\alpha \in [0, 1]$.

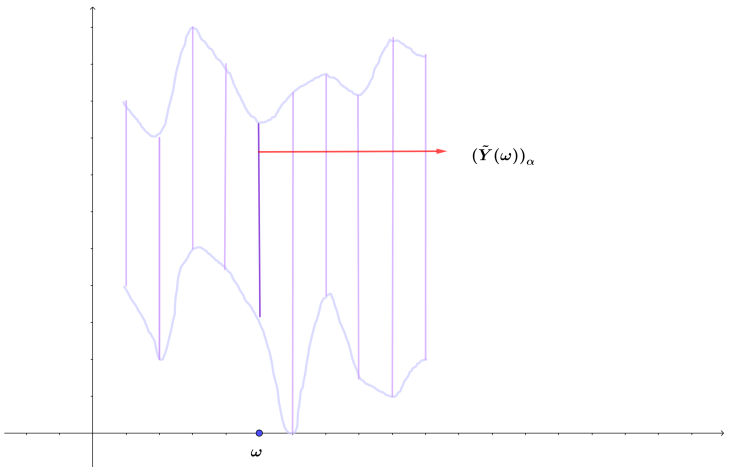

Figura 2.3: Gráfica de $(\tilde{Y})_\alpha$.

La Definición 2.3.1 se ilustra en la Figura 2.3, para algún $\omega \in \Omega$ fijo y $\alpha \in [0, 1]$.

Es decir, una variable aleatoria difusa es una función que asigna a cada resultado de un experimento aleatorio un conjunto difuso de números reales. La siguiente definición será de ayuda para poder definir la esperanza de variables aleatorias difusas.

Definición 2.3.2. Dado un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ una variable aleatoria difusa $\tilde{Y}$ asociada a $(\Omega, \mathcal{A})$ se dice que es una variable aleatoria difusa integrablemente acotada con respecto a $(\Omega, \mathcal{A}, \mathbb{P})$ si existe una función $h : \Omega \to \mathbb{R}$, $h \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ tal que $|x| \leq h(\omega)$, para todo $(\omega, x) \in \Omega \times \mathbb{R}$ con $x \in (\tilde{Y}(\omega))_0 := \tilde{Y}_0(\omega)$.

Definición 2.3.3. Dada una variable aleatoria difusa integrablemente acotada $\tilde{Y}$ asociada con respecto al espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, se define el valor esperado difuso de $\tilde{Y}$ en el sentido de Aumann como el único conjunto difuso de $\mathbb{R}$, $\mathbb{E}^*[\tilde{Y}]$ tal que para cada $\alpha \in [0, 1]$:

$$(\mathrm{E}^*[\tilde{Y}])_\alpha = \left\{ \int_\Omega f(\omega) d\mathbb{P}(\omega) | f : \Omega \to \mathbb{R}, f \in \mathcal{L}^1(\mathbb{P}), f(\omega) \in (\tilde{Y}(\omega))_\alpha \quad [\mathbb{P}] \text{ c.s.} \right\}. \tag{2.7}$$

Ahora que hemos definido la esperanza de una variable aleatoria difusa, podemos considerar un caso específico en el que la variable aleatoria difusa se define en términos de una variable aleatoria clásica.

Lema 2.3.1. Sean $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad fijo y Y una variable aleatoria no negativa asociada a $(\Omega, \mathcal{A}, \mathbb{P})$ tal que $\mathbb{E}[Y]$ existe. Entonces, $\tilde{Y} := Y(B, C, D, F)$ (donde $0 \leq B \leq C \leq D \leq F$) es una variable aleatoria difusa asociada a $(\Omega, \mathcal{A}, \mathbb{P})$ y

$$\mathbb{E}^*[\tilde{Y}] = \mathbb{E}[Y](B, C, D, F). \tag{2.8}$$

Demostración. Sea Y una variable aleatoria discreta no negativa con rango finito o numerable denotado por $Y[\Omega] = \{y_1, y_2, \ldots\}$ y consideremos los conjuntos medibles $[Y = y_j] := \{\omega \in \Omega : Y(\omega) = y_j\}$, $j = 1, 2, \ldots$ Sea $\Gamma = (B, C, D, F)$ con α-cortes $\Gamma_\alpha = [q(\alpha), s(\alpha)]$, $\alpha \in [0, 1]$. Fijamos $\alpha \in [0, 1]$. Consideramos la siguiente multifunción dada por:

$$\begin{aligned} \tilde{Y} &:= (\tilde{Y}(\omega))_\alpha \\ &= (Y(\omega)\Gamma)_\alpha \\ &= Y(\omega)[q(\alpha), s(\alpha)], \qquad \omega \in \Omega. \end{aligned} \tag{2.9}$$

Ahora notemos que

$$\begin{aligned} Gr(\tilde{Y}_\alpha) &= \{(\omega, x) \in \Omega \times \mathbb{R} | x \in \tilde{Y}_\alpha(\omega)\} \\ &= \{(\omega, x) \in \Omega \times \mathbb{R} | x \in Y(\omega)[q(\alpha), s(\alpha)]\} \\ &= \bigcup_j ([Y = y_j] \times y_j[q(\alpha), s(\alpha)]) \,. \end{aligned}$$

Por lo tanto, $Gr(\tilde{Y}_\alpha) \in \mathcal{A} \bigotimes \mathfrak{B}(\mathbb{R})$. Como α es arbitraria, de la Definición 2.3.1 se deduce que $\tilde{Y}$ es una variable aleatoria difusa. Note que, para cada $\omega \in \Omega$, $\tilde{Y}_0(\omega) = Y(\omega)[B, F]$. Definamos $h : \Omega \to \mathbb{R}$ dada por $h(\omega) := Y(\omega)F$, $\omega \in \Omega$. Entonces, $|x| \leq h(\omega)$, $(\omega, x) \in \Omega \times \mathbb{R}$ con $x \in Y(\omega)[B, F]$. Además, claramente $\mathbb{E}[h] = F\mathbb{E}[Y]$ es finito. De la Definición 2.3.2, $\tilde{Y}$ es una variable aleatoria difusa integrable acotada con respecto a $(\Omega, \mathcal{A}, \mathbb{P})$. Ahora, a partir de Definición 3.9, existe un único valor esperado difuso $\mathbb{E}^*[\tilde{Y}]$, para cada α,

$$\begin{aligned} \mathbb{E}[Y]\Gamma_\alpha &= \int_\Omega Y(\omega) d\mathbb{P}(\omega)[q(\alpha), s(\alpha)] \\ &= \int_\Omega Y(\omega) d\mathbb{P}(\omega)\{x : x \in [q(\alpha), s(\alpha)]\} \\ &= \left\{ \int_\Omega Y(\omega) x d\mathbb{P}(\omega) : Y(\omega)x \in Y(\omega)[q(\alpha), s(\alpha)] \right\} \\ &= \Big\{ \int_\Omega f(\omega) d\mathbb{P}(\omega) | f : \Omega \to \mathbb{R}, f \in L^1(\mathbb{P}), \\ &\quad f(\omega) \in (\tilde{Y}(\omega))_\alpha \qquad c.s.[\mathbb{P}] \Big\} \\ &= (\mathbb{E}^*[\tilde{Y}])_\alpha, \end{aligned}$$

por lo que, $(\mathbb{E}^*[\tilde{Y}])_\alpha = \mathbb{E}[Y][q(\alpha), s(\alpha)]$ para cada α, es el α-corte del número trapezoidal dado para $\mathbb{E}[Y](B, C, D, F)$, es decir, $\mathbb{E}^*[\tilde{Y}] = \mathbb{E}[Y](B, C, D, F)$. ■

