Discounted Markov Decision Processes with Fuzzy Rewards Induced by Non-fuzzy Systems

Karla Carrero-Vera[1], Hugo Cruz-Suárez[1] and Raúl Montes-de-Oca[2]
[1]*Benemérita Universidad Autónoma de Puebla, Av. San Claudio y Río Verde,*
Col. San Manuel, CU, Puebla, Pue. 72570, Mexico
[2]*Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Av. San Rafael Atlixco 186,*

Keywords: Markov Decision Processes, Dynamic Programming, Optimal Policy, Fuzzy Sets, Triangular Fuzzy Numbers.

Abstract: This paper concerns discounted Markov decision processes with a fuzzy reward function triangular in shape. Starting with a usual and non-fuzzy Markov control model (Hernández-Lerma, 1989) with compact action sets and reward R, a control model is induced only substituting R in the usual model for a suitable triangular fuzzy function $\tilde{R}$ which models, in a fuzzy sense, the fact that the reward R is "approximately" received. This way, for this induced model a discounted optimal control problem is considered, taking into account both a finite and an infinite horizons, and fuzzy objective functions. In order to obtain the optimal solution, the partial order on the α-cuts of fuzzy numbers is used, and the optimal solution for fuzzy Markov decision processes is found from the optimal solution of the corresponding usual Markov decision processes. In the end of the paper, several examples are given to illustrate the theory developed: a model of inventory system, and two others more in an economic and financial context.

1 INTRODUCTION

In various applied areas, such as engineering, operations research, economics, finance, and artificial intelligence, among others, the data required to propose a mathematical model present ambiguity, vagueness or approximate characteristics of the problem of interest (see, for instance, (Fakoor et al., 2016), (Efendi et al., 2018)). Under this context, it is possible to find in the literature the approach of fuzzy numbers to incorporate this kind of characteristics or assertions to mathematical models. The basic theory on the subject of fuzzy numbers was proposed by L. Zadeh in his seminal article written in 1965, which is entitled: "Fuzzy Sets" (Zadeh, 1965). Subsequently, various research articles and texts referring to the fuzzy theory can be found in the literature on the subject, moreover, it is possible to locate extensions of the theory in other fields of mathematical sciences, such as control theory, see (Driankov et al., 2013).

In this manuscript, the authors provide a Markov decision process (MDP, in plural MDPs) with a finite state space, compact action sets and fuzzy characteristics in its payoff or reward function. The idea is the following: a crisp Markov control model (MCM) is considered, that is, an MCM of the type that has been analyzed in (Hernández-Lerma, 1989), with reward R as a basis, and a new MCM is induced changing only R for a reward function with fuzzy values. Specifically, the authors assume that the fuzzy reward function is triangular. This way, the fuzzy control problem consists of determining a control policy that maximizes the expected total discounted fuzzy reward, where the maximization is made with respect to the partial order on the α-cuts of fuzzy numbers.

It is important to mention that triangular fuzzy numbers have been extensively studied and applied in fuzzy control (Pedrycz, 1994). Furthermore, the triangular fuzzy numbers could be used to approximate an arbitrary fuzzy number (see (Ban, 2009) and (Zeng and Li, 2007)).

The methodology that is followed in this article to guarantee the existence of optimal policies in the fuzzy problem consists in applying the existence of optimal policies and the validity of dynamic programming for the crisp control problem, as well as certain properties of the fuzzy triangular numbers.

To illustrate the theory developed several examples are given: a model of inventory system, and two more in an economic and financial context.

In a short summary, the main contribution of the

Carrero-Vera, K., Cruz-Suárez, H. and Montes-de-Oca, R.
Discounted Markov Decision Processes with Fuzzy Rewards Induced by Non-fuzzy Systems.
DOI: 10.5220/0010231400490059
In *Proceedings of the 10th International Conference on Operations Research and Enterprise Systems (ICORES 2021)*, pages 49-59
ISBN: 978-989-758-485-5

article is to present an extension of the standard discounted MDPs to discounted MDPs with fuzzy rewards. In a general way, a fuzzy reward considered models the fact that a non-fuzzy reward is “approximately” received (in a fuzzy sense), and it is obtained that the optimal control of the fuzzy MDP coincides with the optimal control of the non-fuzzy one and the optimal value function for the fuzzy MDP is “approximately” (in a fuzzy sense) the optimal value function of the non-fuzzy MDP (see Theorem 4.6 and Remark 4.7, below).

Research works related to the topic developed here are the following: (Kurano et al., 2003) and (Semmouri et al., 2020). In (Kurano et al., 2003) a fuzzy control problem with finite state and action spaces is examined. Under this same context, (Semmouri et al., 2020) presents the problem of maximizing the total expected discounted reward through the use of ranking functions.

The paper is organized as follows. Section 2 presents the basic results about fuzzy numbers (arithmetic, metric, order, among others) as well as the notation used in subsequent sections. The following section presents a sketch of definitions and results regarding the control problem under the criterion of expected total discounted reward for both, finite and infinite horizons. Section 4 presents the main results of the paper. In this section the theory on the fuzzy control problem is studied under the criterion of a total expected discounted reward. Finally, in Sections 5 and 6 examples to illustrate the theory developed are given. One of them refers to an inventory control system considered in a fuzzy environment; this example is taken into account with a finite planning horizon. The other two examples refer to finance and economics issues addressed in (Webb, 2007) for the crisp versions, and then the respective fuzzy versions are given in this document. Both examples contemplate an infinite horizon.

2 BASIC THEORY OF FUZZY NUMBERS

In this section definitions and results about the fuzzy theory are presented. The fuzzy set theory was proposed by Zadeh in 1965 (Zadeh, 1965), an interesting feature of using a fuzzy approach is that it allows the use of linguistic variables such as: low, very, high, advisable, highly risky, etc. The following definition describes the concept of a fuzzy number.

Definition 2.1. *Let Θ be a non-empty set. Then a fuzzy set A on Θ is defined in terms of the* membership function *μ, which assigns to each element of Θ a real value from the interval $[0,1]$. Consequently a fuzzy set A can be expressed as a set of ordered pairs: $\{(x,\mu(x)) : x \in \Theta\}$.*

The value $\mu(x)$ in the previous definition represents the degree to which the element x verifies the characteristic property of a set $A \subset \Theta$. Then, using the membership function, a fuzzy number can be defined as follows.

Definition 2.2. *A* fuzzy number *A is a fuzzy set defined on the real numbers $\mathbb{R}$ characterized by means of a membership function μ, $\mu : \mathbb{R} \longrightarrow [0,1]$,*

$$\mu(x) = \begin{cases} 0, & x \le a \\ l(x), & a < x \le b \\ 1, & b < x \le c \\ r(x), & c < x \le d \\ 0, & d < x, \end{cases} \tag{1}$$

where a, b, c, and d are real numbers, l is a non-decreasing function and r is a non-increasing function. The functions l and r are called the left and right side of fuzzy number A, respectively.

In the manuscript the following class of fuzzy numbers are considered.

