




Fuzzy Rewards on the Multi-Armed Bandits Model

Ciria R. Briones-García¹, Raúl Montes-de-Oca²^a, Víctor H. Vázquez-Guevara¹^b
and Hugo Cruz-Suárez¹^c

¹*Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla,
San Claudio y 18 sur San Manuel 72570, Puebla, Mexico*

²*Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Av. Ferrocarril San Rafael Atlixco 186,
Col. Leyes de Reforma 1 A Sección, Alcaldía Iztapalapa 09310, Ciudad de México, Mexico
ciria.briones@alumnos.buap.mx, momr@xanum.uam.mx, {vvazquez, hcs}@fcfm.buap.mx*

Keywords: Armed Bandits Model, Gittins Index, Fuzzy Reward, Trapezoidal Fuzzy Number.

Abstract: In this paper an extension of the Armed Bandits problem is considered under the possibility that reward functions take trapezoidal fuzzy values as the results of a fuzzy affine transformation (which is susceptible of being interpreted as receiving “approximately” a reward located in some interval instead of such reward itself). The main objective is to find an optimal selection strategy that maximizes the fuzzy total expected discounted reward with respect to the partial order based on α -cuts and the one provided by the average ranking. For this, it is obtained that Gittins strategy (that is optimal in the crisp setting) is still optimal at the fuzzy paradigm. In addition, it is found that optimal stopping time associated with crisp Gittins index is the same for its fuzzy counterpart by finding a link between the fuzzy and crisp versions of Gittins index which leads us to demonstrate that fuzzy value function is connected to its crisp analog via some fuzzy affine transformation, with this in mind, it is possible to ensure that value function is approximately in certain interval related to the fuzzy transformation.

1 INTRODUCTION

In this paper it will be taken into account a Simple Family of Armed Bandits with expected discounted total reward criteria in which a trapezoidal fuzzy transformation on reward will be discussed. This topic will lead us to a transition between the theory of Multi-Armed Bandits and the theory of fuzzy numbers.


On the one hand, the first Multi-Armed Bandits problem was discussed in (Thompson, 1933), in which two treatments of unknown efficiency are considered with the goal of adaptively assigning as many patients as possible to the treatment with the greatest success rate. Notably, in (Gittins and Jones, 1974) and (Gittins, 2018) it was developed the so-called Dynamic Allocation index, later renamed as the Gittins index, which is a scalar quantity associated with each process in the Multi-Armed Bandits setting and concludes us to the Gittins optimal strategy: at each stage, the controller must select the process with the highest


Gittins index.


On the other hand, the Fuzzy theory was first introduced by Zadeh in his pioneer work “Fuzzy sets” in 1965 (Zadeh, 1965) where the seminal definitions and operations are stated.

In the best of our knowledge there are not previous works on the Multi-Armed Bandits in a fuzzy environment. However, the Multi-Armed Bandits treated here are related to the fuzzy Markov decision processes (MDPs), specially, with the fuzzy discounted MDPs. Now, concerning the antecedents of fuzzy MDPs, firstly, observe that in (Carrero-Vera et al., 2022) and (Cruz-Suárez et al., 2023b) fuzzy MDPs on discrete spaces and with objective functions other than the total discounted cost are presented. Secondly, in (Carrero-Vera et al., 2020), (Cruz-Suárez et al., 2023a), (Kurano et al., 1996), (Kurano et al., 2003), (Kurano et al., 1998), and (Semmour et al., 2020), discounted MDPs with different fuzzy characteristics have been provided, but, as it was already said, in none of these the specific theme of the Multi-Armed Bandits has been developed.

The main contribution of this paper is to demonstrate that Gittins strategy is also optimal at the fuzzy

^a <https://orcid.org/0000-0002-7632-9190>

^b <https://orcid.org/0000-0001-6602-8733>

^c <https://orcid.org/0000-0002-0732-4943>

framework by finding beforehand a relation between fuzzy and crisp Gittins indexes.

The rest of the paper is organized as follows: Section 2 discusses the preliminary tools of Fuzzy Theory, Section 3 deals with the crisp Armed Bandits Model and the Gittins index, section 4 the Armed Bandits Model with Fuzzy rewards is considered and it contains the main results of this work. Finally, section 5 is devoted to an application setting of the theory presented in section 4.

Notation 1.1. In this article, it will be necessary to establish a difference between crisp and fuzzy operations hence, the standard mathematical symbols will be marked with an asterisk (*) in the fuzzy context. Moreover, some special functions that appear as fuzzy quantities, say, the reward function, the optimal value function, and so on, will be distinguished with a “tilde”; for instance, the fuzzy reward function will be written as \tilde{r} .