Algunas propiedades de la esperanza con variables aleatorias difusas (trapezoidales) se presentan en el siguiente lema.

Lema 2.3.2. Sean $\tilde{X}$ y $\tilde{Y}$ variables aleatorias difusas (trapezoidales). Entonces,

a) $\mathbb{E}[\tilde{X}] \in \mathfrak{F}(\mathbb{R})$,

b) $\mathbb{E}[\tilde{X} + \tilde{Y}] = \mathbb{E}[\tilde{X}] + \mathbb{E}[\tilde{Y}]$,

c) $\mathbb{E}[\lambda\tilde{X}] = \lambda\mathbb{E}[\tilde{X}]$, $\lambda \geq 0$.

La teoría difusa ofrece una herramienta poderosa para modelar y analizar la incertidumbre y la imprecisión en diversos campos. Hemos presentado los conceptos fundamentales de la teorá difusa, incluyendo variables aleatorias difusas y esperanza difusa, que consideramos son una base sólida para los siguientes capítulos.

Capítulo 3

Modelo de Bandidos Armados con Recompensas Difusas Trapezoidales

En este capítulo, se presenta una extensión del problema de Bandidos Armados, en donde las recompensas son representadas por números difusos trapezoidales. Se define el modelo de Bandidos Armados Difusos y se presenta la noción de recompensa total esperada descontada difusa. También se calcula de supremo difuso, aprevechando el orden parcial antes definido, para representar el máximo valor posible de la recompensa total difusa. Finalmente, se establecen las bases para la definición de un índice de Gittins difuso, que permitirá caracterizar la política de selección óptima en este entorno difuso. Éste capítulo está basado en el trabajo "Fuzzy Rewards on the Multi-Armed Bandits Model" [12].

3.1. Modelo BA con recompesa difusa

En muchas ocasiones, las recompensas asociadas con las posibles opciones no son precisas y pueden variar en un rango determinado. Para capturar esta incertidumbre, se pueden utilizar números difusos trapezoidales para representar las recompensas. En primer lugar considere la suposición:

Suposición 3.1.1. Sean $\Delta = (b, d, f, k)$ y $\Delta_1 = (b_1, d_1, f_1, k_1)$ dos números difusos trapezoidales fijos, con $0 \leq b < d \leq f < k$ y $0 \leq b_1 < d - 1 \leq f_1 < k_1$. También se supone que

$$\tilde{r}(x) = r(x)\Delta + \Delta_1, \tag{3.1}$$

con $x \in \mathbb{X}$, y r es una función de recompensa como se consideró en el Capítulo 3, y tal que se cumple la Suposición 1.2.1.

Esta suposición se utiliza para garantizar que las recompensas difusas sean no negativas y tengan un rango finito.

En resumen, $r(x)$ es una función de recompensa que asigna una recompensa a cada estado x, Δ es un número difuso trapezoidal que representa la incertidumbre en la recompensa y Δ_1 es un número difuso trapezoidal que representa la incertidumbre adicional en la recompensa.

La siguiente observación destaca algunas propiedades importantes de la recompensa difusa $\tilde{r}(x)$ definida en la ecuación (3.1).

Observación 3.1.1.

a) Tenga en cuenta que $0 \leq^* r(x), x \in \mathbb{X}$, como una consecuencia de las Suposiciones 1.2.1 y 3.1.1 y Definición 1.

b) Observe que los $\alpha-$cortes de los números difusos trapezoidales Δ y Δ_1 están dados para $\alpha \in (0, 1]$, por:

$\Delta_\alpha = [b(\alpha), c(\alpha)]$ y $(\Delta_1)_\alpha = [b_1(\alpha), c_1(\alpha)]$, donde $b(\alpha) = \alpha(d-b)+b$, $c(\alpha) = k-\alpha(k-f)$, $b_1(\alpha) = \alpha(d_1-b_1)+b_1$ y $c_1(\alpha) = k_1 - \alpha(k_1 - f_1)$.

c) La intuición detrás de la expresión (3.1) es que en lugar de recibir una recompensa $r(x)$, uno obtendrá una recompensa que está aproximadamente dentro del intervalo $[dr(x)+d_1, fr(x)+f_1]$. Esta expresión se utiliza para representar la recompensa total esperada descontada difusa porque permite capturar la incertidumbre asociada con las recompensas, al mismo tiempo que mantiene la estructura básica de la recompensa esperada descontada clásica.

Siguiendo una lógica similar a (1.3), se define; para $\pi \in \Pi$, con Π el conjunto de todas las políticas posibles para el modelo de Bandidos Armados en donde π una política que define la acción a tomar en cada estado $\mathbf{x} \in \mathbb{X}^K$ donde $\mathbb{X}^K$ y X_i significan lo mismo Capítulo **??**. Así, la recompensa total esperada descontada difusa se representa por:

$$\tilde{V}(\pi, \mathbf{x}) := \mathbb{E}^*_{\mathbf{x}} \left[\sum_{t=0}^{\infty}{}^* \beta^t \sum_{i=1}^{K}{}^* \tilde{r}_i(X_i(t)) a_i(t) \right]. \tag{3.2}$$

Donde la esperanza y las sumas son en el sentido difuso 2.2.3.

Entonces, el problema de Bandidos Armados Difusos consiste en determinar una estrategia de activación $\pi_0 \in \Pi$ que maximice para cualquier criterio de orden difuso (basado en alguno de los dos criterios considerados), la recompensa total descontada difusa, es decir, tal que

$$\tilde{V}(\pi, \mathbf{x}) = \sup_{\pi \in \Pi}{}^* \tilde{V}(\pi, \mathbf{x}), \tag{3.3}$$

$\mathbf{x} \in \mathbb{X}^K$. La función $\tilde{V}_0(\mathbf{x}) := \tilde{V}(\pi_0, \mathbf{x})$ se llama función de valor óptimo difuso.