Definition 2.3. *A fuzzy number A is called a* triangular fuzzy number *if its membership function has the following form:*

$$\mu(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{\beta-a}, & a \le x \le \beta \\ \dfrac{\gamma-x}{\gamma-\beta}, & \beta \le x \le \gamma \\ 0, & x > \gamma, \end{cases} \tag{2}$$

i.e. making $l(x) = \dfrac{x-a}{\beta-a}$ and $r(x) = \dfrac{\gamma-x}{\gamma-\beta}$ in (1), where a, γ and β are real numbers such that $a < \beta < \gamma$. In the subsequent sections, a triangular fuzzy number is denoted by $\mu = (a,\beta,\gamma)$.

The next example shows a triangular fuzzy number.

Example 2.4. *Figure 1 illustrates a graphical representation of the triangular fuzzy number $A = (1/2, 3, 7)$.*

Definition 2.5. *Let A be a fuzzy number with a membership function μ and let α be a real number of the interval $[0,1]$. Then the* α-cut *of A, denoted by μ_α, is defined to be the set $\{x \in \Theta : \mu(x) \ge \alpha\}$.*

Remark 2.6. *a) Equivalently to Definition 2.2, a fuzzy number is a fuzzy set with a normal membership function, i.e. there exists $x \in \Theta$ such that*

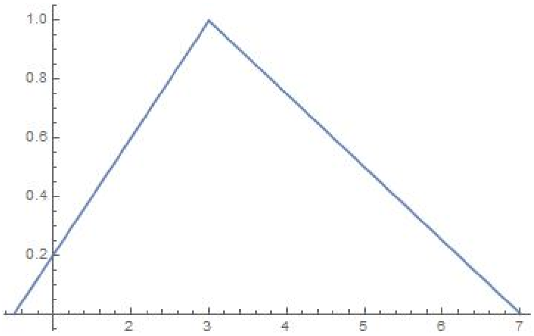


Figure 1: A triangular fuzzy number.

$\mu(x) = 1$ (Klir and Yuan, 1996). Let $\mathfrak{F}(\mathbb{R})$ denote the set of all fuzzy numbers.

b) According to Definition 2.2 and Definition 2.5, for a fuzzy number A, its α-cut set, $\mu_\alpha = [A^-(\alpha), A^+(\alpha)]$ is a closed interval, where $A^-(\alpha) = \inf\{x : \mu_A(x) \geq \alpha\}$ and $A^+(\alpha) = sup\{x : \mu_A(x) \geq \alpha\}$. Consequently, for each $\alpha \in [0,1]$, $(a,\beta,\gamma)_\alpha = [(\beta - a)\alpha + a, \gamma - (\gamma - \beta)\alpha]$ for triangular fuzzy numbers.

Definition 2.7. *Let $\star$ denote any of the four basic arithmetic operations and let A and B be fuzzy numbers. Then, a fuzzy set is defined on $\mathbb{R}$, $A \star B$, by the expression*

$$\mu_{A\star B}(u) = sup_{u=x\star y} \min\{\mu_A(x), \mu_B(y)\}, \tag{3}$$

for all $u \in \mathbb{R}$.

A direct consequence of the previous definition is the following result.

Lemma 2.8. *If $A = (a_l, a_m, a_u)$ and $B = (b_l, b_m, b_u)$ are two triangular fuzzy numbers, then the basic operators for triangular fuzzy numbers are as it follows,*

a) $A \oplus B = (a_l + b_l, a_m + b_m, a_u + b_u)$;

b) $A \ominus B = (a_l - b_u, a_m - b_m, a_u - b_l)$;

c) $A \otimes B = (\min\{a_l b_l, a_l b_u, a_u b_l, a_u b_u\}, a_m b_m, \max\{a_l b_l, a_l b_u, a_u b_l, a_u b_u\})$;

d) $A \oslash B = (\min\{a_l/b_l, a_l/b_u, a_u/b_l, a_u/b_u\}), a_m/b_m, \max\{a_l/b_l, a_l/b_u, a_u/b_l, a_u/b_u\})$.

e) $\lambda A = (\lambda a_l, \lambda a_m, \lambda a_u)$, for each $\lambda \geq 0$.

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

$$d(A,B) = \max(|a_l - b_l|, |a_u - b_u|). \tag{4}$$

It is possible to check that d defines a metric on D and (D,d) is a complete metric space.

Furthermore, for $A, B \in D$ define: $A \lesssim B$ if and only if $a_l \leq b_l$ and $a_u \leq b_u$, where $A = [a_l, a_u]$ and $B = [b_l, b_u]$. Note that "$\lesssim$" is a partial order in D.

Now, define $\hat{d} : \mathfrak{F}(\mathbb{R}) \times \mathfrak{F}(\mathbb{R}) \longrightarrow \mathbb{R}$ by

$$\rho(\mu, \nu) = sup_{\alpha \in [0,1]} d(\mu_\alpha, \nu_\alpha), \tag{5}$$

with $\mu, \nu \in \mathfrak{F}(\mathbb{R})$. It is straightforward to see that ρ is a metric in $\mathfrak{F}(\mathbb{R})$ (Kurano et al., 2003).

Furthermore, for $\mu, \nu \in \mathfrak{F}(\mathbb{R})$ define

$$\mu \preccurlyeq \nu \ if \ and \ only \ if \ \mu_\alpha \lesssim \nu_\alpha \tag{6}$$

with $\alpha \in [0,1]$.

Remark 2.9. *Observe that "$\preccurlyeq$" corresponds to a partial order of $\mathfrak{F}(\mathbb{R})$. A partial order is a reflexive, transitive and antisymmetric binary relation (Aliprantis and Border, 2006). In this case, $(\mathfrak{F}(\mathbb{R}), \preccurlyeq)$ is a partially ordered set or poset. Moreover, if $\tilde{x}$ satisfies that $x \preccurlyeq \tilde{x}$ for each $x \in \mathfrak{F}(\mathbb{R})$, then $\tilde{x}$ is an upper bound for $\mathfrak{F}(\mathbb{R})$. If the set of upper bounds of $\mathfrak{F}(\mathbb{R})$ has a least element, then this element is called the supremum of $\mathfrak{F}(\mathbb{R})$ (Topkis, 1998).*

The proof of the following result can be consulted in (Puri et al., 1993).

Lemma 2.10. *The metric space $(\mathfrak{F}(\mathbb{R}), \rho)$ is complete.*

Definition 2.11. *A sequence $\{l_n\}$ of fuzzy numbers is said to be convergent to the fuzzy number l, written as $\lim_{n \longrightarrow \infty} l_n = l$, if for every $\varepsilon > 0$ there exists a positive integer N such that $\rho(l_n, l) < \varepsilon$ for $n > N$.*

The following result is an extension of Lemma 2.8 and its proof is straightforward.

