KYBERNETIKA — VOLUME 57 (2021), NUMBER 1, PAGES 1–14

BI-PERSONAL STOCHASTIC TRANSIENT MARKOV GAMES WITH STOPPING TIMES AND TOTAL REWARD CRITERION

Victor M. Martínez–Cortés

The article is devoted to a class of Bi-personal (players 1 and 2), zero-sum Markov games evolving in discrete-time on Transient Markov reward chains. At each decision time the second player can stop the system by paying terminal reward to the first player. If the system is not stopped the first player selects a decision and two things will happen: The Markov chain reaches next state according to the known transition law, and the second player must pay a reward to the first player. The first player (resp. the second player) tries to maximize (resp. minimize) his total expected reward (resp. cost). Observe that if the second player is dummy, the problem is reduced to finding optimal policy of a transient Markov reward chain. Contraction properties of the transient model enable to apply the Banach Fixed Point Theorem and establish the Nash Equilibrium. The obtained results are illustrated on two numerical examples.

Keywords: two-person Markov games, stopping times, stopping times in transient Markov decision chains, transient and communicating Markov chains

Classification: 91A50, 91A05

1. INTRODUCTION

This contribution is devoted to optimality in Markov games evolving on transient (not necessarily) communicating Markov chains with two-players with opposite aims. We focus attention on the model where the first player tries to maximize his total reward and the second player tries to stop the game and receive the reward. This model is an extension of the problem studied by Cavazos-Cadena and Hernández-Hernández (see [2]). Observe that if the second player is dummy, the problem is reduced to finding optimal policy of the Markov Decision Process (MDP) introduced by Howard (for details see the monograph Puterman [10]). The main goal of the paper is to find a Nash Equilibrium in zero-sum transient stochastic games with stopping times and total reward optimality.

Recall that the considered MDP is transient if and only if the spectral radius of any admissible transition probability matrix is less than one. In particular, if the transition probability matrix is irreducible (i. e. if all states are communicating) then the model is transient if for at least one state probability of reaching any other state of the system is

DOI: 10.14736/kyb-2021-1-0001

less than one. Obviously, models with discounting are a very special case of the transient MDP.

During the 50's stochastic games were introduced by their pioneer, Shapley [13], who for first time proposed the brilliant idea to repeat a game but not necessarily the same one at each state. This opens a whole bunch of applications on different subjects, some of which can be viewed in [8]. One branch of the stochastic games are the transient games which consist in determining the end of a game based on a specific characteristic of the transition law.

In general, Markov Decision Processes (MDPs) can be seen as Stochastic Games with only one player but this is a whole topic that can be checked in [10]. For the objective of this article the line developed in: [3, 11, 18, 20, 21], and [15, 22] was followed and for the topic of the Stopping Times [14].

In the first Section we introduce the basic knowledge for this article, for the second Section we develop all the conditions we are going to use for the first player while on the third Section we introduce the second player and our main problem. On the next Section we develop the main result, for this topic we look for a fixed point on a specific set and after we need to demonstrate that this fixed point its exactly reached with the strategies which follow the Nash Equilibrium. For the last Section we show some examples: first one with a single Nash Equilibrium and the second with Multiple Equilibria and we finish the paper given some concluding remarks and the references.

2. PRELIMINARES

Definition 2.1. A *Markov Control Model* (MCM), [10] is a quintuple

$$M := \{X, A, \{A(i)|i \in X\}, Q, R\} \text{ that consists on:}$$

1. X is a finite set that will be referred to as the *state space*; whereas "i" stands for an arbitrary element of X.
2. A is a finite set called the *action space* or the *control space*.
3. $\{A(i)|i \in X\}$ a family of non-empty subsets $A(i)$ of A where $A(i)$ represents the subset of *admisible controls* for the state $i \in X$. $\mathbb{K} := \{(i, a)|i \in X, a \in A(i)\}$ is the *space of admisible state-action pairs*.
4. $Q(B|i, a)) := P(X_{t+1} \in B|X_t = i, A_t = a)$, $B \in X$ where $t = 0, 1, 2, \ldots$.
5. $R : \mathbb{K} \to \mathbb{R}$ as the *reward function* in the sense of the result obtained by applying the action a when the state was "i". Notice that we are going to consider that $0 \leq R(i, a) < \infty$, $\forall\, i, \forall\, a$, where $\mathbb{R}$ denote the real numbers.

We consider a control stochastic system and we suppose that the system can be observed at each epoch, i. e. the MCM represent the previous consideration with space state X and space of controls A, this system is observed at each time $t = 0, 1, \ldots$. So we can denote by X_t and A_t the *state* of the system and the *action* at time t. Then, the development of the system can be described as follows: If the system is in state $X_t = i \in X$ at time t and it takes the control $A_t \in A(i)$ then we obtain a response to the system $R(i, a)$ as

a consequence of the action $A_t = a$ on the state i then, the system moves to the next state X_{t+1} which is a random variable X-valued with distribution $Q(\cdot|i,a)$. So once the process is in the new state, we will have the same conditions to choose another control and the process continue. A MCM has the characteristic that at any state the response of the system and the transition law only depend of the current state of the system.