Ésta función representa el máximo valor posible de la recompensa total descontada difusa que se puede obtener siguiendo la política óptima π_0 con estado inicial $\mathbf{x}$. Es decir, la función de valor óptimo difuso proporciona la recompensa total descontada difusa máxima que se puede obtener en cada estado inicial $\mathbf{x}$, asumiendo que se sigue la política óptima π_0. Esto permite evaluar la eficiencia de diferentes políticas en términos de la recompensa total descontada difusa.

El objetivo ahora es proponer una versión difusa del índice de Gittins para caracterizar la política de selección óptima en el marco difuso. Para esto, considere los siguientes conjuntos:

$$S_F := \{\omega \in \Omega : \sum_{t=0}^{\infty} r(X(t,\omega)) < \infty\} \text{ y}$$

$$S_I := \{\omega \in \Omega : \sum_{t=0}^{\infty} r(X(t,\omega)) = \infty\}$$

. Luego, definimos la variable aleatoria:

$$Y(\omega) = \begin{cases} \sum_{t=0}^{\tau-1} \beta^t r(X(t,\omega)) & \text{si} \quad \omega \in [\tau < \infty] \bigcup ([\tau = \infty] \bigcap S_F) \\ 0 & si \qquad \omega \in [\tau = \infty] \bigcap S_I \end{cases} \tag{3.4}$$

Observe que $Y(\omega)$ es una función bien definida debido a

$$\Omega = [\tau < \infty] \bigcup ([\tau = \infty] \cap S_F) \bigcup ([\tau = \infty] \cap S_I) \text{ y}$$

$$\mathbb{P}([\tau = \infty] \cap S_I) = 0.$$

Como consecuencia,

$$\sum_{t=0}^{\tau-1} \beta^t r(X(t)) = Y, \mathbb{P}_X - \text{c.s.} \tag{3.5}$$

En la siguiente subsección, vamos a definir una variable aleatoria difusa $\tilde{Y}$ que nos permita modelar la recompensa total descontada difusa.

3.2. Índice de Gittins difuso

Se presentará ahora alguna notación necesaria para discutir de forma apropiada sobre el rango de la variable aleatoria Y dada por (3.4). Es decir, Y está definida como la suma de las recompensas instantáneas descontadas hasta el tiempo τ, o como 0 si el proceso estocástico $X(t,\omega)$ no converge a una recompensa finita en el infinito.

Para esto, tenga en cuenta que $[\tau < +\infty] = \bigcup_{n=1}^{+\infty} [\tau = n]$. Por lo tanto, el comportamiento de Y restringido a cada uno de los elementos de la partición del espacio muestral son los siguientes:

(a) En $[\tau = n]$; para $n \geq 1$, la imagen de Y se denota por $Y[\tau = n] = \{y_1^n, y_2^n, \ldots\}$.

(b) Además, $Y\left[[\tau = +\infty] \cap S_F\right] = Y\left[\bigcup_{n=1}^{+\infty} [\tau = n]\right]$ que es un conjunto numerable.

1. Y, finalmente, $Y\left[[\tau = +\infty] \cap S_I\right] = \{0\}$.

Ahora, consideremos la variable aleatoria:

$$Z := \sum_{t=0}^{\tau-1} \beta^t < \infty, \, \mathbb{P}_x - \text{c.s.}, \tag{3.6}$$

cuyo rango adopta la notación $Z[\tau = n] = \{z_n\}$, $n = 1, 2, \ldots$, y $Z[\tau = +\infty] = \{z\}$.

Los hechos anteriores se utilizarán en la prueba del siguiente Lema, que es el primer paso para proponer una versión difusa del índice de Gittins.

Lema 3.2.1. Sea

$$\tilde{Y}(\omega) := \begin{cases} \sum_{t=0}^{\tau-1}{}^{*} \beta^t \tilde{r}(X(t,\omega)) & \text{si} \quad \omega \in [\tau < \infty] \bigcup ([\tau = \infty] \cap S_F) \\ z\Delta_1 & si \qquad \omega \in [\tau = \infty] \cap S_I \end{cases} \tag{3.7}$$

Entonces $\tilde{Y}$ es una variable aleatoria difusa.

Demostración. Fijemos $\alpha \in [0,1]$. En primer lugar observe que $\tilde{Y}$ definida en (3.7) se puede representar como

$$\tilde{Y} = Y\Delta +^* Z\Delta_1 \tag{3.8}$$

como consecuencia de (3.4) y (3.6).

Entonces, según la Observación 3.1.1, el $\alpha-$corte de $\tilde{Y}$ esta dado por

$$\begin{aligned} \tilde{Y}_\alpha &= Y\Delta_\alpha + Z(\Delta_1)_\alpha \\ &= [Yb(\alpha) + Zb_1(\alpha), Yc(\alpha) + Zc_1(\alpha)]. \end{aligned}$$

Así, la gráfica de $\tilde{Y}_\alpha$ se puede escribir de la siguiente manera:

$$\begin{aligned} Gr(\tilde{Y}_\alpha) &= \{(\omega,x) \in \Omega \times \mathbb{R} : x \in \tilde{Y}(\omega)_\alpha\} \\ &= \left\{(\omega,x) \in [\tau < +\infty] \bigcup ([\tau = +\infty] \cap S_F) \times \mathbb{R} : x \in \tilde{Y}(\omega)_\alpha\right\} \\ &\quad \bigcup \{(\omega,x) \in ([\tau = +\infty] \cap S_I) \times \mathbb{R} : x \in \tilde{Y}(\omega)_\alpha\} \\ &= \bigcup_{n=1}^{\infty} \bigcup_{j=1}^{\infty} \left(\left([Y = y_j^n] \cap [Z = z_n]\right) \times [y_j^n b(\alpha) + z_n b_1(\alpha), y_i^n c(\alpha) + z_n c_1(\alpha)]\right) \\ &\quad \bigcup \left(([Y = 0] \cap [Z = z]) \times [zb_1(\alpha), zc_1(\alpha)]\right). \end{aligned}$$

En consecuencia, $Gr(\tilde{Y}_\alpha) \in \mathcal{F} \bigotimes \mathfrak{B}(\mathbb{R})$. Como α es arbitraria, se obtiene como consecuencia de la Definición 2.3.1 que $\tilde{Y}$ es una variable aleatoria difusa.