Lemma 2.12. *For triangular fuzzy numbers the following statements hold:*

a) If $\{(a_l^n, a_m^n, a_u^n) : 1 \leq n \leq N\}$ where N is a positive integer, then

$$\bigoplus_{n=1}^{N} (a_l^n, a_m^n, a_u^n) = (\sum_{n=1}^{N} a_l^n, \sum_{n=1}^{N} a_m^n, \sum_{n=1}^{N} a_u^n).$$

b) If $u_n = \{(a_l^n, a_m^n, a_u^n) : 1 \leq n\}$ and $\sum_{n=1}^{\infty} a_i^n < \infty$, $i \in \{l, m, u\}$, then $S_n := \bigoplus_{m=1}^{n} X_m, n \geq 1$, converges to the triangular fuzzy number $(\sum_{n=1}^{\infty} a_l^n, \sum_{n=1}^{\infty} a_m^n, \sum_{n=1}^{\infty} a_u^n)$.

The next remark provides the Zadeh's extension principle which provides a general method for fuzzification of non-fuzzy mathematical concepts.

Remark 2.13 (Zadeh's Extension Principle)**.** *Let L be a function such that $L : X \longrightarrow Z$ and let A be a fuzzy subset of Θ with a membership function μ.* Zadeh's extension *of L is the function $\hat{L}$ which, applied to A*

gives the fuzzy subset $\hat{L}(A)$ of Z with the membership function given by

$$\hat{\mu}(z)=\begin{cases} sup_{x\in L^{-1}(z)}\mu(x), & L^{-1}(\{z\})\neq\varnothing \\ 0, & L^{-1}(\{z\})=\varnothing. \end{cases} \tag{7}$$

Observe that, if A is a fuzzy subset of Θ, with the membership function μ, and if L is bijective, then the membership function of $\hat{L}(A)$ is given as follows

$$\begin{aligned}\hat{\mu}(z) &= sup_{\{x:L(x)=z\}}\mu(x)\\ &= sup_{\{x\in L^{-1}(z)\}}\mu(x)\\ &= \mu(L^{-1}(z)).\end{aligned}$$

Now, a fuzzy random variable will be defined. In this case, the definition proposed in (Puri et al., 1993) will be adopted.

Definition 2.14. *Let $(\Omega,\mathcal{F})$ 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 $\tilde{X}:\Omega\longrightarrow\mathfrak{F}(\mathbb{R})$ such that for all $(\alpha,B)\in[0,1]\times\mathcal{B}(\mathbb{R})$, $\{\omega\in\Omega:\tilde{X}_\alpha\cap B\neq\varnothing\}\in\mathcal{F}$. Equivalently, $\tilde{X}$ must be viewed as a generalized interval with a membership function μ and α-cut: $X(\omega)_\alpha=[X^-(\omega),X^+(\omega)]$.*

Definition 2.15. *Let $(\Omega,\mathcal{F},P)$ be a probability space and let $\tilde{X}$ be a discrete fuzzy random variable with the range $\{\tilde{s}_1,\tilde{s}_2,...,\tilde{s}_l\}\subseteq\mathfrak{F}(\mathbb{R})$. The mathematical expectation of $\tilde{X}$ is a fuzzy number, $E(\tilde{X})$, such that*

$$E(\tilde{X})=\bigoplus_{i=1}^{l}\tilde{s}_iP(\tilde{X}=\tilde{s}_i). \tag{8}$$

A proof of the following result should be consulted in (Puri et al., 1993).

Lemma 2.16. *Let $\tilde{X}$ and $\tilde{Y}$ be discrete fuzzy random variables with finite range. Then*

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

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

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

3 DISCOUNTED MARKOV DECISION PROCESSES WITH FUZZY REWARD FUNCTIONS

In this section the theory on Markov decision processes necessary for this article is introduced. This kind of processes are used to model dynamic systems in a discrete time. Firstly the optimal control problem with a crisp reward is presented, later the reward function is changed by a fuzzy reward function and the new optimal control problem is given.

3.1 Markov Decision Models

Detailed literature on the theory of Markov decision processes can be consulted in the references: (Hernández-Lerma, 1989) and (Puterman, 1994).

A *Markov decision model* is characterized by the following five-tuple:

$$M:=(X,A,\{A(x):x\in X\},Q,R), \tag{9}$$

where

a) X is a finite set, which is called the *state space*.

b) A is a Borel space, A is denominated the *control* or *action space*.

c) $\{A(x):x\in X\}$ is a family of nonempty subsets $A(x)$ of A, whose elements are the *feasible actions*.

c) Q is the *transition law*, which is a stochastic kernel on X given $\mathbb{K}:=\{(x,a):x\in X,a\in A(x)\}$, $\mathbb{K}$ is denominated the set of feasible state-actions pairs.

d) $R:\mathbb{K}\longrightarrow\mathbb{R}$ is the one-step *reward function*.

Now, given a Markov control Model M, the concept of policy will be introduced. A *policy* is a sequence $\pi=\{\pi_t:t=0,1,...\}$ of stochastic kernels π_t on the control set A given the history $\mathbb{H}_t$ of the process up to time t, where $\mathbb{H}_t:=\mathbb{K}\times\mathbb{H}_{t-1}, t=1,2,...$ and $\mathbb{H}_0=X$. The set of all policies will be denoted by Π. A *deterministic Markov policy* is a sequence $\pi=\{f_t\}$ such that $f_t\in\mathbb{F}$, for $t=0,1,...$, where $\mathbb{F}$ denotes the set of functions $f:X\longrightarrow A$ such that $f(x)\in A(x)$, for all $x\in X$. A Markov policy $\pi=\{f_t\}$ is said to be *stationary* if f_t is independent of t, i.e., $f_t=f\in\mathbb{F}$, for all $t=0,1,...$. In this case, π is denoted by f and $\mathbb{F}$ is denominated the set of *stationary policies*.

Let $(\Omega,\mathcal{F})$ be the measurable space consisting of the canonical sample space $\Omega=\mathbb{H}_\infty:=(X\times A)^\infty$ and $\mathcal{F}$ be the corresponding product σ-algebra. The elements of Ω are sequences of the form $\omega=(x_0,a_0,x_1,a_1,...)$ with $x_t\in X$ and $a_t\in A$ for all $t=0,1,...$. The projections x_t and a_t from Ω to the sets X and A are called state and action variables, respectively.

Let $\pi=\{\pi_t\}$ be an arbitrary policy and μ be an arbitrary probability measure on X called the initial distribution. Then, by the theorem of C. Ionescu-Tulcea (Hernández-Lerma, 1989), there is a unique probability measure P^π_μ on $(\Omega,\mathcal{F})$ which is supported on $\mathbb{H}_\infty$, i.e., $P^\pi_\mu(\mathbb{H}_\infty)=1$. The stochastic process $(\Omega,\mathcal{F},P^\pi_\mu,x_t)$ is called a discrete-time Markov control process or a Markov decision process.

The expectation operator with respect to P^π_μ is denoted by E^π_μ. If μ is concentrated at the initial state $x\in X$, then P^π_μ and E^π_μ are written as P^π_x and E^π_x, respectively.