2 PRELIMINARIES ON FUZZY THEORY

In this section, introductory fuzzy theory will be displayed, for a general discussion see (Diamond and Kloeden, 1994) and (Zadeh, 1965).

Let Λ be a non-empty set, which denotes the universal set of the discourse. A fuzzy set Γ on Λ is defined in terms of its membership function m_Γ , which assigns to each element of Λ a real value on $[0, 1]$, i.e. $\Gamma = \{(g, m_\Gamma(g)) : g \in \Lambda\}$. The α -cut of Γ , denoted by Γ_α , is defined as the set $\Gamma_\alpha = \{x \in \Lambda \mid m_\Gamma(x) \geq \alpha\}$ for $0 < \alpha \leq 1$, and Γ_0 is the closure of $\{x \in \Lambda \mid m_\Gamma(x) > 0\}$ denoted by $cl\{x \in \Lambda \mid m_\Gamma(x) > 0\}$. In the sequel, it will be assumed that $\Lambda = \mathbb{R}$.

Definition 2.1. A fuzzy number Γ is a fuzzy set defined on the set of real numbers \mathbb{R} , which satisfies:

- a) m_Γ is normal, i.e. there exists $x_0 \in \mathbb{R}$ with $m_\Gamma(x_0) = 1$;
- b) m_Γ is convex, i.e. Γ_α is convex for all $\alpha \in [0, 1]$;
- c) m_Γ is upper-semicontinuous;
- d) Γ_0 is compact.

The set of fuzzy numbers will be denoted by $\mathfrak{F}(\mathbb{R})$.

The manuscript will focus its attention on trapezoidal fuzzy numbers, which are typically considered when the degree of membership for particular values is known to be asymmetric. This class of fuzzy numbers has been successfully applied; for example, to transportation problems (Raj et al., 2023) and portfolio selection (Pahade and Jha, 2021).

Definition 2.2. A fuzzy number Γ is called a trapezoidal fuzzy number if its membership function has the following form:

$$m_\Gamma(x) = \frac{x-l}{m-l}I_{(l,m]}(x) + I_{(m,n]}(x) + \frac{p-x}{p-n}I_{(n,p]}(x)$$

where l, m, n and p are real numbers, with $l < m \leq n < p$ and I_A is the indicator function of A . Hence, a trapezoidal fuzzy number will be represented by (l, m, n, p) .

Remark 2.3. For a trapezoidal fuzzy number $\Gamma = (l, m, n, p)$ its corresponding α -cuts are given by

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

$\alpha \in [0, 1]$ (Rezvani and Molani, 2014).

It may be demonstrated that the following operations between trapezoidal fuzzy numbers hold (Rezvani and Molani, 2014):

Lemma 2.4. If $H = (a_l, a_m, a_n, a_p)$ and $I = (b_l, b_m, b_n, b_p)$ are trapezoidal fuzzy numbers and letting λ be a positive number, it follows that

- a) $\lambda H = (\lambda a_l, \lambda a_m, \lambda a_n, \lambda a_p)$, and
- b) $H +^* I = (a_l + b_l, a_m + b_m, a_n + b_n, a_p + b_p)$.

Let \mathbb{D} denote the set of all closed bounded intervals on the real line \mathbb{R} . For $\Psi = [a_l, a_u], \Phi = [b_l, b_u] \in \mathbb{D}$ define

$$d(\Psi, \Phi) = \max(|a_l - b_l|, |a_u - b_u|).$$

It is possible to verify that (\mathbb{D}, d) is a complete metric space (see (Puri and Ralescu, 1986)). Now, if $\tilde{\eta} \in \mathfrak{F}(\mathbb{R})$, then, as its membership function satisfies b), c) and d) of Definition 2.1, it follows that $\eta_\alpha \in \mathbb{D}$. Therefore, it is defined $\hat{d} : \mathfrak{F}(\mathbb{R}) \times \mathfrak{F}(\mathbb{R}) \rightarrow \mathbb{R}$ by

$$\hat{d}(\tilde{\eta}, \tilde{\mu}) = \sup_{\alpha \in [0,1]} d(\eta_\alpha, \mu_\alpha),$$

for $\tilde{\eta}, \tilde{\mu} \in \mathfrak{F}(\mathbb{R})$.