2.1. Policies or strategies

Considering a MCM for each $t = 0, 1, 2, \ldots$, we define the space H_t of admissible histories until time t as follows: $H_t := \mathbb{K}^t \times X = \mathbb{K}^t \times H_{t-1}$ for $t = 1, 2, \ldots$ and $H_0 := X$. A generic element of H_t is denoted by $h_t = (i_0, a_0, \ldots, i_j, a_j, \ldots, i_t)$, where $a_j \in A(i_j)$.

Definition 2.2. A *Policy or a Decision Strategy* [10] is a sequence $\pi = \{\pi_t, t = 0, 1, \ldots\}$ of probability measures π_t over the set of control A given H_t which satifies $\pi_t(A(i_t)|h_t) = 1$ for all $h_t \in H_t$, $t = 0, 1, \ldots$.

Definition 2.3. Let F be the set of all the functions $f : X \to A$ such that $f(i) \in A(i)$ for all $i \in X$. We will called *Stationary Policies* to the ones that there is a function $f \in F$ such that $\pi_t(\cdot|h_t)$ its concentrated on $f(i_t) \in A(i_t)$ for all $h_t \in H_t$ and $t = 0, 1 \ldots$.

Observe that the set of all *Policies* is denoted by $\mathcal{P}$ and the set of all *Deterministic Stationary Policies* by $\mathbb{F}$. Its clear that $\mathbb{F} \subset \mathcal{P}$.

Given the policy π and the initial state $X_0 = i$, a unique probability measure P_i^π is determined in the product space $\mathbb{H} := \prod_{t=0}^{\infty} \mathbb{K}$ of all possible realizations of the state-action process $\{(X_t, A_t)\}$ ([1], [10]); the corresponding expectation operator is denoted by E_i^π.

An *Objective Function* is a function that accomplish that:

$$V : \mathcal{P} \times X \to \mathbb{R},$$

which is a way to measure the result through the whole process.

Definition 2.4. Given a MCM $\{X, A, \{A(i)|i \in X\}, Q, R\}$, the set of policies $\mathcal{P}$ and the objective function V. The *Optimal Control Problem* consists on determine a policy $\pi \in \mathcal{P}$ if this exists, such that:

$$V(\pi^*, i) = \sup_{\pi \in \mathcal{P}} V(\pi, i), i \in X.$$

We define the *Total Expected Reward Optimal Value Function* as:

$$\mathcal{V}(i) = \sup_{\pi \in \mathcal{P}} E_i^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \text{ for all } i \in X, \text{ for all } \pi \in \mathcal{P},$$

and in the same sense we will say that π^* is *Optimal* if

$$\mathcal{V}(i) = V(\pi^*, i), \text{ for all } i \in X,$$

where the *Total Expected Reward Objective Function* is defined by:

$$V(\pi, i) = E_i^\pi \left[\sum_{t=0}^{\infty} R(X_t, A_t) \right], \text{ for all } i \in X.$$

3. TRANSIENT MARKOV MODEL

Definition 3.1. A *Transient Markov Control Model* (TMCM) [4] is a Markov Control Model with the addition of the following condition: there is a state $N \in X$ that $Q(\{N\}|N,a) = 1$ and $R(N,a) = 0$ for all $a \in A(N)$.

The existence of the absorbing state N on a TMCM implies that the Total Reward is bounded if the state N can be reached from any state of X. Notice that, for any $\beta \in (0,1)$ a MCM endowed with the discounted criterion $V_\beta(\pi, i) = E_i^\pi \left[\sum_{t=0}^{\infty} \beta R(X_t, A_t)\right]$, for all $i \in X, \pi \in \mathcal{P}$, can be seen as a very special case of a TMCM.

Remark 3.2. Analogously, a Transient Markov Control Model on game environment (see [4], page 161), is a model for which $\forall i \in X$, $\forall\ \pi \in \mathcal{P}$, the following *transient condition* holds:

$$\sum_{t=0}^{\infty} P_i^\pi(X_t \neq N) < \infty,$$

which is equivalent to

$$\forall i \in X, \forall\ \pi \in \mathcal{P}, \sum_{t=0}^{\infty} \sum_{j=1}^{N-1} P_i^\pi(X_t = j) < \infty.$$

knowing as the *Transient Condition*.