The transition law of a Markov control process (see 9) is often specified by a difference equation of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \tag{10}$$

$t = 0,1,2,...,$ with $x_0 = x \in X$ known, where $\{\xi_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with values in a finite space S and a common distribution Δ, independent of the initial state x_0. In this case, the transition law Q is given by

$$Q(B|x,a) = E[I_B(F(x,a,\xi))],$$

$B \subseteq X$, $(x,a) \in K$, E is the expectation with respect to distribution Δ, ξ is a generic element of the sequence $\{\xi_t\}$ and $I_B(\cdot)$ denotes the indicator function of the set B.

Definition 3.1. *Let $(X, A, \{A(x) : x \in X\}, Q, R)$ be a Markov model, then the expected total discounted reward is defined as follows:*

$$v(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \beta^t R(x_t, a_t) \right], \tag{11}$$

$\pi \in \Pi$ and $x \in X$, where $\beta \in (0,1)$ is a given discount factor. Furthermore, the T-stage expected total discounted reward, for each $x \in X$ and $\pi \in \Pi$, is defined as follows:

$$v_T(\pi, x) := E_x^\pi \left[\sum_{t=0}^{T-1} \beta^t R(x_t, a_t) \right], \tag{12}$$

where T is a positive integer.

Definition 3.2. *The* optimal value function *is defined as*

$$V(x) := sup_{\pi \in \Pi} V(\pi, x), \tag{13}$$

$x \in X$. Then the optimal control problem *is to find a policy $\pi^* \in \Pi$ such that*

$$v(\pi^*, x) = V(x),$$

$x \in X$, in which case, π^* is said to be the optimal policy. *Similar definitions can be stated analogously for v_T. In this case, V_T denotes the optimal value function.*

Assumption 3.3. *a) For each $x \in X$, $A(x)$ is a compact set on $\mathcal{B}(A)$, where $\mathcal{B}(A)$ is the Borel σ-algebra of space A.*

b) The reward function R is a non negative and bounded function.

c) For every $x, y \in X$, the mappings $a \mapsto R(x,a)$ and $a \mapsto Q(y|x,a)$ are continuous in $a \in A(x)$.

The proof of the following theorem can be consulted in (Hernández-Lerma, 1989).

Theorem 3.4 (Dynamic Programming)**.** *Under Assumption 3.3 the following statements hold:*

a) Define $W_T(x) = 0$ and for each $n = T-1, ..., 1, 0,$ consider

$$W_n(x) = \max_{a \in A(x)} \{R(x,a) + \beta E[W_{n+1}(F(x,a,\xi))]\}. \tag{14}$$

$x \in X$. Then for each $n = 0, 1, ..., T-1$ there exists $f_n \in \mathbb{F}$ such that

$$W_n(x) = R(x, f_n(x)) + \beta E[W_{n+1}(F(x, f_n(x), \xi))],$$

$x \in X$. In this case, $\pi^* = \{f_0, ..., f_{T-1}\}$ is a Markovian optimal policy and $v_T(\pi^*, x) = W_0(x)$, $x \in X$.

b) The optimal value function V, satisfies the following dynamic programming equation:

$$V(x) = \max_{a \in A(x)} \{R(x,a) + \beta E[V(F(x,a,\xi))]\}, \tag{15}$$

$x \in X$.

c) There exists a policy $f^* \in \mathbb{F}$ such that the control $f^*(x) \in A(x)$ attains the maximum in (15), i.e. for all $x \in X$,

$$V(x) = R(x, f^*(x)) + \beta E[V(F(x, f^*(x), \xi))]. \tag{16}$$

d) Define the value iteration functions *as follows:*

$$V_n(x) = \min_{a \in A(x)} \{c(x,a) + \beta E[V_{n-1}(F(x, f^*(x), \xi))]\}, \tag{17}$$

for all $x \in X$ and $n = 1, 2, \ldots$, with $V_0(\cdot) = 0$. Then the value iteration functions converge point-wise to the optimal value function V, i.e.

$$\lim_{n \to \infty} V_n(x) = V(x),$$

$x \in X$.

3.2 Objective Functions

Consider a Markov decision model $M = (X, A, \{A(x) : x \in X\}, Q, \tilde{R})$, where the first four components are the same as in the model given in (9). The fifth component corresponds to a fuzzy reward function on $\mathbb{K}$.

The evolution of a stochastic fuzzy system is as follows: if the system is in the state $x_t = x \in X$ at time t and the control $a_t = a \in A(x)$ is applied, then two things happen:

a) a fuzzy reward $\tilde{R}(x,a)$ is obtained.

b) the system jumps to the next state x_{t+1} according to the transition law Q, i.e.

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

with $B \subseteq X$.

For each policy $\pi \in \Pi$ and state $x \in X$, let

$$\tilde{v}_T(\pi, x) := \bigoplus_{t=0}^{T-1} \beta^t \tilde{E}_x^\pi \left[\tilde{R}(x_t, a_t)\right], \qquad (18)$$

where T is a positive integer and $\tilde{E}_x^\pi$ is the expectation with respect to $\tilde{P}_x^\pi$ and its expectation of a fuzzy random variable is defined by (8). The expression given in (18) is called the T-stage fuzzy reward. Furthermore, the following objective function will be considered:

$$\tilde{v}(\pi, x) := \bigoplus_{t=0}^{\infty} \beta^t \tilde{E}_x^\pi \left[\tilde{R}(x_t, a_t)\right]. \qquad (19)$$

In this way, the control problem of interest is the maximization of the finite/infinite horizon expected total discounted fuzzy reward (see (18) and (19), respectively). In the next section it will be proved that (18) converges to the objective function (19) with respect to the metric ρ (see (5)). The following assumption is considered for the reward function of fuzzy model M.

Assumption 3.5. *Let B,C and D be real numbers, such that $0 < B < C < D$. It will be assumed that the fuzzy reward is a triangular fuzzy number (see Definition 2.3), specifically*

$$\tilde{R}(x, a) = (BR(x, a), CR(x, a), DR(x, a)) \qquad (20)$$

for each $(x, a) \in \mathbb{K}$, where $R : \mathbb{K} \longrightarrow \mathbb{R}$ is the reward function of the model introduced in Section 3.1.

Remark 3.6. *Observe that, under Assumption 3.5 and Lemma 2.12, the T-stage fuzzy reward (18) is a triangular fuzzy number.*

4 OPTIMAL CONTROL PROBLEM WITH FUZZY REWARDS

In this section results will be presented which refer to the convergence of the T-stage fuzzy reward (18) to the infinite horizon expected total discounted fuzzy reward (19). Later, the existence of optimal policies and validity of dynamic programming will be verified.

Lemma 4.1. *For each $\pi \in \Pi$ and $x \in X$,*

$$\lim_{T \longrightarrow \infty} \rho(\tilde{v}_T, \tilde{v}) = 0,$$

where ρ is the Hausdorff metric (see (5)).