■

Lema 3.2.2. En el contexto antes descrito se obtiene que $\tilde{Y}$ es una variable aleatoria difusa integrable acotada con respecto a $(\Omega, \mathcal{F}, \mathbb{P}_x)$ y además

$$\mathbb{E}_x^*[\tilde{Y}] = \mathbb{E}_x[Y]\Delta + \mathbb{E}_x[Z]\Delta_1,\ x \in \mathbb{X}. \tag{3.9}$$

Demostración. Sea $x \in \mathbb{X}$ fija. Para demostrar que $\tilde{Y}$ es integrablemente acotada, note que de (3.8) y la Observación 3.1.1, se cumplen las siguientes identidades

$$\begin{aligned}(\tilde{Y}(\omega))_0 &= (Y(\omega)\Delta + Z(\omega)\Delta_1)_0 \\ &= (Y(\omega)(b,d,f,k) + Z(\omega)(b_1,d_1,f_1,k_1))_0 \\ &= (Y(\omega)b + Z(\omega)b_1, Y(\omega)d + Z(\omega)d_1, Y(\omega)f + Z(\omega)f_1, Y(\omega)k + Z(\omega)k_1)_0\end{aligned}$$

para cada $\omega \in \Omega$. Por lo tanto, $(\tilde{Y}(\omega))_0 = [bY(\omega)+Z(\omega)b_1, Y(\omega)k+ Z(\omega)k_1], \omega \in \Omega$.

Definimos la función $h : \Omega \to \mathbb{R}$ dada por,

$$h(\omega) = Y(\omega)k + Z(\omega)k_1,\ \omega \in \Omega.$$

Entonces, si $(\omega, x) \in \Omega \times \mathbb{R}$ con $x \in \tilde{Y}_0(\omega)$, se obtiene que:

$$bY(\omega) + Z(\omega)b_1 \leq x \leq Y(\omega)k + Z(\omega)k_1,$$

esto implica que $|x| \leq h(\omega)$. Dado que $\mathbb{E}_x[h] < \infty$, en primer lugar observe que para $\alpha \in [0,1]$, las siguientes igualdades son válidas.

$$\begin{aligned}(\mathbb{E}_x^*[\tilde{Y}])_\alpha &= \left\{\int_\Omega f d\mathbb{P} | f \in L^1, f(\omega) \in (\tilde{Y}(\omega))_\alpha \mathbb{P}_x - \text{c.s.}\right\} \\ &= \left\{\int_\Omega f d\mathbb{P} | f : \Omega \to \mathbb{R}, f \in L^1,\ f(\omega) \in [Y(\omega)b(\alpha) + \right. \\ &\quad \left. Y(\omega)c(\alpha) + Z(\omega)c_1(\alpha)]\right\}.\end{aligned}$$

Se tiene entonces que,

$$Y(\omega)b(\alpha)+Z(\omega)b_1(\alpha) \leq f(\omega) \leq Y(\omega)c(\alpha)+Z(\omega)c_1(\alpha),\ \mathbb{P}_x-\text{c.s.},$$

al tomar esperanzas encontramos que

$$\mathbb{E}_x[Y]b(\alpha) + \mathbb{E}_x[Z]b_1(\alpha) \leq \int f d\mathbb{P}_x \leq \mathbb{E}_x[Y]c(\alpha) + \mathbb{Z}c_1(\alpha). \tag{3.10}$$

Sea I_α el $\alpha-$corte del número difuso $\mathbb{E}_x[Y]\Delta + \mathbb{E}_x[Z]\Delta_1$, es decir,

$$I_\alpha := [\mathbb{E}_x[Y]b(\alpha) + \mathbb{E}_x[Z]b_1(\alpha), \mathbb{E}_x[Y]c(\alpha) + \mathbb{E}_x[Z]c_1(\alpha)].$$

En consecuencia, de (3.10) se deduce que

$$(\mathbb{E}_x[\tilde{Y}])_\alpha \subseteq I_\alpha. \tag{3.11}$$

Por otro lado, sean $y \in I_\alpha$ y $f : \Omega \to \mathbb{R}$ dada por $f(\omega) = y$, $\omega \in \Omega$. Ahora, supongamos que $\forall\omega \in \Omega$ $y \notin [Y(\omega)b(\alpha) + Z(\omega)b_1(\alpha), Y(\omega)c(\alpha) + Z(\omega)c_1(\omega)]$, entonces

$$\int_\Omega f d\mathbb{P}_x = y \notin I_\alpha,$$

lo cual es una contradicción. Por tanto, existe $\bar{\omega} \in \Omega$ tal que:

$$y \in [Y(\bar{\omega})b(\alpha) + Z(\bar{\omega})b_1(\alpha), Y(\bar{\omega})c(\alpha) + Z(\bar{\omega})c_1(\alpha)]$$

sin embargo, dado que $f(\omega) = y$, para todo $\omega \in \Omega$, esto implica que:

$$y \in [Y(\omega)b(\alpha) + Z(\omega)b_1(\alpha), Y(\omega)c(\alpha) + Z(\omega)c_1(\alpha)],\ \forall\omega \in \Omega.$$

En consecuencia, $y \in (\mathbb{E}^*_x[\tilde{Y}])_\alpha$, es decir,

$$I_\alpha \subseteq (\mathbb{E}^*[\tilde{Y}])_\alpha. \tag{3.12}$$

Entonces, de (3.11) y (3.12) se obtiene que, $(\mathbb{E}^*[\tilde{Y}])_\alpha = I_\alpha$. Dado que $x \in \mathbb{X}$ es arbitrario, los argumentos anteriores demuestran la validez de (3.9). ■

En el Capítulo 1 vimos que el índice de Gittins es un concepto fundamental en el modelo de Bandidos Armados, que se utiliza para evaluar la eficiencia de diferentes estrategias de decisión en entornos estocásticos. En el contexto de la teoría de Bandidos Armados difusa, el índice de Gittins difuso se define como una extensión del índice de Gittins clásico, que

incorpora la indertidumbre y la imprecisión de la información disponible. Esto permite evaluar la eficiencia de diferentes estrategias de decisión en entornos estocásticos difusos, donde la información disponible es incierta o imprecisa.

Ahora se definirá una versión difusa del índice de Gittins analizado en el Capítulo 3.

Definición 3.2.1. Sea τ un tiempo de paro asociado al proceso estocástico $\{X(t)\}$. El índice difuso de Gittins se define como:

$$\tilde{\mathbb{G}}(x) = \sup_{\tau>0}{}^{*} \tilde{G}(x,\tau) = \sup_{\tau>0}{}^{*} \frac{\mathbb{E}_x^* \left[\sum_{t=0}^{\tau-1}{}^{*} \beta^t \tilde{r}(X(t)) \right]}{\mathbb{E}_x \left[\sum_{t=0}^{\tau-1} \beta^t \right]} \tag{3.13}$$

para todo $x \in \mathbb{X}$.

El siguiente teorema relaciona las versiones nítida y difusa del índice de Gittins. Además, muestra que el tiempo óptimo de paro en ambos paradigmas coincide.