Now, for being able of establish comparison between trapezoidal fuzzy numbers, let us consider $\tilde{\eta}, \tilde{\mu} \in \mathfrak{F}(\mathbb{R})$, with α -cuts $\tilde{\eta}_\alpha = [a_\alpha, b_\alpha]$ and $\tilde{\mu}_\alpha = [c_\alpha, d_\alpha]$, $\alpha \in [0, 1]$, respectively and define the partial order on $\mathfrak{F}(\mathbb{R})$ “(\leq^*)” by

$$\tilde{\eta} \leq^* \tilde{\mu} \text{ if and only if } a_\alpha \leq c_\alpha \text{ and } b_\alpha \leq d_\alpha, \tag{1}$$

for all $\alpha \in [0, 1]$ (Furukawa, 1997).

It is also possible to compare Fuzzy numbers via the so-called “average ranking”, given for the fuzzy number $\tilde{\eta} = (a_l, a_m, a_n, a_p)$ by

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

Hence, it will be said that trapezoidal fuzzy numbers $\tilde{\eta} = (a_l, a_m, a_n, a_p)$ and $\tilde{\mu} = (a_l^1, a_m^1, a_n^1, a_p^1)$ are such that $\tilde{\eta} \leq^{**} \tilde{\mu}$ if and only if $R(\tilde{\eta}) \leq R(\tilde{\mu})$.

For providing a formal treatment of random entities on fuzzy environments, some concepts on random fuzzy variables are presented.

Definition 2.5. Let (Ω, \mathcal{F}) be a measurable space and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the measurable space of the real numbers. A fuzzy random variable is a function $\tilde{X} : \Omega \rightarrow \mathfrak{F}(\mathbb{R})$ such that $Gr(\tilde{Y}_\alpha) := \{(\omega, u) \in \Omega \times \mathbb{R} : u \in (\tilde{Y}(\omega))_\alpha\} \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R})$, for all $\alpha \in [0, 1]$.

Definition 2.6. Given a probability space (Ω, \mathcal{A}, P) , a fuzzy random variable \tilde{Y} associated to (Ω, \mathcal{A}) is said to be an integrably bounded fuzzy random variable with respect to (Ω, \mathcal{A}, P) if there is a function $h : \Omega \rightarrow \mathbb{R}, h \in L^1(\Omega, \mathcal{A}, P)$ such that $|u| \leq h(\omega)$, for all $(\omega, u) \in \Omega \times \mathbb{R}$ with $u \in (\tilde{Y}(\omega))_0 := \tilde{Y}_0(\omega)$.

Definition 2.7. Given an integrably bounded fuzzy random variable \tilde{Y} associated with respect to the probability space (Ω, \mathcal{A}, P) , then the fuzzy expected value of \tilde{Y} in Aumann's sense is the unique fuzzy set of $\mathbb{R}, E^*[\tilde{Y}]$ such that for each $\alpha \in [0, 1]$:

$$\left(E^*[\tilde{Y}]\right)_\alpha = \left\{ \int_\Omega f(\omega) dP(\omega) \mid f : \Omega \rightarrow \mathbb{R}, \right. \\ \left. f \in L^1(P), f(\omega) \in (\tilde{Y}(\omega))_\alpha \text{ a.s. } [P] \right\}.$$

3 FUZZY GITTINS INDEX

A Simple Family of Armed Bandits Processes is a particular class of discrete-time Markov control process. In this class of process, the controller (player) must play one among $K \in \mathbb{Z}^+$ available Armed Bandits processes. In this way, at each time $t \in \{0, 1, \dots\}$; for each Bandit Process, the controller has two possible actions: either play it or not. In case of playing the m -th Armed Bandit Process, the controller receives a reward (from m -th process) and only the m -th system moves to another state according to a Markovian dynamic. The objective is to determine an activation policy so as to maximize the total expected reward of the overall selecting process. Formally:

Let $\{\mathbf{X}(t) = (X_1(t), \dots, X_K(t)) : t = 0, 1, \dots\}$ the state of a Simple Family of Armed Bandits Processes, where for each $t \in \{0, 1, \dots\}$, $X_i(t)$ is a discrete random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values on a finite-nonempty set \mathbb{X} called the state space. Let $\mathbf{a}(t) = (a_1(t), a_2(t), \dots, a_K(t))$ the action selected by the controller at time t . For each $t \in \{0, 1, \dots\}$, $a_i(t)$ is a binary function, which takes the value 1, if the player chooses the i -th Armed Bandit process, otherwise

$a_i(t) = 0$. Hence, the action space can be defined for each time t as $\mathbb{A}(t) := \{(a_1(t), a_2(t), \dots, a_K(t)) : \sum_{i=1}^K a_i(t) = 1\}$. Consider for each $i = 1, 2, \dots, K$, $r_i : \mathbb{X} \rightarrow \mathbb{R}$ the corresponding reward function and let $P_i = [p_i(x, y)]$ the markovian transition law of the stochastic processes $\{X_i(t)\}$.