Theorem 3.3. The Transient Condition guarantees that the Total Reward Problem with infinite horizon results to be finite, i. e. for all initial state $i \in X$, $\forall\ \pi$ stationary, it holds that

$$V(\pi, i) = \sum_{t=0}^{\infty} E_i^\pi R(X_t, A_t) < \infty.$$

P r o o f .

$$\begin{aligned} |E_i^\pi R(X_t, A_t)| &= |\sum_{k \in X} \sum_{a \in A(k)} R(k,a) P_i^\pi(X_t = k, A_t = a)| \\ &= |\sum_{k=1}^{N-1} \sum_{a \in A(k)} R(k,a) P_i^\pi(X_t = k, A_t = a)| \\ &\leq \mathbf{S} P_i^\pi(X_t \neq N) \end{aligned}$$

where $\mathbf{S} = \max_{k \in X, a \in A(k)} |R(k,a)|$. Then

$$\sum_{t=0}^{\infty} |E_i^\pi R(X_t, A_t)| \leq \mathbf{S} \sum_{t=0}^{\infty} P_i^\pi(X_t \neq N) < \infty,$$

by using Remark 3.2. □

Given a stochastic system if we denote by τ the *Stopping Time* until $X_t = N$ that means

$$\tau = \inf\{t | X_t = N\} \text{ or } \tau = \infty \text{ if } X_t \neq N \text{ for all } t.$$

Theorem 3.4. If we have that our model acomplish the Transient Condition then it holds that $P_i^\pi(\tau < \infty) = 1$ for all policy π stationary and any state i.

P r o o f . Without loosing generality, let us consider the previous system changing the reward function by this $R(i,a) = 1$ if $i \neq N$ and $R(N,a) = 0$. As we can observe, the new stochastic system is still being transient, because the Transient Condition does not depend of the reward function. So, with the new system if there is $w \in [\tau = \infty]$, it holds that:

$$\sum_{t=0}^{\infty} R(X_t(w), A_t(w)) = \infty,$$

then, if for some initial state i, there exists a policy π which that accomplish $\tau = \infty$ with positive probability $P_i^\pi(\tau = \infty) > 0$ we obtain that:

$$\sum_{t=0}^{\infty} E_i^\pi R(X_t, A_t) = \infty.$$

Then, we obtain a contradiction to the Transient Condition. As a consequence, $\tau < \infty$. □

Remark 3.5. For our case, as we have that the set of the states X and $A(i)$ are finite for all $i \in X$, then also the set $\mathbb{K}$ is finite, as a consequence, R is trivially a continuous function, even more, $\| R \|$ is finite, where $\| R \| := \sup_{k\in\mathbb{K}} \mid R(k) \mid$, also we have that $\forall \pi \in \mathcal{P}$ the Transient Condition holds, then

$$\sum_{t=0}^{\infty} E_i^\pi R(X_t, A_t) < \infty.$$

Its easy to observe that:

$$\sum_{t=0}^{\infty} |E_i^\pi R(X_t, A_t)| \leq \| R \| \sum_{t=0}^{\infty} P_i^\pi(X_t \neq N) < \infty.$$

Even more, by using $R(i,a) = 1$ on the previous equation for the set $\mathbb{K} \setminus \{(N,a) | a \in A(N)\}$, this is equivalent to $\sum_{t=0}^{\infty} P_i^\pi(X_t \neq N) = E_i^\pi[\tau] < \infty$.

4. STOCHASTIC GAMES WITH STOPPING TIMES

Given the previous context, now we can propose two players: The first player has a Transient Markov Control Model (TMCM) and for the second player we can propose the concept of a Stopping Time.

As a brief, this topic concerns of a class of discrete-time, bi-personal zero sum games (that means what a player wins is what the other losses) and Markov transitions on a finite space. The idea behind this game it is that at each decision time, Player 2 can stop the system paying a Terminal Reward to Player 1 and if the system is not stopped, then Player 1 selects an action that moves the system into the next state receiving a

reward from Player 2. In this work the system is going to be measured by the Total Reward criterion.

Let $\mathcal{G} = (X, A, \{A(i)\}_{i\in X}, P, R, G)$ stands for a zero-sum stopping sequential game with two players, 1 and 2, where the state space X is a finite set endowed with the discrete topology and the action set A is a finite space. Player 1 is playing a TMCM $\{X, A, \{A(i) \mid i \in X\}, P, R\}$ on the game environment and Player two is deciding to stop the game at the price of paying a Terminal Reward G to the Player 1. Introducing the *Terminal Reward* G for player 2 as $G : X \to \mathbb{R}$ which is the result to decide to stop the game at state "i" we may assume $G(i) \geq R(i) \geq 0$.

Note: Given a topological space $\mathbb{K}$, the *Banach Space* $B(\mathbb{K})$ consists of all continuous functions $\hat{R} : \mathbb{K} \to \mathbb{R}$ whose supremum norm $\| \hat{R} \|$ is finite, analogous we define $B(X)$. Notice, that for our work we are going to concentrate on $R \geq 0$ and $G \geq 0$, and we have that $R \in B(\mathbb{K})$ and $G \in B(X)$.