Proof. Let $\pi \in \Pi$ and $x \in X$ fixed. To simplify the notation in this proof $v = v(\pi, x)$ and $v_T = v_T(\pi, x)$ will be denoted. Then, according to (18) and (20) the α-cut of the fuzzy reward function, is given by

$$\begin{aligned}\Delta^T :&= (Bv_T, Cv_T, Dv_T)_\alpha \\ &= [B(1-\alpha)v_T + \alpha C v_t, D(1-\alpha)v_T + \alpha C v_T].\end{aligned}$$

Analogously, the α-cut of (19) is given by

$$\begin{aligned}\Delta :&= (Bv, Cv, Dv)_\alpha \\ &= [B(1-\alpha)v + \alpha C v, D(1-\alpha)v + \alpha C v].\end{aligned}$$

Hence, by (5), it is obtained that

$$\rho(\Delta^T, \Delta) = sup_{\alpha \in [0,1]} d(\Delta_\alpha^T, \Delta_\alpha).$$

Now, due to the identity $\max(c, b) = (c + b + |b - c|)/2$ with $b, c \in \mathbb{R}$, it yields that

$$d(\Delta_\alpha^T, \Delta_\alpha) = (1-\alpha)D(v - v_T) + \alpha C(v - v_T).$$

Then,

$$\begin{aligned}\rho(\Delta^T, \Delta) &= sup_{\alpha \in [0,1]}(v - v_T)(D - \alpha(D - C)) \\ &= (v - v_T)D.\end{aligned} \qquad (21)$$

Therefore, when T goes to infinity in (21), it concludes that

$$\begin{aligned}\lim_{T \longrightarrow \infty} \rho(\tilde{v}_T, \tilde{v}) &= \lim_{T \longrightarrow \infty} (v - v_T)D \\ &= 0.\end{aligned}$$

The second equality is a consequence of the dominated convergence theorem (see (11) and (12)). □

Definition 4.2. *The optimal control fuzzy problem consists in determining a policy $\pi^* \in \Pi$ such that*

$$\tilde{v}(\pi, x) \preccurlyeq \tilde{v}(\pi^*, x),$$

for all $\pi \in \Pi$ and $x \in X$. In consequence,

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

for all $x \in X$ (see Remark 2.9). In this case, the optimal fuzzy value function is defined as follows:

$$\tilde{V}(x) = \tilde{v}(\pi^*, x),$$

$x \in X$ and π^* is called the optimal policy of the fuzzy optimal control problem. Similar definitions can be stated for $\tilde{v}_T$, the T-stage fuzzy reward, in this case the optimal fuzzy value is denoted by $\tilde{V}_T$.

A direct consequence of a previous definition and Theorem 3.4 is the next result.

Theorem 4.3. *Under Assumptions 3.3 and 3.5 the following statements hold.*

a) The optimal policy π^* of the crisp finite optimal control problem (see (12)) is the optimal policy for $\tilde{v}_T$, i.e. $\tilde{v}_T(\pi^*, x) = sup_{\pi \in \Pi} \tilde{v}_T(\pi, x)$ for all $\pi \in \Pi$ and $x \in X$.

b) The optimal fuzzy value function is given by

$$\tilde{V}_T(x) = (BV_T(x), CV_T(x), DV_T(x)), \qquad (22)$$

$x \in X$, where $\tilde{V}_T(x) = sup_{\pi \in \Pi} \tilde{v}_T(\pi, x)$, $x \in X$.

Theorem 4.4. *Under Assumptions 3.3 and 3.5 the following statements hold:*

a) *The optimal policy of the fuzzy control problem is the same as the optimal policy of the optimal control problem.*

b) *The optimal fuzzy value function is given by*

$$\tilde{V}(x) = (BV(x), CV(x), DV(x)), x \in X. \qquad (23)$$

Proof. a) Let $\pi \in \Pi$ and $x \in X$ be fixed. First, observe that (19) is equivalent to

$$\tilde{v}(\pi,x) := (Bv(\pi,x), Cv(\pi,x), Dv(\pi,x)),$$

as a consequence of Assumption 3.5. Then, the α-cut of $\tilde{v}(\pi,x)$ is given by

$$\tilde{v}(\pi,x)_\alpha = [Bv(\pi,x) + \alpha(C-B)v(\pi,x), Dv(\pi,x) + \alpha(D-C)v(\pi,x)].$$

Now, by Theorem 3.4, there exists $f^* \in \mathbb{F}$ such that

$$Bv(\pi,x) + \alpha(C-B)v(\pi,x) \leq Bv(f^*(x),x) + \alpha(C-B)v(f^*(x),x),$$
$$Bv(\pi,x) + \alpha(D-C)v(\pi,x) \leq Dv(f^*(x),x) + \alpha(D-C)v(f^*(x),x)$$

and since $x \in X$ and $\pi \in \Pi$ are arbitrary, the result follows, due to Definition 4.2.

b) By Theorem 4.4 a), it follows that

$$\tilde{V}(x) = (Bv(f^*(x),x), Cv(f^*(x),x), Dv(f^*(x),x)),$$

for each $x \in X$, thus applying Theorem 3.4, it is concluded that

$$\tilde{V}(x) = (BV(x), CV(x), DV(x)), x \in X.$$

□

It is important to observe that Assumption 3.5 could be changed for the following one:

Assumption 4.5. *Let B and D be real numbers, such that $0 < B < R(x,a) < D$ for all $x \in X$ and $a \in A(x)$. It will be assumed that the fuzzy reward is a triangular fuzzy number of the type:*

$$\tilde{R}(x,a) = (B, R(x,a), D) \qquad (24)$$

for each $(x,a) \in \mathbb{K}$, where $R : \mathbb{K} \longrightarrow \mathbb{R}$ is the reward function of the model introduced in Section 3.1.

Note that it is possible to prove with similar ideas to the ones given in the proof of Theorem 3.5 the following result:

Theorem 4.6. *Under Assumptions 3.3 and 4.5 the following statements hold:*

a) *The optimal policy of the fuzzy control problem is the same as the optimal policy of the non-fuzzy optimal control problem.*

b) *The optimal fuzzy value function is given by*

$$\tilde{V}(x) = (\frac{B}{1-\beta}, V(x), \frac{D}{1-\beta}), \qquad (25)$$

$x \in X$.

Remark 4.7. *Using Theorem 4.6, the main interpretation of the fuzzy extension presented here is obtained: the original non- fuzzy reward R is substituted by the fuzzy reward $\tilde{R}(\cdot,\cdot) = (B, R(\cdot,\cdot), D)$ which modelled the fact that "approximately" R, measured in a fuzzy sense is gotten, and it is deduced that the optimal control of the fuzzy MDP coincides with the optimal control of the non-fuzzy one and the optimal value function for the fuzzy MDP is "approximately" (in a fuzzy sense) the optimal value function of the non-fuzzy MDP.*

5 A FUZZY INVENTORY CONTROL SYSTEM

In this section, first a classical example of inventory control system (Puterman, 1994) will be presented, later a triangular fuzzy inventory control system will be introduced. The optimal solution of the fuzzy inventory is obtained by an application of Theorem 4.3 and the solution of the crisp inventory system.