Teorema 3.2.1. En el contexto abordado, las siguientes afirmaciones son válidas.

(a) $\tilde{G}(x,\tau) = G(x,\tau)\Delta +^{*} \Delta_1$, $x \in \mathbb{X}$ y τ un tiempo de paro respecto al proceso estocástico $\{X(t)\}$.

(b) Si τ_0 es un tiempo de paro tal que $\mathbb{G}(x) = G(x,\tau_0)$, $x \in \mathbb{X}$ entonces $\tilde{\mathbb{G}}(x) = \tilde{G}(x,\tau_0)$, $x \in \mathbb{X}$.

Demostración. Sean $x \in \mathbb{X}$ y τ un tiempo de paro fijo. La

validez de (a) es una consecuencia de las siguientes igualdades:

$$\begin{aligned}
\tilde{G}(x,\tau) &= \frac{\mathbb{E}_x^*\left[\sum_{t=0}^{\tau-1}{}^{*}\,\beta^t\tilde{r}(X(t))\right]}{\mathbb{E}_x\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&= \frac{\mathbb{E}_x[Y]\Delta +^* \mathbb{E}_x[Z]\Delta_1}{\mathbb{E}_x[Z]} \\
&= \frac{\mathbb{E}_x[Y]}{\mathbb{E}_x[Z]}\Delta +^* \Delta_1 \\
&= G(x,\tau)\Delta +^* \Delta_1.
\end{aligned}$$

Es decir, el índice de Gittins difuso es una extensión del índice de Gittins clásico, en él se incorpora la incertidumbre y la imprecisión de la información disponible a través de los números difusos Δ y Δ_1.

Para probar (b) observe que

$$(\tilde{G}(x,\tau))_\alpha = [G(x,\tau)b(\alpha) + b_1(\alpha), G(x,\tau)c(\alpha) + c_1(\alpha)].$$

Entonces,

- $G(x,\tau)b(\alpha) + b_1(\alpha) \leq G(x,\tau_0)b(\alpha) + b_1(\alpha) = \mathbb{G}(x)b(\alpha) + b_1(\alpha)$ y
- $G(x,\tau)c(\alpha) + c_1(\alpha) \leq G(x,\tau_0)c(\alpha) + c_1(\alpha) = \mathbb{G}(x)c(\alpha) + c_1(\alpha)$.

Por lo tanto, dado que x y τ son arbitrarios, el resultado se sigue como consecuencia de la relación $\tilde{G}(x,\tau_0) = \mathbb{G}(x)\Delta +^* \Delta_1$ observado de la parte (a) de este Teorema y de la Definición 3.2.1.

De manera similar, se puede proporcionar la prueba correspondiente considerando el método de rankig promedio. Para

ello observe que

$$\begin{aligned} R(\tilde{G}(x,\tau)) &= R(G(x,\tau)\Delta + \Delta_1) \\ &= \frac{BG(x,\tau) + DG(x,\tau) + FG(x,\tau) + KG(x,\tau) + B_1 + D_1 + F_1 + K_1}{4} \\ &= R(\Delta)G(x,\tau) + R(\Delta_1) \\ &\leq R(\Delta)\mathbb{G}(x) + R(\Delta_1). \end{aligned}$$

Con esto se concluye la demostración del Teorema. ■

Este Teorema posee implicaciones relevantes para completar la estructura difusa del modelo de Bandidos Armados con recompensa difusa.

Corolario 3.2.1. Las consecuencias inmediatas del Teorema 3.2.1 son

1. El tiempo de paro óptimo de los modelos nítido y difuso coinciden.

2. La estrategia de asignación para continuar con el proceso Bandido con el índice más alto (con respecto a cualquier orden considerado en este trabajo), es óptima.

3. Las funciones de valor óptimo difusa y nítida se relacionan de la siguiente manera

$$\tilde{V}_0(\mathbf{x}) = V_0(\mathbf{x})\Delta +^* \Delta_1$$

con $\mathbf{x} \in \mathbb{X}^K$.

Las consecuencias inmediatas del Corolario 3.2.1 son importantes porque proporcionan una conexión directa entre la versión clásica y la versión difusa del índice de Gittins. Esto permite aplicar los resultados y técnicas desarrollados en el contexto clásico a problemas que involucran incertidumbre e imprecisión.

Observación 3.2.1. La parte 3 del Corolario anterior afirma que la función de valor óptimo está aproximadamente en el intervalo $[dV_0(\mathbf{x}) + d_1, fV_0(\mathbf{x}) + f_1]$ para $\mathbf{x} \in \mathbb{X}^K$.

3.3. Ejemplo

En esta Sección se trabajará un problema de programación propuesto por Gittins [23]. Para este fin considere una máquina capaz de procesar un trabajo a la vez de un total de $N \geq 1$. Al inicio de cada periodo de procesamiento; digamos t,se debe seleccionar uno de estos trabajos para ser ejecutado independientemente de que el trabajo anterior se haya completado o no. Asociado con el trabajo i y su correspondiente tiempo de procesamiento (el número de veces que el trabajo i ha sido elegido para su ejecución durante las primeras t unidades de tiempo) $t_i \geq 1$, existe una probabilidad $P_i(t_i)$ de que $t_i + 1$ periodos sean necesarios para completar dicho trabajo dado que se necesitan más de t_i periodos. Por lo tanto, para un trabajo incompleto, su estado coincide con su tiempo de procesamiento.

Además, se supondrá que, si en el tiempo de procesamiento $t_i \geq 1$ el trabajo i no se ha completado y ha sido seleccionado para ser procesado, entonces se recibirá la recompensa difusa siguiente:

$$P_i(t_i)V_i\Delta + \Delta_1 \tag{3.14}$$

en donde $\Delta = (b, d, f, k)$ y $\Delta_1 = (b_1, d_1, f_1, k_1)$ son númers difusos tales que $0 \leq b$ y $0 \leq b_1$, con $V_i \geq 0$ para cada i. Una vez que se haya terminado algún trabajo,éste se producirá una recompensa nula en caso de que se elija este trabajo para ser procesado.

Es importante destacar que la recompensa difusa definida en la ecuación (3.14) combina la probabilidad de completar el trabajo con la recompensa asociada a él.

De aquí en adelante, se considerará que la Suposición 1.2.1 se cumple para todos los trabajos susceptibles de ser procesados, esto nos permitirá aplicar la teoría de procesos de Bandidos Armados difusos y obtener resultados óptimos en entornos

estocásticos difusos. Algunas situaciones en las que se cumple tal supuesto son:

1. (Paciencia finita) Para cada trabajo i, existe un tiempo máximo para su conclusión (M_i); no necesariamente el mismo. Esta suposición implica que todo espacio de estados es finito.