The model $\mathcal{G}$ is interpreted as follows: at each time $t = 0, 1, 2, \ldots$, Players one and two observe the state of the system, say $X_t = i \in X$, and Player 2 can decide to stop the system at the expense of paying a Terminal Reward $G(i)$ to Player 1, or else Player 2 can decide to let the system continue its evolution. For these case, Player 1 uses the history of previously observed states and actions applied, as well as the current state $X_t = i$, to select an action (control) $A_t = a \in A(i)$ to drive the system. As a consequence, Player 1 gets a reward $R(i, a)$ from Player 2 and, regardless of the previous states and actions, the state of the system at time $t + 1$ will be $X_{t+1} = j \in X$ with probability $p_{ij}(a)$.

4.1. Strategies for $\mathcal{G}$

In the case of Player one, the strategies (policies) are the inherited by the TMCM. For notation, let $\mathcal{P}$ the set of all the strategies.

Let $\mathcal{F}_t := \bar{\sigma}(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$, where all the elements used are based on the strategies of the Player 1. ($\bar{\sigma}(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$ represents the σ algebra which is generated by the elements $(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$).

The strategy set for Player two is the space $\mathcal{T}$ which consists of all the *Stopping Times* $\tau : \mathbb{H} \to \mathbb{N}$, ($\mathbb{H} := \prod_{t=0}^{\infty} \mathbb{K}_t$, where $\mathbb{K}_t \equiv \mathbb{K}$ and $\mathbb{N} := \{1, 2, 3, \ldots, \} \bigcup \{\infty\}$) with respect to the filtration $\{\mathcal{F}_t\}$, i. e. for each non-negative integer t, the event $[\tau = t]$ belongs to $\mathcal{F}_t$. ($\mathbb{K} := \{(i, a) \mid i \in X, a \in A(i)\}$ known as the available state-action pairs).

Let's observe that the game $\mathcal{G}$ is played on the product space $\mathbb{H}$ which can be built on his natural canonical way [1]; given a policy π and a state $i \in X$, there is a unique determinate probability measure on $\mathbb{H}$ that can be denoted as P_i^π (a unique probability induced on the natural product space) in which, also we can define the expected operator as E_i^π. Notice also that the Stopping Time τ also belongs to $\mathbb{H}$. We define the *Pair of Strategies* for the game $\mathcal{G}$ as: (π, τ) where $\pi \in \mathcal{P}$ and $\tau \in \mathcal{T}$.

Definition 4.1. Given an initial state $i \in X$, the *Expected Total Reward* of Player 1 corresponding to the pair $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$ is given by:

$$\begin{aligned} V(i;\pi,\tau): &= E_i^{\pi}\left[\sum_{t=0}^{\tau-1} R(X_t,A_t) + G(X_\tau) I\left[\tau < +\infty\right]\right] \\ &= V(\pi,i) - E_i^{\pi}\left[\sum_{t=\tau}^{\infty} V(\pi,i)\right] + E_i^{\pi}[G(X_\tau) I\left[\tau < +\infty\right]]. \end{aligned}$$

Remark 4.2. Notice that:

$$\mid V(i;\pi,\tau) \mid \leq E_i^{\pi}\left[\sum_{t=0}^{\tau-1} \mid R(X_t,A_t) \mid + \parallel G(X_\tau) I\left[\tau < +\infty\right] \parallel\right] \leq \parallel R \parallel E_i^{\pi}[\tau] + \parallel G \parallel < \infty,$$

as a consequence, $V(i;\pi,\tau)$ is well defined for each $\pi \in \mathcal{P}$ and for each $\tau \in \mathcal{T}$.

When Player 2 uses the strategy τ, the *Best Expected Total Reward* of Player 1 is $\sup_{\pi\in\mathcal{P}} V(i;\pi,\tau)$, and the *Value Function* of the game is:

$$V^*(i) := \inf_{\tau\in\mathcal{T}}\left[\sup_{\pi\in\mathcal{P}} V(i;\pi,\tau)\right],\ i \in X. \tag{1}$$

Interchanging the order, the *Lower-Value Function* of the game has the following form:

$$V_*(i) := \sup_{\pi\in\mathcal{P}}\left[\inf_{\tau\in\mathcal{T}} V(i;\pi,\tau)\right],\ i \in X. \tag{2}$$

Definition 4.3. A pair $(\pi^*,\tau^*) \in \Pi \times \mathcal{T}$ is a *Nash Equilibrium* if

$$V(i;\pi,\tau^*) \leq V(i;\pi^*,\tau^*),\ i \in X,\ \pi \in \mathcal{P},$$

and

$$V(i;\pi^*,\tau) \geq V(i;\pi^*,\tau^*),\ i \in X,\ \tau \in \mathcal{T}.$$

5. STOCHASTIC TRANSIENT GAMES WITH STOPPING TIMES

This kind of games are using in order to model some problems that in somehow we cannot bound a priori the number of steps in which the game will finish but it accomplish that the number of steps is finite with probability one, i. e. there is a Stopping Time on this games.