Define $\Pi := \{\mathbf{a}(t) : t = 0, 1, \dots : \mathbf{a}(t) \in \mathbb{A}(t)\}$ the set of all admissible policies (or activation strategies). Then for $\pi \in \Pi$ and $\mathbf{x} \in \mathbb{X}^K$ the expected total discounted value function is defined as follows:

$$V(\pi, \mathbf{x}) := E_{\mathbf{x}}^\pi \left[\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^K r_i(X_i(t)) a_i(t) \right], \quad (2)$$

where $0 < \beta < 1$ is a discount factor. The expectation operator $E_{\mathbf{x}}^\pi$ is associated with the product measure $P_{\mathbf{x}}^\pi$ defined on $(\Omega^K, \mathcal{F}^K)$ induced by the Ionescu-Tulcea theorem (Puterman, 2014). Then the Armed Bandits problem consists in determining an activation strategy $\pi_o \in \Pi$ such that maximizes the total expected discounted reward, i.e.

$$V(\pi_o, \mathbf{x}) = \sup_{\pi \in \Pi} V(\pi, \mathbf{x}),$$

$\mathbf{x} \in \mathbb{X}^K$. The function $V_o(\mathbf{x}) := V(\pi_o, \mathbf{x})$ is called, optimal value

One approach to solve this problem was proposed in (Gittins and Jones, 1974), where for each bandit process one may compute the dynamic allocation index (or simply Gittins index), which depends only on that process, and then at each time the player operates on the bandit process with the highest index. The Gittins index is defined as follows (Gittins, 2018):

$$G(x) = \sup_{\tau > 0} G(x, \tau), x \in \mathbb{X}, \quad (3)$$

where $G(x, \tau) := \frac{\mathbb{E}_x[\sum_{t=0}^{\tau-1} \beta^t r_i(X_i(t))]}{\mathbb{E}_x[\sum_{t=0}^{\tau-1} \beta^t]}$ and τ is a stopping time associated with the stochastic process $\{X_i(t)\}$, i.e. for each $n \in \mathbb{N}$, $[\tau = n] \in \mathcal{F}_n := \sigma(X_i(1), X_i(2), \dots, X_i(n))$.

In the remainder of the manuscript, the subscript i will remove for the sake of simplicity in argumentation. Moreover, the following assumption on the reward function is considered.

Assumption 3.1. (a) $r(\cdot) \geq 0$ and $r(\cdot)$ has finite support.

$$(b) \mathbb{E}_x \left[\sum_{t=0}^{\infty} r(X(t)) \right] < \infty, \text{ for all } x \in \mathbb{X}.$$

Remark 3.2. (a) Assumption 3.1 (b) is satisfied, for instance under a Transient Markov condition (Martínez-Cortés, 2021).

(b) Moreover, observe that Assumption 3.1 (b) implies that

$$0 \leq \sum_{t=0}^{\infty} \beta^t r(X(t)) \leq \sum_{t=0}^{\infty} r(X(t)) < \infty, \quad (4)$$

\mathbb{P}_x almost surely.

4 FUZZY ARMED BANDITS

This section is devoted to a fuzzy version of the Armed Bandits problem introduced in the previous section. Firstly, consider the following assumption.

Assumption 4.1. Let $\Delta = (b, d, f, k)$ and $\Delta_1 = (b_1, d_1, f_1, k_1)$ two fixed trapezoidal fuzzy numbers, with $0 \leq b < d \leq f < k$ and $0 \leq b_1 < d_1 \leq f_1 < k_1$. It is also supposed that

$$\tilde{r}(x) = r(x)\Delta + \Delta_1, \tag{5}$$

with $x \in \mathbb{X}$, where r is a reward function as it was considered in the previous section, and such that Assumption 3.1 holds.

Intuition behind expression (5) is that instead of receiving reward $r(x)$, one will obtain a reward that is approximately within interval $[dr(x) + d_1, fr(x) + f_1]$. In a similar way to (2), it is defined; for $\pi \in \Pi$ and $\mathbf{x} \in \mathbb{X}^K$, the fuzzy total expected discounted reward by:

$$\tilde{V}(\pi, \mathbf{x}) := E_{\mathbf{x}}^* \left[\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^K \tilde{r}_i(X_i(t)) a_i(t) \right]. \tag{6}$$

Then the fuzzy Armed Bandits problem consists in determining an activation strategy $\pi_o \in \Pi$ that maximizes with respect to any considered fuzzy order (based on α -cuts and average ranking), the fuzzy total expected discounted reward, i.e. such that

$$\tilde{V}(\pi_o, \mathbf{x}) = \sup_{\pi \in \Pi} \tilde{V}(\pi, \mathbf{x}), \tag{7}$$

$\mathbf{x} \in \mathbb{X}^K$. The function $\tilde{V}_o(\mathbf{x}) := \tilde{V}(\pi_o, \mathbf{x})$ is called, fuzzy optimal value function.