2. (Probabilidad constante de conclusión) Para cada trabajo i, y cada tiempo de procesamiento; t_i, se tiene que $P_i(t_i) = P_i$.

3. (Capacidad limitada de conclusión) Para cada trabajo, existe un valor máximo de probabilidad para completarlo.

Formalmente, para el trabajo i, sea $S_i = \{0, 1, 2, \ldots, L_i\} \bigcup \{C\}$ donde $L_i = M_i$ o $L_i = \infty$ y C es un estado absorbente asociado con la finalización del trabajo correspondiente.

Además considere que:
$\mathbb{P}[\{C\}|x_i = C] = 1$, $\mathbb{P}[\{C\}|x_i = t_i]$, $\mathbb{P}[\{t_i + 1\}|x_i = t_i] = 1 - P_i(t_i)$, $\tilde{r}_i(C) = \Delta_1$, $\tilde{r}_i(ti) = P_i(t_i)V_i\Delta + \Delta_1$, $t_i \geq 1$.

Si además, bajo el paradigma de la "paciencia finita", se obtiene que $\mathbb{P}[\{M_i\}|x_i = M_i] = 1$ y $\tilde{r}_i(M_i) = \Delta_1$.

3.3.1. Caso con deterioro

En algunos casos, se puede considerar que mientras más intentos se realicen para llevar a cabo un trabajo, la probabilidad de que se termine disminuye: $\forall i$ and $\forall t : P_i(t-1) \geq P_i(t)$.

En este caso, para cada trabajo no completado (y que aún continúa en el caso de "paciencia finita":$t < M_i$) el índice de Gittins difuso coincide con la recompensa inmediata, ya que en el contexto tanto clásico como difuso, el tiempo óptimo de detención para el trabajo i es $\tau_i = 1$.

A continuación, tenemos el índice de Gittins difuso del trabajo i en el tiempo de procesamiento t_i, denotado por denotado por ${}^i\tilde{G}(t_i)$.

$$ {}^i\tilde{G}(t_i) = {}^iG(t_i)\Delta + \Delta_1 = P_i(t_i)V_i\Delta + \Delta_1 $$

y ${}^i\tilde{G}(C) = \Delta_1$, donde $t_j \neq C$ (y $t_j \neq M_i$ en el marco del "entorno de paciencia finita"). Se puede ver que el índice de Gittins nítido es $P_i(t_i)V_i$.

Por lo tanto, la estrategia de Gittins difusa es: seleccionar un trabajo no completado con el mayor valor, con respecto a ambos órdenes considerados de: $P_i(t_i)V_i\Delta + \Delta_1$, de manera que no se haya alcanzado el tiempo límite (en caso de que aplique). O de manera equivalente, dado $\tilde{0} \leq \Delta$ y $\tilde{0} \leq \Delta_1$, se selecciona un trabajo incompleto de manera que $P_i(t_i)V_i$ sea el mayor.

Observación 3.3.1. El marco de deterioro se vuelve importante en vista del teorema 2D en [26], que establece una equivalencia entre cualquier Proceso de Bandidos Armados (bajo condiciones adecuadas) y con deterioro, en el sentido de que el valor de ambos bandidos es igual y las clases de estrategias óptimas para ambos bandidos coinciden.

3.3.2. Resultados numéricos

En esta sección, se presentará una implementación numérica de la estrategia óptima de selección de trabajos en el contexto difuso con probabilidades no decrecientes de alcanzar el estado C en el caso de "paciencia finita" con $\beta = 0.3$ donde $S_1 = \{0, 1, 2, 3, 4, 5, 6, 7\} \cup \{C\}$ y $S_2 = \{0, 1, 2, 3, 4, 5, 6\} \cup \{C\}$.

Para esto, considere dos procesos de bandidos y dos números difusos Δ y Δ_1. Además, considere las siguientes matrices de transición correspondientes a los trabajos 1 y 2, respectivamente:

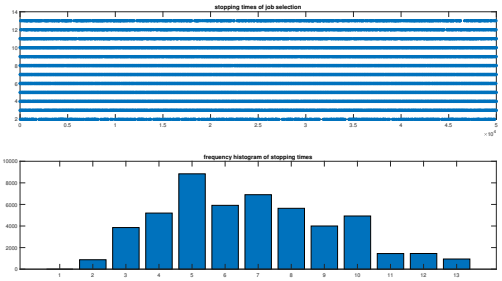

Figura 3.1: Descripción de los tiempos de paro.

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.35 & 0 & 0 & 0.65 & 0 & 0 & 0 & 0 & 0 \\ 0.15 & 0 & 0 & 0 & 0.85 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0.05 & 0 & 0 & 0 & 0 & 0 & 0.95 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0.99 & 0 \\ 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

y

$$P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.19 & 0 & 0.81 & 0 & 0 & 0 & 0 & 0 \\ 0.22 & 0 & 0 & 0.78 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0.35 & 0 & 0 & 0 & 0.65 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0.99 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Además, las recompensas difusas pueden ser condensadas en los siguientes vectores:

$$\tilde{r}_1 := (\Delta_1, 0.1\Delta + \Delta_1, 0.35\Delta + \Delta_1, 0.15\Delta + \Delta_1, \\ 0.5\Delta + \Delta_1, 0.05\Delta + \Delta_1, 0.01\Delta + \Delta_1, 0.8\Delta + \Delta, \Delta_1)$$

y

$$\tilde{r}_2 := (\Delta_1, 0.19\Delta + \Delta_1, 0.22\Delta + \Delta_1, 0.5\Delta + \Delta_1, \\ 0.35\Delta + \Delta_1, 0.6\Delta + \Delta_1, 0.01\Delta + \Delta_1, \Delta_1).$$

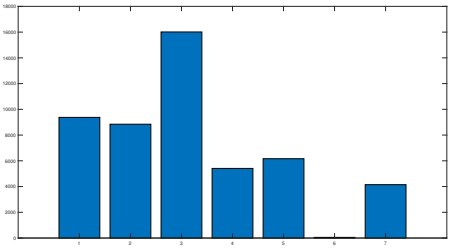

Figura 3.2: Descripción de los tiempos de paro para el trabajo 1 (frecuencia absoluta).