5.1. Main Theorem

In this part we are giving a characterization of the Nash Equilibrium for the Bi-personal Stochastic Transient with Stopping Time.

By remembering Remark 3.2, the Transient Condition can be rewritten as follows [16]:

$$\sup_{i\in X,\pi\in\mathcal{P}} \sum_{t=0}^{\infty} P_i^{\pi}[X_t \neq N] < \infty. \tag{3}$$

This condition is equivalent to the Simultaneous Doeblin Condition [16], [17]

$$\sup_{i\in X,\pi\in\mathcal{P}} E_i^{\pi}[\tau] < \infty. \tag{4}$$

Proposition 5.1. There exists a positive integer K such that:

$$\sum_{t=K}^{\infty} P_i^{\pi}[X_t \neq N] < 1/2, \;\; i \in X, \pi \in \mathcal{P}.$$

P r o o f . The left-hand side of the inequality is $P_i^{\pi}[\tau > K]$ and by using Markov's inequality, (3) implies that the desired conclusion holds if $K > 2M$. □

Proposition 5.2. Let $B'(X)$ be the class of all functions $W : X \to \mathbb{R}$ which satisfy $W(N) = 0$ and we define the operator $\hat{C}$ on $B'(X)$ as follows: For each $W \in B'(X)$ the function $\hat{C}[W]$ is determined by

$$\hat{C}\,[W]\,(i) = \min\left\{G(i),\; \sup_{a\in A(i)}\left[R(i,a) + \sum_{j=1}^{N} P_{ij}(a)W(j)\right]\right\}.$$

The operator $\hat{C}^K$ is contractive, where K is as in Proposition 5.1.

P r o o f . For $W, V \in B'(X)$ the following inequality holds:

$$|\hat{C}[W] - \hat{C}[V]|(i) \leq \sup_{a\in A(i)}\left[\sum_{j=1}^{N} P_{ij}(a)|W(j) - V(j)|\right].$$

Since X and A are finite this relation implies that there is a policy $f \in \mathbb{F}$ such that

$$\begin{aligned} |\hat{C}[W](i) - \hat{C}[V](i)| &\leq E_i^f[|W(X_1) - V(X_1)|] \\ &= E_i^f[|W(X_1) - V(X_1)|I[X_1 \neq N]] \\ &= E_i^f[|W(X_1) - V(X_1)|I[\tau > 1]], \end{aligned}$$

where the second equality stems from the inclusion $W, V \in B'(X)$, and the definition of τ was used in the last step. Via an induction argument if follows that there exists a Markov policy π such that. for every $i \in X$,

$$\begin{aligned} |\hat{C}^K[W](i) - \hat{C}^K[V](i)| &\leq E_i^{\pi}[|W(X_K) - V(X_K)|I[\tau > K]] \\ &\leq \| W - V \| P_i^{\pi}[\tau > K] \leq \| W - V \| /2, \end{aligned}$$

where the last inequality is due to Proposition 5.1 □

So we showed in the Proposition 5.2 that $\hat{C}^K$ is contractive, so it holds the conditions of the Banach Fixed Point Theorem [5], that means $\hat{C}$ has a fixed point $W^* \in B'(X)$. Even more, this fixed point its the only one which acommplish:

$$W^*(i) = \min\left\{G(i), \sup_{a\in A(i)}\left[R(i,a) + \sum_{j=1}^{N} P_{ij}(a)W^*(j)\right]\right\},\ i \in X.$$

For our model, W^* is exactly V^*. For each $i \in X$, let $f^*(i) \in A(i)$ be a maximizer of the term in square brackets in the above display and define the stopping time τ^* as follows: $\tau^*(i) = 1$ (stop) if $V^*(i) = G(i)$, $\tau(i) = 0$ (continue) when $G(i) > V^*(i)$.

Theorem 5.3. For the stochastic game with stopping time introduced in the Section 4, the pair (π^*, τ^*) given by Proposition 5.2 is a Nash Equilibrium.

P r o o f. Let $S^* = \{i \in X | V^*(i) = G(i)\}$ and let $\tau^*(h) = \min\{t \geq 0 | X_t \in S^*\}$, $h = (X_0, A_0, X_1, A_1, \ldots) \in \mathbb{H}$. We need to prove two inequalities for being a Nash Equilibrium.