The following example is addressed in (Puterman, 1994), below there is a summary of the points of interest to introduce its fuzzy version. Consider the following situation, in a warehouse where every certain period of time the manager carries out an inventory to determine the quantity of product stored. Based on such information, a decision is made whether or not to order a certain amount of additional product from a supplier. The manager's goal is to maximize the profit obtained. The demand for the product is assumed to be random with known probability distribution. The following assumptions will be treated to propose the mathematical model.

Inventory Assumptions.

a) The decision to additional order is made at the beginning of the period and is delivered immediately.

b) Product demands are received throughout the period of time but are fulfilled in the last instant of the time of the period.

c) There are no unfilled orders.

c) Revenues and the distribution of demand do not vary with the period.

d) The product is only sold in whole units.

e) The warehouse has a capacity for M units, where M is a positive integer.

Then, under previous assumption, the state space is given by $X := \{0,1,2,...,M\}$, the action space and admissible action set are given by $A := \{0,1,2,...\}$ and $A(x) := \{0,1,2,...,M-x\}), x \in X$, respectively.

Now, consider the following variables: let x_t denote the inventory at time $t = 0,1,...$, the evolution of the system is modeled by the following dynamic system Lindley kind:

$$x_{t+1} = (x_t + a_t - D_{t+1})^+, \qquad (26)$$

with $x_0 = x \in X$ known, where

a) a_t denotes the control or decision applied in the instant t and it represents the quantity ordered by the inventory manager (or decision maker).

b) The sequence $\{D_t\}$ is conformed by independent and identically distributed non-negative random variables with common distribution $p_j := \mathbb{P}(D = j), j = 0,1,...$, where D_t denotes the demand within the period of time t.

Observe that the difference equation given in (26) induces a stochastic kernel defined on X given $\mathbb{K} := \{(x,a) : x \in X, a \in A(x)\}$, as follows

$$Q(X_{t+1} \in (-\infty, y])|X_t = x, a_t = a) = 1 - \Delta(x+a-y),$$

where Δ is the distribution of D with $x \in X$, $y,a \in \{0,1,...\}$ and $Q(X_{t+1} \in (-\infty,y])|X_t = x, a_t = a) = 0$, if $x \in X, a \in \{0,1,...\}$ and $y < 0$. Then it follows that

$$Q(X_{t+1} = y|x,a) = \begin{cases} 0 & if \quad M \geq y > x+a \\ p_{x+a-y} & if \quad M \geq x+a \geq y > 0 \\ q_{x+a} & if \quad M \geq x+a, y = 0 \end{cases}$$

The step reward function is given by $R(x,a) = E[H(x+a-(x+a-\mathcal{D})^+)]$, $(x,a) \in \mathbb{K}$, where $H : \{0,1,...\} \to \{0,1,...\}$ is the revenue function, which is a known function and $\mathcal{D}$ is a generic element of the sequence $\{D_t\}$. Equivalently, $R(x,a) = F(x+a)$, $(x,a) \in \mathbb{K}$, where

$$F(u) := \sum_{k=0}^{u-1} H(k)p_k + H(u)q_u, \qquad (27)$$

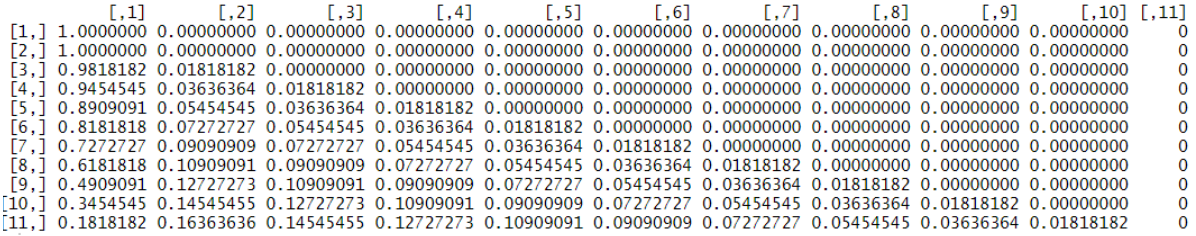
with $q_u := \sum_{k=u}^{\infty} p_k$. The objective in this section is to maximize the total discounted reward with a finite horizon, see (18).

In particular, suppose that the horizon is $T = 50$, the state space $X = \{0,1,...,10\}$, the revenue function $H(u) = 10u$ and the transition law is given in Figure 2. Then, in accordance with Theorem 4.3, which was programmed in the statistical software R using the next algorithm:

```
Algorithm: To calculate the optimal value and optimal policy.
  Input: MDP
  Output: The optimal value vector.
An optimal policy
  Initialize $W_N(x,A) = 0$, $W_N^*(x) = 0$,
    $K_N(x) = W_N^*(x)$.
    $t = N-1$
repeat
    for $x \in S$ do
        $f_x = 0$
        $a(x) = f_x$
        $W(x,a(x)) = R(x,a(x)) +$
                $\beta \sum_{i=0}^{Z} Q(y|x+a(x)) W_{t+1}(y,0)$

        $A(x) = 1,...,M-x$
        for $a \in A(x)$ do
            $W_t(x,a) = R(x,a) +$
                    $\beta \sum_{y=0}^{Z} Q(y|x+a) W_{t+1}(y,0)$

            if $W_t(x,a) \geq W(x,a(x))$ do
                $W(x,a(x)) = W_t(x,a)$
                $f_x = a$

        end for

        $W_t(x) = W_t(x,f_x)$

        $W_t(x,0) = W_t(x)$

        if $W_t(x) \geq K_{t+1}(x)$ do
            $K_t(x) = W_t(x)$

        $W^*(x) = K_t(x)$
    end for

    $t = n-1$
until $t = 0$
```

In consequence, the output of the program is obtained as illustrated in Figure 3. In this matrix the last column represents the optimal policy and the penultimate column the value function, for each state $x \in \{0,1,...,10\}$. The other input of the matrix represents the following:

$$G(x,a) := R(x,a) + \alpha E[W_1(F(x,a,D))],$$



```
         [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]       [,8]       [,9]      [,10] [,11]
 [1,] 1.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [2,] 1.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [3,] 0.9818182 0.01818182 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [4,] 0.9454545 0.03636364 0.01818182 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [5,] 0.8909091 0.05454545 0.03636364 0.01818182 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [6,] 0.8181818 0.07272727 0.05454545 0.03636364 0.01818182 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000     0
 [7,] 0.7272727 0.09090909 0.07272727 0.05454545 0.03636364 0.01818182 0.00000000 0.00000000 0.00000000 0.00000000     0
 [8,] 0.6181818 0.10909091 0.09090909 0.07272727 0.05454545 0.03636364 0.01818182 0.00000000 0.00000000 0.00000000     0
 [9,] 0.4909091 0.12727273 0.10909091 0.09090909 0.07272727 0.05454545 0.03636364 0.01818182 0.00000000 0.00000000     0
[10,] 0.3454545 0.14545455 0.12727273 0.10909091 0.09090909 0.07272727 0.05454545 0.03636364 0.01818182 0.00000000     0
[11,] 0.1818182 0.16363636 0.14545455 0.12727273 0.10909091 0.09090909 0.07272727 0.05454545 0.03636364 0.01818182     0
```

Figure 2: Transition law.