Por lo tanto, al suponer que $\mathbf{x}(0) = (0, 0)$, se obtiene que los índices de Gittins difusos correspondientes son

$$\tilde{\mathbb{G}}_1 := (\Delta_1, 0.1565\Delta + \Delta_1, 0.35\Delta + \Delta_1, 0.2211\Delta + \Delta_1, \\ 0.5\Delta + \Delta_1, 0.088\Delta + \Delta_1, 0.191\Delta + \Delta_1, 0.8\Delta + \Delta_1, \Delta_1)$$

y

$$\tilde{\mathbb{G}}_2 := (\Delta_1, 0.2106\Delta + \Delta_1, 0.277\Delta + \Delta_1, 0.5\Delta + \Delta_1, \\ .3908\Delta + \Delta_1, 0.6\Delta + \Delta_1, 0.01\Delta + \Delta_1, \Delta_1).$$

Se consideran $50,000$ repeticiones del proceso de selección de trabajos bajo las condiciones anteriores. En la Figura 3.1 se puede observar un diagrama de puntos en el que los valores centrales tienen líneas más densas, una situación que se corrobora en el histograma a continuación; además, el valor promedio de los tiempos de paro observados asociados con el proceso global es de 6.763.

Ahora presentamos un análisis de la información relacionada con el tiempo de paro de cada trabajo por separado.

Para el trabajo 1, en la Figura 3.2 se muestra un histograma donde se presenta la frecuencia absoluta del tiempo de

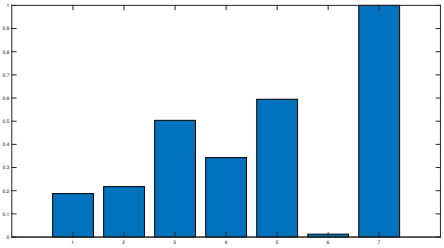

Figura 3.3: Descripción de los tiempos de paro para el trabajo 1 (frecuencia relativa).

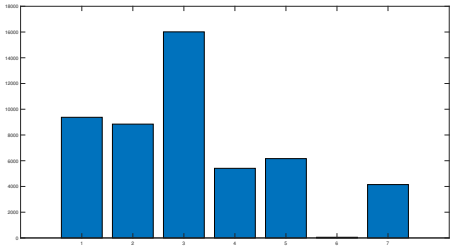

Figura 3.4: Descripción de los tiempos de paro para el trabajo 2 (frecuencia absoluta).

proceso en el que dicho trabajo se detiene. Se observa un pico en el valor "2", ya que es altamente probable que el trabajo 1 no se termine en la primera selección y existe una probabilidad de 0.35 de que el trabajo 1 se termine en su segunda elección. Además, la Figura 3.2 ilustra el histograma correspondiente de frecuencia relativa asociada a las frecuencias absolutas de la Figura 3.2, donde se puede interpretar la altura de cada barra como una aproximación a la probabilidad de finalizar el trabajo 1 dado que no ha sido completado anteriormente (la primera columna de la matriz de transición P_1) Por otro lado, se presenta una descripción gráfica similar para el trabajo 2,

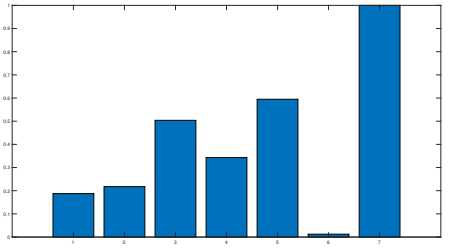

Figura 3.5: Descripción de los tiempos de paro para el trabajo 2 (frecuencia absoluta).

que se encuentra en las Figuras 3.4 y 3.5.

La Figura 3.6 ilustra una trayectoria de la evolución de los estados de ambos trabajos durante el procedimiento de selección, donde el estado asociado con la finalización de cualquier trabajo está etiquetado con “1”.

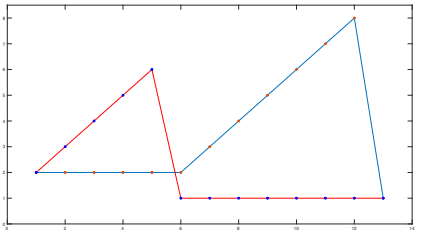

Figura 3.6: Evolución vs estados.

omo complemento de la Figura 3.6, la Figura 3.7 ilustra el desarrollo de la selección de trabajos a través de una trayectoria de muestra.

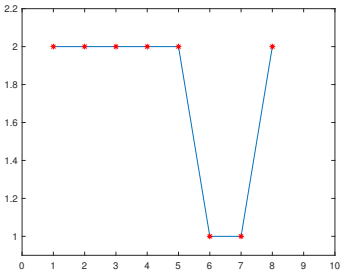

Figura 3.7: Evolución de estados.

Conclusiones

Zadeh nos da las bases en [42] del estudio de la teoría difusa, que ha sido de gran interés en diferentes áreas debido a su utilidad para trabajar con problemas que involucran incertidumbre. Desde entonces nos hemos enfocado en su desarrollo y aplicación en diferentes campos de la matemática.

Por ejemplo en la probabilidad difusa, conceptos como variables aleatorias, probabilidad condicional, esperanza, se han extendido al contexto difuso, en particular en [14], [20], [28], [27], [29] y [36] encontramos procesos de Decisión de Markov difusos con diferentes características. Esto nos ha permitido ampliar el repertorio de herramientas disponibles para dar solución a problemas difusos a través de problemas nítidos.

La importancia de este trabajo radica en el estudio de Bandidos Armados, si bien el modelo de Bandidos Armados es un caso particular de los Procesos de Decisión de Markov, no existen trabajos previos que nos den resultados con una relación entre la solución del modelo nítido y difuso..

Nosotros aprovechamos los resultados dados para los Procesos de Decisión de Markov para extenderlo a los modelos de Bandidos Armados difusos, lo que nos permite utilizar como herramienta el índice de Gittins en este entorno y definir también su versión difusa. Es decir, logramos encontrar la solución óptima de un modelo de Bandidos Armados difuso mediante su versión nítida, lo que simplifica el proceso.

Como trabajo a futuro se buscarán problemas susceptibles de ser modelados mediante Bandidos Armados difusos. También se puede explorar otros órdenes difusos y la difuminación en otros elementos del modelo de Bandidos Armandos.

Bibliografía

[1] Aliprantis C.D. y Border K.C., Infinite Dimensional Analysis A Hitchhiker's Guide. Springer, 3a edición, 2006.

[2] Bellman R. A Problem in the Sequential Design of Experiments. The Indian Journal of Statistics (1933-1960), Vol. 16, No. 3/4, 221-229, 1956.

[3] Bellman and Zadeh, Decision.making in a fuzzy environment, Management Sciences, vol. 17, no. 4, 1970.

[4] Bellman, R. The Theory of dynamic programming. Bull. Amer. Math Soc. 60(1954), no. 6, 503-515.

[5] Berry D.A., Bandit Problems with random discounting. Mathematical learning. Models-Theory and Algorithms. Springer Verlag (1983).

[6] Berry D.A. and Fristedt B., Bandit Problems, Chapman and Hall, 1985.

[7] Bertsekas D.P., Dynamic programming and optimal control. Athena Scientific. 2012.