Lets consider the first one $V(i, \pi, \tau^*) \leq V(i, \pi^*, \tau^*)$, $i \in X$, $\pi \in \mathcal{P}$, where we are using π^* the strategy which is the stream of the $f^*(i)$ as on the previous proposition.

Case 1. If $i \in S^*$ then $[\tau^* = 0]$ has probability 1 with respect P_i^{π} and $P_i^{\pi^*}$ and on this cases then both of the rewards are $G(i) = V^*(i)$.

Case 2. If $i \notin S^*$ then $V^*(i) \geq R(i,a) + \sum_{j\in X} P_{ij}(a)V^*(j)$ so for each $\pi \in \mathcal{P}$ and as $i \in X \setminus S^*$ then

$$\begin{aligned} V^*(i) &\geq E_i^{\pi}[R(X_0, A_0) + V^*(X_1)] \\ &= E_i^{\pi}\left[R(X_0, A_0)I[\tau^* > 0] + G(X_{\tau^*})I[\tau^* < 1] + I[\tau^* \geq 1]V^*(X_1)\right], \end{aligned}$$

where the second inequality is due to the relation $P_i^{\pi}[\tau^* \geq 1] = 1$, by an induction argument for each integer n and $\pi \in \mathcal{P}$ we obtain that:

$$\begin{aligned} V^*(i) &\geq E_i^{\pi}\left[\sum_{t=0}^{n-1} R(X_t, A_t)I[\tau^* > t]\right] \\ &\quad + E_i^{\pi}\left[G(X_{\tau^*})I[\tau^* < n]\right] + E_i^{\pi}\left[I[\tau^* \geq n]V^*(X_n)\right],\ i \in X \setminus S^* \end{aligned}$$

taking the limit as $n \to \infty$, via the Bounded Convergence Theorem, it follows from Definition 4.1 and Theorem 3.3 that:

$$\begin{aligned} V^*(i) &\geq E_i^{\pi}\left[\sum_{t=0}^{\infty} R(X_t, A_t)I[\tau^* > t]\right] + E_i^{\pi}\left[G(X_{\tau^*})I[\tau^* < +\infty]\right] \\ &= E_i^{\pi}\left[\sum_{t=0}^{\tau^*-1} R(X_t, A_t)\right] + E_i^{\pi}\left[G(X_{\tau^*})I[\tau^* < +\infty]\right] \\ &= V(i; \pi, \tau^*); \end{aligned}$$

Case 1 together with Case 2 leads us to obtain the desire inequality.

To complete the proof, we need to show that: $V(i,\pi^*,\tau) \geq V(i,\pi^*,\tau^*), i \in X, \tau \in \mathcal{T}$. To obtain this, lets consider the game $\hat{G} := (X, A, \{\hat{A}(i)\}_{i\in X}, P, R, G)$ obtained by shrinking the actions sets $A(i)$ to $\hat{A}(i) = \{\pi^*(i)\} = \{f^*(i)\}$, $i \in X$, and restricting the domain of $R(\cdot)$ and each $P_{ij}(\cdot)$ to $\hat{A}(i)$. For this new model, the corresponding class $\hat{\mathcal{P}}$ of strategies for player one is the singleton $\{f^*\}$ so that the (upper) value function associated with $\hat{G}$ is given by

$$\hat{V}^*(i) = \inf_{\tilde{\tau}\in\mathcal{T}} V(i;\pi^*,\tilde{\tau}),\ i \in X. \tag{5}$$

This equation can be obtained by (1) replacing $\mathcal{P}$ for $\hat{\mathcal{P}}$. Applying Proposition 5.2 to this reduce game $\hat{G}$, the function $\hat{V}^*$ is characterized as the unique solution on $B'(X)$ of the equilibrium equation

$$\hat{V}^*(i) = \min\left\{G(i), \left[R(i,f^*(i)) + \sum_{j\in X} P_{ij}(f^*(i))\hat{V}^*(j)\right]\right\},\ i \in X.$$

so we can replace $\hat{V}^*(i)$ by $V^*(i)$ using the uniqueness of the result of Proposition 5.2. Combining the last observation with (5), it follows that, for each $\tau \in \mathcal{T}$ and $i \in X$,

$$\begin{aligned} V(i;\pi^*,\tau) &\geq \inf_{\tilde{\tau}} V(i;\pi^*,\tilde{\tau}) \\ &= \hat{V}^*(i) \\ &= V^*(i), \end{aligned}$$

So we obtain that $V(i;\pi,\tau^*) \leq V^*(i) \leq V(i;\pi^*,\tau)$ so we have that $V(i;\pi^*,\tau^*) = V^*(i)$, $\forall i \in X$ and that the pair (π^*,τ^*) is a *Nash Equilibrium* and optimal for each player. □

6. EXAMPLES

In order to illustrate the previous work we will show some models with his respective details that will reveal the applications of this theory.