$(x,a) \in \mathbb{K}$.

In conclusion, the optimal value function is $V_T(x) = 693.39$ for each $x \in X$ and the optimal policy is given by $f(x) = M - x$, $x \in X$ with $M = 10$.

Now, considering that in operations research it is often difficult for a manager to control inventory systems, due to the fact that data in each stage of observation is not always is certain, then a fuzziness approach should be applied. In this way, take into account the previous inventory system in a fuzzy environment, that is, the reward function given in Assumption 3.5 will be considered:

$$\tilde{R}(x,a) = (BR(x,a), CR(x,a), DR(x,a)),$$

with $0 < B < C < D$. Then, by Theorem 4.3, it follows that the optimal policy of the fuzzy optimal control problem is given by

$$\tilde{\pi}^* = \{f_0, \dots, f_{T-1}\},$$

where $f_t(x) = M - x$, $x \in X$ and the optimal value function is given by $\tilde{V}_T(x) = (BV_T(x), CV_T(x), DV_T(x))$, $x \in X$.

6 ECONOMIC/FINANCIAL APPLICATIONS

In this section, two examples on applications in Economics and Finance are presented. Firstly the crisp model of both examples and the respective solution is proposed. Later a fuzzy version of these problems is introduced.

6.1 Example 1

Let $X = \{\chi_0, \chi_1\}$, $0 < \chi_0 < \chi_1$, $A(\chi) = [0,1]$, $\chi \in X$. The transition law is given by

$$Q(\{\chi_0\}|\chi_0, a) = p, \tag{28a}$$
$$Q(\{\chi_1\}|\chi_0, a) = 1 - p, \tag{28b}$$
$$Q(\{\chi_1\}|\chi_1, a) = q, \tag{28c}$$
$$Q(\{\chi_0\}|\chi_1, a) = 1 - q, \tag{28d}$$

for all $a \in [0,1]$, where $0 \leq p \leq 1$ and $0 \leq q \leq 1$. The reward is given by a function $R(\chi, a)$, $(\chi, a) \in \mathbb{K}$ that met:

Assumption 6.1. **(a)** *R depends only of a, that is $R(\chi,a) = U(a)$, for all $(\chi,a) \in \mathbb{K}$, where U is nonnegative and continuous.*

(b) *There is $a^* \in [0,1]$ such that*

$$max_{a \in [0,1]} U(a) = U(a^*), \tag{29}$$

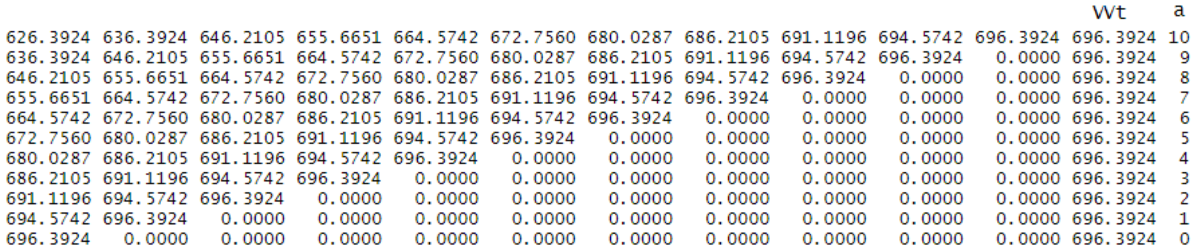
for all $\chi \in X$.

An interpretation of this example is given in the following remark.

Remark 6.2. *The states χ_0 and χ_1 could represent the behavior of a certain stock market, which is bad ($\equiv \chi_0$) and good ($\equiv \chi_1$). It is assumed that , for each a and $t = 0, 1, \cdots$, the probability of going from χ_0 to χ_0 is p (resp. the probability of going from χ_0 to χ_1 is $1-p$); moreover, for each a and $t = 0, 1, \cdots$, the probability of going from χ_1 to χ_1 is q (resp. the probability of going from χ_1 to χ_0 is $1-q$). Now, specifically, suppose that in a dynamic portfolio choice problem, two assets are available to an investor. One is risky-free, and the risk-rate $r > 0$ is assumed known and constant over time. The other asset is risky with a stochastic return having mean μ and a variance σ^2. Following Example 1.24 in (Webb, 2007), the expected utility of the investor could be given for the expression:*

$$U(a) = a\mu + (1-a)r - \frac{k}{2} a^2 \sigma^2, \tag{30}$$

where $a \in [0,1]$ is the fraction of its money that the investor invests in the risky asset and the remainder $1-a$, he/she invests in the riskless asset. In (30), k represents the value that the investor places on the variance relative to the expectation. Observe that if $\mu > \frac{k\sigma^2}{2}$, then U defined in (30) is positive in $[0,1]$ (in fact, in this case $U(0) = r > 0$ and $U(1) = \mu - \frac{k\sigma^2}{2} > 0$); moreover, it is possible to prove (see (Webb, 2007)) that if $0 < \mu - r < k\sigma^2$, then $max_{a \in [0,1]} U(a)$ is attained for $a^* \in (0,1)$ given by



```
                                                                                                     Wt      a
626.3924 636.3924 646.2105 655.6651 664.5742 672.7560 680.0287 686.2105 691.1196 694.5742 696.3924 696.3924 10
636.3924 646.2105 655.6651 664.5742 672.7560 680.0287 686.2105 691.1196 694.5742 696.3924   0.0000 696.3924  9
646.2105 655.6651 664.5742 672.7560 680.0287 686.2105 691.1196 694.5742 696.3924   0.0000   0.0000 696.3924  8
655.6651 664.5742 672.7560 680.0287 686.2105 691.1196 694.5742 696.3924   0.0000   0.0000   0.0000 696.3924  7
664.5742 672.7560 680.0287 686.2105 691.1196 694.5742 696.3924   0.0000   0.0000   0.0000   0.0000 696.3924  6
672.7560 680.0287 686.2105 691.1196 694.5742 696.3924   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  5
680.0287 686.2105 691.1196 694.5742 696.3924   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  4
686.2105 691.1196 694.5742 696.3924   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  3
691.1196 694.5742 696.3924   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  2
694.5742 696.3924   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  1
696.3924   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 696.3924  0
```

Figure 3: V_0 state-action matrix.

$$a^* = \frac{\mu - r}{k\sigma^2}. \tag{31}$$

Hence, taking $R(\chi,a) = U(a)$, $\chi \in X$, and $a \in [0,1]$, where U is given by (30), and considering the last two inequalities given in the previous paragraph, Assumption 6.1 holds.