The objective now is to propose a fuzzy version of the Gittins index in order to characterize the optimal selection policy at the fuzzy frame. For this, consider the following sets: $S_F := \{\omega \in \Omega : \sum_{t=0}^{\infty} r(X(t, \omega)) < \infty\}$ and $S_I := \{\omega \in \Omega : \sum_{t=0}^{\infty} r(X(t, \omega)) = \infty\}$. Then, define the random variable:

$$Y(\omega) := \begin{cases} \sum_{t=0}^{\tau-1} \beta^t r(X(t, \omega)) & \text{if } \omega \in A(r, \tau) \\ 0 & \text{if } \omega \in B(r, \tau). \end{cases} \tag{8}$$

where $A(r, \tau) := [\tau < \infty] \cup ([\tau = \infty] \cap S_F)$ and $B(r, \tau) := [\tau = \infty] \cap S_I$.

Observe that $Y(\omega)$ is a well-defined function due to $\Omega = [\tau < \infty] \cup ([\tau = \infty] \cap S_F) \cup ([\tau = \infty] \cap S_I)$ and $\mathbb{P}([\tau = \infty] \cap S_I) = 0$, see Remark 3.2 (b). In consequence,

$$\sum_{t=0}^{\tau-1} \beta^t r(X(t)) = Y, \mathbb{P}_x - a.s. \tag{9}$$

Now, it will be presented some notation about the rank of the random variable Y given by (8). For this, note that $[\tau < +\infty] = \bigcup_{n=1}^{+\infty} [\tau = n]$. Hence, the characteristics of Y restricted to each of the elements of the so-called partition of sample space is as follows:

(a) On $[\tau = n]$; for $n \geq 1$, image of Y is denoted by $Y[\tau = n] = \{y_1^n, y_2^n, \dots\}$.

(b) In addition, $Y[[\tau = +\infty] \cap S_F] = Y \left[\bigcup_{n=1}^{+\infty} [\tau = n] \right]$ which is a denumerable set.

(c) And, finally $Y[[\tau = +\infty] \cap S_I] = \{0\}$.

In a similar way for the random variable

$$Z := \sum_{t=0}^{\tau-1} \beta^t < \infty, \mathbb{P}_x - a.s. \tag{10}$$

whose rank adopts the notation $Z[\tau = n] = \{z_n\}$, $n = 1, 2, \dots$, and $Z[\tau = +\infty] = \{z\}$.

Previous facts will be utilized during the proof of the following lemma, which is the first step in order to propose a fuzzy version of the Gittins index.

Lemma 4.2. Let

$$\tilde{Y}(\omega) := \begin{cases} \sum_{t=0}^{\tau-1} \beta^t \tilde{r}(X(t, \omega)), & \text{if } \omega \in A(r, \tau) \\ z\Delta_1, & \text{if } \omega \in B(r, \tau). \end{cases} \tag{11}$$

Then \tilde{Y} is a fuzzy random variable.

Proof. Fix $\alpha \in [0, 1]$. Firstly observe that \tilde{Y} defined in (11) can be represented as

$$\tilde{Y} = Y\Delta + z\Delta_1, \tag{12}$$

as a consequence of (8) and (10). Then, the α -cut of \tilde{Y} is given by

$$\tilde{Y}_\alpha = [Yb(\alpha) + Zb_1(\alpha), Yc(\alpha) + Zc_1(\alpha)].$$

Thus the graph of \tilde{Y}_α can be written as follows:

$$\begin{aligned} Gr(\tilde{Y}_\alpha) &= \bigcup_{n=1}^{\infty} \bigcup_{j=1}^{\infty} ([Y = y_j^n] \cap [Z = z_n]) \\ &\times [y_j^n b(\alpha) + z_n b_1(\alpha), y_j^n c(\alpha) + z_n c_1(\alpha)] \\ &\cup (([Y = 0] \cap [Z = z]) \times [z b_1(\alpha), z c_1(\alpha)]). \end{aligned}$$

Consequently, $Gr(\tilde{Y}_\alpha) \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R})$. Hence \tilde{Y} is a fuzzy random variable. □

Lemma 4.3. \tilde{Y} is an integrably bounded fuzzy random variable with respect to $(\Omega, \mathcal{F}, \mathbb{P}_x)$ and

$$\mathbb{E}_x^*[\tilde{Y}] = \mathbb{E}_x[Y]\Delta + \mathbb{E}_x[Z]\Delta_1, x \in \mathbb{X}. \quad (13)$$

Proof. Let $x \in \mathbb{X}$ fixed. To prove that \tilde{Y} is integrably bounded, note that, $(\tilde{Y}(\omega))_0 = [bY(\omega) + Z(\omega)b_1, Y(\omega)k + Z(\omega)k_1], \omega \in \Omega$.