6.1. Example with a Unique Nash Equilibrium

Let N be a non negative number fixed and $p \in [0,1]$ then the five elements for the TMCM are the following:

i. $X := \{0,1,2,\ldots,N\}$.

ii. $A := \{0,1,2,\ldots,\lfloor N/2\rfloor\}$ where $\lfloor z\rfloor$ represents the floor of z.

iii. For each $i \in X$, $A(i) = \{1,2,\ldots,\min\{i, N-i\}\}$.

iv. The transition law is defined by $P = [p_{ij}(a)]$ for $i \in X$ and $a \in A(i)$ as : $p_{ii+a}(a) = p$, $p_{ii-a}(a) = q$ where $q = 1-p$, $p_{N0}(a) = 1$, $p_{00}(a) = q$, $p_{01}(a) = p$.

v. The reward function R for epoch as:

$$R(i,a) = 1,\ i \neq N;\ R(N,a) = 0.$$

Let $v_i(0)$ the reward for the state i and let $v_i(n)$ the maximum expected reward for the problem on the nth period using the it was started on the state i.

Using Bellman's Optimal Principle [10] (Section 4.2 Finite-Horizon Policy Evaluation), we have that the following recursion holds for $v_i(n)$

$$v_i(n) = \max_{a \in A} \left\{ R(i,a) + \sum_{j=1}^{N} p_{ij}(a) v_j(n-1) \right\}. \tag{6}$$

The last inequality it can be rewritten in matrix notation using the policies terminology as follows:

$$v(n) = \max_{\pi \in \Pi} \{R(\pi) + P(\pi)v(n-1)\}, \qquad n \in \mathbb{N}, \tag{7}$$

where $v(n)$ represents the vector with components $v_i(n)$, $i = 1, 2, \ldots N$, introducing a simple constant as [7] we can obtain:

$$\begin{pmatrix} v(n) \\ 1 \end{pmatrix} = \max_{\pi \in \Pi} \left\{ \begin{pmatrix} P(\pi) & R(\pi) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v(n-1) \\ 1 \end{pmatrix} \right\}, \qquad n \in \mathbb{N}. \tag{8}$$

Let $\mathcal{G} = (X, A, \{A(i)\}_{i \in X}, R, G, P)$ the elements for this game are: $X := \{0,1,2,3,4\}$; this implies $N = 4$, $A(i) := \{A(0) = \{1\}, A(1) = \{1\}, A(2) = \{2\}, A(3) = \{1\}, A(4) = \{0\}\}$.

Let P and R as previously defining. The reward function G for the Player 2 as follows:

$$G(X_t) = \begin{cases} 356 & \text{if } t \leq 560, \\ 356 + \sum_{k=560}^{t} (1/2)^{(k-560)} & \text{if } t > 560, \end{cases}$$

with $p = .2$ and $q = .8$.

For this particular example we have two policies that we can represent as:

$$\pi_1 = (1,1,1) \text{ and } \pi_2 = (1,2,1).$$

In which (1,1,1) represents that $A(1) = \{1\}$, $A(2) = \{1\}$ and $A(3) = \{1\}$ (respectively the same notation for the π_2). The following matrices represent the policies π_1 and π_2 respectively, by adding a simple variable and the reward (for this part we are using some ideas that are developing on [23] about the non-negative matrices and the formula (8).

$$P(\pi_1) := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ q & 0 & p & 0 & 0 & 1 \\ 0 & q & 0 & p & 0 & 1 \\ 0 & 0 & q & 0 & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P(\pi_2) := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ q & 0 & p & 0 & 0 & 1 \\ q & 0 & 0 & 0 & p & 1 \\ 0 & 0 & q & 0 & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We can observe that this two matrices has only one absorbing state if we observe the sub-matrix $P(\pi_i)$ removing the reward column, then we can analyze the absorption time. Lets take $P(\pi_1)$ and $P(\pi_2)$ restricted (sub-matrix without absorbing state), calculating the absorption time we obtained 560 and 155 respectively. (The absorption time can be calculated using the fundamental matrix [12]).

Remark 6.1. On this paper we are using the formula (8) in order to obtain the results by using computational codes. However, the process of the Total Expected Reward can be done using (6) following the references in [23].

Lemma 6.2. For Example 1 the pair ($\pi^* = \pi_1,\ \tau^* = 560$) is the Nash Equilibrium for the game $\mathcal{G}$.

Proof. Lets take the pair $(\pi_1, 560)$. The first inequality of being Nash Equilibrium holds $V(\pi_1, 560) \leq V(\pi_1, \tau)$ for all $\tau \in \mathbb{N}$ just for how the $G(X_t)$ was defined. We need to observe that π_1 is optimal for this game. So we have to check that $V(x; \pi_2, \tau^*) \leq V(x; \pi_1, \tau^*)$, we obtain that the values are (152, 356) of the Total Expected Reward respectively, we suppose that the initial state i was the first state. This can be calculate for any initial state just by doing simple calculations, so it can be observed that the inequality holds then the pair (π_1, τ^*) is the Nash Equilibrium for the Stochastic Transient Game with Stopping Time for two players. □

Remark 6.3. The pair $(\pi_2, 155)$ is not a Nash Equilibrium.