[8] Omar Besbes, Yonatan Gur, Assaf Zeevi (2019) Optimal Exploration- Exploration in a Multi-armed Bandit Problem with Non-stationary Rewards. Stochastic Systems 9(4):319-337.

[9] Blanco-Mesa, Merigó and Gil-Lafuente, Fuzzy decision-making: Abibliometric-based review, Juumla of Intelligent Fuzzy Systems, vol. 32, no. 3, 2017.

[10] Boucherie, Richardus J. (Editor); Van Dijk, N.M. (Editor)./ Markov Decision Processes in Practice. Cham:Springer, 2017. (International Series in Operations Research and Management Science).

[11] Boneffouf D., Rish I. and Aggarwal C., Survey on Applications of Multi-Armed and Contextual Bandits, 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, United Kingdmon, 2020, pp,1-8, doi:10.1109/CEC48606.2020.9185782.

[12] Briones-García C. R., Montes-de-Oca R., Vázquez-Guevara V.H. y Cruz-Suárez H., Fuzzy Rewards on the Multi-Armed Bandits Model. ICORES, 2024, por publicar.

[13] Carpin, S., Chow, Y.-L., and Pavone, M., Risk aversion in finite markov decision processes using total cost criteria and average value at risk. In 2016 ieee international conference on robotics and automation (icra), pages 335–342. IEEE.

[14] Carrero-Vera, K., Cruz-Suarez, H. y Montes-de Oca, R., (2020). Finite-horizon and infinite-horizon Markov decision processes with trapezoidal fuzzy discounted rewards. In International Conference on Operations Research and Enterprise Systems, pages 171–192. Springer.

[15] Carrero-Vera K., Cruz-Suárez H. and Montes-de-Oca R., Markov decision processes on finite spaces with fuzzy total rewards. Kybernetika (Prague), 58(2),180-199,2022.

[16] Chakraborty, A.; Mondal, S.P.; Ahmadian,; Senu, N.;De, D.; Salahshour, S. The Pentagonal Fuzzy Number: Its Dif-

ferent Representations, Properties, Ranking, Defuzzification and Application in Game Problems. Symmetry, 2019, 11, 248.

[17] Chakravorty, J. and Mahajan, A., Multi-Armed Bandits, Gittins Index, and its Calculation. Methods and Applications of Statistics in Clinical Trials, 2014, vol.2, 416-435.

[18] Cruz-Suárez H., Ilhuicatzi-Roldán R. and Montes-de-Oca R., Markov Decision Processes on Borel Spaces with Total Cost and Random Horizon, Journal of Optimization Theory and Applications 161(1), 329-346, 2013.

[19] Cruz- Suárez H., Montes-de-Oca R. and Ortega-Gutiérrez R.I., An extended version of average Markov decision processes on discrete spaces under fuzzy environment. Kybernetika (Prague), 59(1), 160-178,2023.

[20] Cruz- Suárez H., Montes-de-Oca R. and Ortega-Gutiérrez R.I., Deterministic Discounted Markov Decision Processes with Fuzzy Rewards/Costs. Fuzzy Information and Engineering, 2023, 15(3), 274-290.

[21] Furukawa, N., PArametric orders on fuzzy numbers and their roles in fuzzy optimatization problems. Optimization, 40(2): 171-192, 1997.

[22] Frostig, E. and Weiss, G. Four proofs of Gittins' multiarmed bandit theorem. Ann Oper Res 241, 127–165 (2016).

[23] Gittins, J. and Jones, D., A dynamic allocation index for the sequential desing of experiments. Progress in statistics (edited by j.gani), 241-266, 1974.

[24] Gittins, J. C. (1979) Bandit processes and dynamic allocation indices. Sankhy$\bar{a}$: The Indian Journal of Statistics, 16(3-4), 221-229.

[25] Hernández-Lerma, O. and LAsserre, J.B., Discrete-time Markov control processes. Springer-Verlag, 1996.

[26] Kaspi, H. y Mandelbaum, A. (1998). Multi-armed bandits in discrete and continuous time. The Annals of Applied Probability, 8(4):1270–1290.

[27] Kurano, M., Song, J., Hosaka, M. y Huang, Y. (1998). Controlled Markov set-chains with discounting. Journal of applied probability, 35(2):293–302.

[28] Kurano, M., Yasuda, M., Nakagami, J.-i. y Yoshida, Y. (1996). Markov-type fuzzy decision processes with a discounted reward on a closed interval. European Journal of Operational Research, 92(3):649–662.

[29] Kurano, M., Yasuda, M., Nakagami, J.-i. y Yoshida, Y. (2003). Markov decision processes with fuzzy rewards. Journal of Nonlinear and Convex Analysis, 4(1):105–116.

[30] Martínez-Cortes, V. M. (2021). Bi-personal stochastic transient markov games with stopping times and total reward criterion. Kybernetika, 57(1):1–14.

[31] Montes-de-Oca, R., Saavedra, P., Zacarías-Espinoza, G., Cruz-Suárez, D., Markov Decision Processes Applied to the Payment of Dividends of a Reserve Process. In International Conference on Operations Research and Enterprise System Springer, Cham. (pp. 84-104), 2017.

[32] Puterman, M.L., Markov decision processes discrete stochastic dynamic programming. John Wiley & Sons. 2014.

[33] Ramík J. and Rimánek J., Inequality relation between fuzzy numbers and its use in fuzzy optimization. Fuzzy Sets and Systems, 16, 123-138, 1985.

[34] Robbins H., Some Aspects of the Sequential Desing of Experiments.Bulletin of the American Mathematical Society, 1952, 58(5), 527-535.

[35] Rudin, W., Principles of Mathematical Analysis. McGraw-Hill, 3rd edition, 1976.

[36] Semmouri, A., Jourhmane, M., and Belhallaj, Z. (2020). Discounted Markov decision processes with fuzzy costs. Annals of Operations Research, 295:769–786.

[37] Slivkins A., A Survey on Multi-Armed Bandits. arXiv preprint arXiv:1904.07228.

[38] Sutton, R. S. y Barto, A. G.(2018) Reinforcement learning: An Introduction. MIT Press.

[39] Sutton, R. S. y Barto, A. G.(2018) Reinforcement learning: An Introduction. MIT Press. Segunda edición.

[40] Thompson W. R. On the likelihood one unknown probability exceeds another in view of the evidence of two samples. Biometrika, vol. 25, no 3-4, 285-295, 1933.

[41] V. H. Vázquez-Guevara, H. Cruz-Suárez and F. Velasco-Luna, Optimal assignment of Sellers in a store with a random number of clients via the Armed Bandid model, RAIRO-Oper. Res. 51 (2017) 1119-1132.

[42] Zadeh L., Fuzzy Sets, Information and Control, 8, pp. 338-353, 1965.