Define $h : \Omega \rightarrow \mathbb{R}$ given by,

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

Then, if $(\omega, u) \in \Omega \times \mathbb{R}$ with $u \in \tilde{Y}_0(\omega)$, it yields that:

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

this implies that $|u| \leq h(\omega)$. Since $\mathbb{E}_x[h] < \infty$, due to (9) and (10), it is concluded from Definition 2.6 that \tilde{Y} is integrably bounded.

Now, the expected value for \tilde{Y} is to be calculated, firstly noted that for $\alpha \in [0, 1]$ the following identities are valid

$$\left(\mathbb{E}_x^*[\tilde{Y}] \right)_\alpha = \left\{ \int_\Omega f d\mathbb{P} \mid f : \Omega \rightarrow \mathbb{R}, f \in L^1, f(\omega) \in [Y(\omega)b(\alpha) + z(\omega)b_1(\alpha), Y(\omega)c(\alpha) + Z(\omega)c_1(\alpha)] \right\}.$$

Then,

$$Y(\omega)b(\alpha) + Z(\omega)b_1(\alpha) \leq f(\omega) \leq Y(\omega)c(\alpha) + Z(\omega)c_1(\alpha),$$

\mathbb{P}_x -almost surely. Hence, by taking expectations, we find that

$$\begin{aligned} \mathbb{E}_x[Y]b(\alpha) + \mathbb{E}_x[Z]b_1(\alpha) &\leq \int f d\mathbb{P}_x \\ &\leq \mathbb{E}_x[Y]c(\alpha) + \mathbb{E}_x[Z]c_1(\alpha). \end{aligned} \quad (14)$$

Let I_α be the α -cut of the fuzzy number $\mathbb{E}_x[Y]\Delta + \mathbb{E}_x[Z]\Delta_1$, i.e.

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

Consequently, it follows from (14) that

$$(\mathbb{E}_x[\tilde{Y}])_\alpha \subseteq I_\alpha. \quad (15)$$

On the other hand, let $y \in I_\alpha$ and $f : \Omega \rightarrow \mathbb{R}$ given by: $f(\omega) = y, \omega \in \Omega$. Now, suppose that $\forall \omega \in \Omega, y \notin [Y(\omega)b(\alpha) + Z(\omega)b_1(\alpha), Y(\omega)c(\alpha) + Z(\omega)c_1(\alpha)]$, then

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

which is a contradiction. Hence, there exists $\bar{\omega} \in \Omega$ such that:

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

however, given that $f(\omega) = y$, for all $\omega \in \Omega$, it implies that $\forall \omega \in \Omega$:

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

In consequence, $y \in (\mathbb{E}_x^*[\tilde{Y}])_\alpha$, i.e.

$$I_\alpha \subseteq (\mathbb{E}_x^*[\tilde{Y}])_\alpha. \quad (16)$$

Then, from (15) and (16), $(\mathbb{E}_x^*[\tilde{Y}])_\alpha = I_\alpha$. Since $x \in \mathbb{X}$ is arbitrary, the above arguments demonstrate the validity of (13). \square

Now, a fuzzy version of the Gittins index discussed in Section 3 will be defined.

Definition 4.4. Let τ be a stopping time associated with the stochastic process $\{X(t)\}$. The fuzzy Gittins index is defined; for all $x \in \mathbb{X}$, as

$$\tilde{\mathbb{G}}(x) = \sup_{\tau > 0}^* \tilde{\mathbb{G}}(x, \tau), \quad (17)$$

$$\text{where } \tilde{\mathbb{G}}(x, \tau) := \frac{\mathbb{E}_x^* \left[\sum_{t=0}^{\tau-1} \beta^t \tilde{r}(X(t)) \right]}{\mathbb{E}_x^* \left[\sum_{t=0}^{\tau-1} \beta^t \right]}.$$

The following theorem relates the crisp and fuzzy versions of the Gittins index. Furthermore, it shows that the optimal stopping time in both paradigms coincides.