Lemma 6.3. *Suppose that Assumption 6.1 holds. Then, for Example 2, $V(\chi) = \dfrac{U(a^*)}{1-\alpha}$ and $f^*(\chi) = a^*$, for all $\chi \in X$.*

Proof. Firstly, the value iteration functions will be found: V_n, for $n = 1,2,\ldots$.

By definition,

$$V_1(\chi_0) = \max_{a\in[0,1]} U(a), \tag{32}$$

this implies that $V_1(\chi_0) = U(a^*)$. In a similar way, it is possible to obtain that $V_1(\chi_1) = U(a^*)$.

Now, for $n = 2$,

$$\begin{aligned} V_2(\chi_0) &= \max_{a\in[0,1]} \{U(a) + \beta[V_1(\chi_1)(1-p) + V_1(\chi_0)p]\} \\ &= U(a^*) + \beta[V_1(\chi_1)(1-p) + V_1(\chi_0)p] \\ &= U(a^*) + \beta[U(a^*)(1-p) + U(a^*)p] \\ &= U(a^*) + \beta U(a^*). \end{aligned}$$

Analogously, $V_2(\chi_1) = U(a^*) + \beta U(a^*)$. Continuing this way, it is obtained that

$$V_n(\chi_0) = V_n(\chi_0) = U(a^*) + \beta U(a^*) + \ldots + \beta^{n-1} U(a^*), \tag{33}$$

for all $n = 1,2,\ldots$.

By Theorem 3.4 d), $V_n(\chi) \to V(\chi)$, $n \to \infty$, $\chi \in X$, which implies that $V(\chi) = \dfrac{U(a^*)}{1-\beta}$, $\chi \in X$. And, from the Dynamic Programming Equation (see (15)), it follows that $f^*(\chi) = a^*$, for all $\chi \in X$. □

Lemma 6.4. *For the fuzzy version of Example 2, following Theorem 4.6, it results that*

$$\widetilde{V}(\chi) = \left(0, \frac{U(a^*)}{1-\beta}, \frac{D}{1-\beta}\right),$$

$\chi \in X$, with $\tilde{R}(x,a) = (0,U(a),D)$, $\chi \in X$, $a \in A(\chi)$, where $D > U(a^*)$ and $f^*(\chi) = a^*$, for all $\chi \in X$.

6.2 Example 2

Let $X = \{\chi_0, \chi_1\}$, $0 < \chi_0 < \chi_1$, $A(\chi) = [0,\chi]$, $\chi \in X$. The transition law is given by

$$Q(\{\chi_1\}|\chi_0,a) = 1, \tag{34a}$$
$$Q(\{\chi_0\}|\chi_1,a) = 1, \tag{34b}$$

for all $\chi \in X$ and $a \in [0,\chi]$.

For this example, consider a person who will have some kind of resource available to him/her at each period of time; he/she will receive a profit depending on the amount of resource consumed. Then, χ_0 and χ_1 are the amounts available to the person at each time t. The reward R will be specified in Assumption 6.5 below.

Assumption 6.5. *R is given by*

$$R(\chi,a) = U(a) = a^\gamma, \tag{35}$$

$a \in [0,\chi]$, $\chi \in X$ and $0 < \gamma < 1$. Observe that U is non-negative, concave and continuous (taking, as it is common $U(0) = 0$), and that

$$max_{a\in[0,\chi]} U(a) = U(\chi) = \chi^\gamma, \tag{36}$$

for all $\chi \in X$.

Lemma 6.6. *Suppose that Assumption 6.5 holds. Then, for Example 3,*

$$V(\chi_0) = \frac{\chi_0^\gamma}{1-\beta^2} + \frac{\beta\chi_1^\gamma}{1-\beta^2}, \tag{37a}$$
$$V(\chi_1) = \frac{\chi_1^\gamma}{1-\beta^2} + \frac{\beta\chi_0^\gamma}{1-\beta^2}, \tag{37b}$$

and $f^*(\chi) = \chi$, for all $\chi \in X$.

Proof. Similar to the proof of Lemma 6.3. □

Lemma 6.7. *For the fuzzy version of Example 3, following Theorem 4.6, it results that $\widetilde{V}(\chi) = (0, V(\chi), \frac{D}{1-\beta})$, $\chi \in X$, with $\tilde{R}(\chi,a) = (0,U(a),D)$, $\chi \in X$, $a \in A(\chi)$, where $D > max\{\chi_1^\gamma, \chi_2^\gamma\}$ and $f^*(\chi) = \chi$, for all $\chi \in X$.*

7 CONCLUSION

In this article, Markov decision processes were studied under the total expected discounted reward criterion and considering for each of them a fuzzy reward function, specifically of the triangular type. These processes were induced from crisp processes taking into account some of their properties to cause certain properties in the fuzzy case, and the main interpretation of them is given in Theorem 4.6 and Remark 4.7. The theory was illustrated by three interesting examples from applied areas, as operations research, economics, and finance. Future work in the direction of this paper consists in applying the methodology used to other criteria of optimality such as the average case or the risk-sensitive criterion. Moreover, it is possible to contemplate the extension to another class of fuzzy payment function, for example, to trapezoidal numbers.

REFERENCES

Aliprantis, C. D. and Border, K. (2006). *Infinite dimensional analysis*. Springer.

Ban, A. I. (2009). Triangular and parametric approximations of fuzzy numbers inadvertences and corrections. *Fuzzy Sets and Systems*, 160(21):3048–3058.

Driankov, D., Hellendoorn, H., and Reinfrank, M. (2013). *An introduction to fuzzy control*. Springer Science & Business Media.

Efendi, R., Arbaiy, N., and Deris, M. M. (2018). A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. *Information Sciences*, 441:113–132.

Fakoor, M., Kosari, A., and Jafarzadeh, M. (2016). Humanoid robot path planning with fuzzy Markov decision processes. *Journal of Applied Research and Technology*, 14(5):300–310.

Hernández-Lerma, O. (1989). *Adaptive Markov control processes*, volume 79. Springer Science & Business Media.

Klir, G. J. and Yuan, B. (1996). Fuzzy sets and fuzzy logic: theory and applications. *Possibility Theory Versus Probab. Theory*, 32(2):207–208.

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.

Pedrycz, W. (1994). Why triangular membership functions? *Fuzzy Sets and Systems*, 64(1):21–30.

Puri, M. L., Ralescu, D. A., and Zadeh, L. (1993). Fuzzy random variables. In *Readings in Fuzzy Sets for Intelligent Systems*, pages 265–271. Elsevier.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Semmouri, A., Jourhmane, M., and Belhallaj, Z. (2020). Discounted Markov decision processes with fuzzy costs. *Annals of Operations Research*, pages 1–18.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton university press.

Webb, J. N. (2007). *Game theory: decisions, interaction and Evolution*. Springer Science & Business Media.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zeng, W. and Li, H. (2007). Weighted triangular approximation of fuzzy numbers. *International Journal of Approximate Reasoning*, 46(1):137–150.