Theorem 4.5. The following statements hold

- (a) $\tilde{\mathbb{G}}(x, \tau) = \mathbb{G}(x, \tau)\Delta +^* \Delta_1, x \in \mathbb{X}$ and τ an stopping time with the stochastic process $\{X(t)\}$.
- (b) If τ_o is a stopping time such that $\mathbb{G}(x) = \mathbb{G}(x, \tau_o), x \in \mathbb{X}$ then $\tilde{\mathbb{G}}(x) = \tilde{\mathbb{G}}(x, \tau_o), x \in \mathbb{X}$.

Proof. Let $x \in \mathbb{X}$ and τ a stopping time be fixed. The validity of (a) is a consequence of the following identities:

$$\begin{aligned} \tilde{\mathbb{G}}(x, \tau) &= \frac{\mathbb{E}_x[Y]\Delta +^* \mathbb{E}_x[Z]\Delta_1}{\mathbb{E}_x[Z]} \\ &= \mathbb{G}(x, \tau)\Delta +^* \Delta_1. \end{aligned}$$

To prove (b) observe that

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

Then

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

Therefore, since x and τ are arbitrary, the result follows as a consequence of $\tilde{\mathbb{G}}(x, \tau_o) = \mathbb{G}(x)\Delta +^* \Delta_1$ (see (a) of this theorem) and Definition 4.4.

In a similar way, the corresponding proof by considering the average ranking approach can be provided. To this end, observe that

$$\begin{aligned} R(\tilde{G}(x, \tau)) &= R(G(x, \tau)\Delta + \Delta_1) \\ &= R(\Delta)G(x, \tau) + R(\Delta_1) \\ &\leq R(\Delta)\mathbb{G}(x) + R(\Delta_1). \end{aligned}$$

This concludes the proof of the theorem. □

5 APPLICATION SETTING

In this section, a fuzzy-reward version of a scheduling example due to Gittins (Gittins, 2018) is taken into account. To this end, consider a machine capable of processing one job at a time from a total of $N \geq 1$. At the beginning of each processing period; t , one of such jobs must be selected in order to be executed whether or not the previous one was completed. Associated with job i and process time $t_i \geq 1$ (the number of times that job i has been chosen for execution during the first t units of time), there exists a probability of $P_i(t_i)$ that $t_i + 1$ periods are necessary to complete such job given that more than t_i periods are needed.

In addition, if at process time $t_i \geq 1$ job i has not been completed and has been selected for performing, then the following fuzzy reward will be received

$$P_i(t_i)V_i\Delta + \Delta_1, \tag{18}$$

where $\Delta = (b, d, f, k)$ and $\Delta_1 = (b_1, d_1, f_1, k_1)$ are fuzzy numbers such that $0 \leq b$ and $0 \leq b_1$ with $V_i > 0$ for each i . Once a job has been terminated it will produce a null reward.

At this point, it will be considered that Assumption 3.1 holds for every job. Some situations in which such supposition is satisfied are, for example:

1. (Finite patience) For each job i , there exists a maximum time for its conclusion (M_i).
2. (Constant probability of conclusion) For each job; i , and every process time; t_i we may assume additionally that $P_i(t_i) = P_i$.
3. (Limited ability for conclusion) For every job, there exists a maximum value of probability for completing it.

Formally, for job i , set $S_i = \{0, 1, 2, \dots, L_i\} \cup \{C\}$ where $L_i = M_i$ or $L_i = \infty$ and C is an absorbing state associated with completion of corresponding job. Additionally, consider that $\mathbb{P}[\{C\}|C] = 1$, $\mathbb{P}[\{C\}|t_i] = P_i(t_i)$, $\mathbb{P}[\{t_i + 1\}|t_i] = 1 - P_i(t_i)$, $\tilde{r}_i(C) = \Delta_1$, $\tilde{r}_i(t_i) = P_i(t_i)V_i\Delta + \Delta_1$, $t_i \geq 1$.

If in addition, under the “finite patience” paradigm, it is obtained that $\mathbb{P}[\{M_i\}|M_i] = 1$ and $\tilde{r}_i(M_i) = \Delta_1$.

5.1 The Deteriorating Case

In some situations, it might be appropriate to consider that each job is such that, the more attempts are made to complete it, the less likely it is that the job will be terminated: $\forall i$ and $\forall t : P_i(t-1) \geq P_i(t)$.

In this case, for each uncompleted (and still going in the “finite patience” case: $t < M_i$) job, the fuzzy Gittins index coincides with immediate reward because at the crisp and fuzzy setting optimal stopping time for job i is $\tau_i = 1$:

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

and ${}^i\tilde{G}(C) = \Delta_1$.

Hence, the fuzzy Gittins strategy is: Select an uncompleted job with the greatest value; with respect to both considered orders, of $P_i(t_i)V_i\Delta + \Delta_1$, such that its limit time has not been reached (if applicable). Or, equivalently; given that $\tilde{\theta} \leq \Delta$ and $\tilde{\theta} \leq \Delta_1$, choose an unfinished job such that $P_i(t_i)V_i$ is the biggest.

Remark 5.1. *Deteriorating framework becomes important in view of theorem 2D in (Kaspi and Mandelbaum, 1998) which states an equivalence between any Armed Bandits Process (under suitable conditions) and a deteriorating one in the sense that the value of both bandits equals and the classes of optimal strategies for both bandits coincide.*

6 CONCLUSIONS

In this paper, a fuzzy affine transformation on reward functions associated with a Multi-Armed Bandits Model was considered. Such transformation led us to model the possibility that rewards lie approximately in certain interval. Furthermore, it guided us to formalize the Fuzzy-random structure of its components and to find relationships with their crisp analogues. In particular, it was found that the policy dictated by Gittins strategy as well as the corresponding stopping time are also optimal at the considered fuzzy setting. Further development may be possible by considering a fuzzy rewards scheme with a denumerable number of Bandit Processes with denumerable state spaces. Another way of action may arise by taking into account reward functions with countable support or by finding application settings that satisfy Assumption 3.1.

ACKNOWLEDGMENTS

This work was partially supported by Proyecto CONAHCYT: “Procesos de Decisión de Markov en ambiente difuso”, Ciencia de Frontera 2023, CF-2023-I-1362.

REFERENCES

- Carrero-Vera, K., Cruz-Suárez, H., and 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.
- Carrero-Vera, K., Cruz-Suárez, H., and Montes-de Oca, R. (2022). Markov decision processes on finite spaces with fuzzy total rewards. *Kybernetika (Prague)*, 58(2):180–199.
- Cruz-Suárez, H., Montes de Oca, R., and Ortega Gutiérrez, R. (2023a). Deterministic discounted Markov decision processes with fuzzy rewards/costs. *Fuzzy Information and Engineering*, 15(3):274–290.
- Cruz-Suárez, H., Montes-de Oca, R., and Ortega-Gutiérrez, R. (2023b). An extended version of average Markov decision processes on discrete spaces under fuzzy environment. *Kybernetika (Prague)*, 59(1):160–178.
- Diamond, P. and Kloeden, P. (1994). *Metric Spaces of Fuzzy Sets: Theory and Applications*. WORLD SCIENTIFIC.
- Furukawa, N. (1997). Parametric orders on fuzzy numbers and their roles in fuzzy optimization problems. *Optimization*, 40(2):171–192.
- Gittins, J. and Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics* (edited by J. Gani), 241–266.
- Gittins, J. C. (2018). Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164.
- Kaspi, H. and Mandelbaum, A. (1998). Multi-armed bandits in discrete and continuous time. *The Annals of Applied Probability*, 8(4):1270–1290.
- Kurano, M., Song, J., Hosaka, M., and Huang, Y. (1998). Controlled Markov set-chains with discounting. *Journal of applied probability*, 35(2):293–302.
- Kurano, M., Yasuda, M., Nakagami, J.-i., and 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.
- Kurano, M., Yasuda, M., Nakagami, J.-i., and Yoshida, Y. (2003). Markov decision processes with fuzzy rewards. *Journal of Nonlinear and Convex Analysis*, 4(1):105–116.
- Martínez-Cortés, V. M. (2021). Bi-personal stochastic transient Markov games with stopping times and total reward criterion. *Kybernetika*, 57(1):1–14.
- Pahade, J. K. and Jha, M. (2021). Credibilistic variance and skewness of trapezoidal fuzzy variable and mean–variance–skewness model for portfolio selection. *Results in Applied Mathematics*, 11:100159.
- Puri, M. L. and Ralescu, D. A. (1986). Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, 114(2):409–422.
- Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- Raj, M. E. A., Sivaraman, G., and Vishnukumar, P. (2023). A novel kind of arithmetic operations on trapezoidal fuzzy numbers and its applications to optimize the transportation cost. *International Journal of Fuzzy Systems*, 25(3):1069–1076.
- Rezvani, S. and Molani, M. (2014). Representation of trapezoidal fuzzy numbers with shape function. *Ann. Fuzzy Math. Inform*, 8(1):89–112.
- Semmouri, A., Jourhmane, M., and Belhallaj, Z. (2020). Discounted Markov decision processes with fuzzy costs. *Annals of Operations Research*, 295:769–786.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.