Proof. It is easy to check $V(\pi_2, 155) \leq V(\pi_2, \tau)$ just for how the $G(X_t)$ was defined but $V(x; \pi_1, 155) \leq V(x; \pi_2, 155)$, it is not accomplished because the Expected Reward for the TMCM knowing that the Player 2 will stop at $\tau^* = 155$ has Total Expected Reward (136, 98) respectively π_1 and π_2 taking that the initial state was the first. Then, the pair $(\pi_2, 155)$ is not a Nash Equilibrium. □

6.2. Example with Multiple Nash Equilibria

Specifically, following the idea of Example 1 using the matrix extension. Let's take $N = 4$, the reward function $R(i, a) = 1$, $i \neq N$ and $R(N, a) = 0$; but with transition law as follows: $p_{ii+a}(a) = p$, $p_{ii-a+1}(a) = q$, $p_{N0}(a) = 1$.

It is important to mention that doing some numerical work we can calculate the policies and their respectively Total Expected Reward. This calculation gave us 4 policies $\pi_1 = (1, 1, 1, 1),\ \pi_2 = (1, 1, 2, 1),\ \pi_3 = (1, 2, 1, 1),\ \pi_4 = (1, 2, 2, 1)$, but only 2 are optimal respect to the TMCM. This policies are:

$$\pi_2 = (1, 1, 2, 1),\ \pi_4 = (1, 2, 2, 1).$$

On this case $\pi_2 = (1, 1, 2, 1)$, $A(0) = \{1\}$, $A(1) = \{1\}$, $A(2) = \{2\}$ and $A(3) = \{1\}$. As the previous example we can observe that the associate matrices respect to the policies are:

$$P(\pi_2) := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ 0 & q & p & 0 & 0 & 1 \\ 0 & q & 0 & 0 & p & 1 \\ 0 & 0 & 0 & q & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P(\pi_4) := \begin{pmatrix} q & p & 0 & 0 & 0 & 1 \\ q & 0 & 0 & p & 0 & 1 \\ 0 & q & 0 & 0 & p & 1 \\ 0 & 0 & 0 & q & p & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Let $\mathcal{G} = (X, A, \{A(i)\}_{i\in X}, R, G, P)$, using the previous assumptions and that the reward function G for Player 2 is:

$$G(X_t) = \begin{cases} 35 & \text{if } t \leq 95, \\ 35 + \sum_{k=95}^{t} (1/2)^{k-95} & \text{if } t > 95, \end{cases}$$

with $p = .2$ and $q = .8$.

For Example 2, we have two Nash Equilibria.

For the previous part we have that $\pi_2 = (1, 1, 2, 1)$, $\pi_4 = (1, 2, 2, 1)$, are optimal for the TMCM then as a result of the calculations the Stopping Time for both is $\tau^* = 95$. Then the pairs $(\pi_2, 95)$ and $(\pi_4, 95)$ are Nash Equilibria.

7. CONCLUDING REMARKS

It is important to recognize that the next step can be done in the sense of the generalization for multi-players. Another option is to weaken the conditions for the reward function G. With respect to the player one, it is also possible to introduced some risk sensitive criterion on the reward function R. Even more, we need to develop a criterion of how to choose a Nash Equilibrium when there are multiple of them.

8. ACKNOWLEDGMENT

The author thanks all the anonymous referees for their helpful comments and suggestions which were used to aid the improvement of this paper.

The author wants to make an special mention to Dr. Raúl Montes-de-Oca and Dr. Karel Sladký for all the suggestions and guideance during the work of this paper, also an special mention to CONACYT and Universidad Autónoma Metropolitana-Iztapalapa for the grants to aid to achieve this goal.

(Received November 11, 2019)

REFERENCES

[1] E. Ash: Real Analysis and Probability. Academic Press, 1972.

[2] R. Cavazos-Cadena and D. Hernández-Hernández: Nash equilibria in a class of Markov stopping games. Kybernetika *48* (2012), 1027–1044.

[3] R. Cavazos-Cadena and R. Montes-de-Oca: Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. Math. Methods Oper. Res. *27* (2000), 137–167. DOI:10.4064/am-27-2-167-185

[4] J. A. Filar and O. J. Vrieze: Competitive Markov Decision Processes. Springer Verlag, Berlin 1996. DOI:10.1007/978-1-4612-4054-9

[5] A. Granas and J. Dugundji: Fixed Point Theory. Springer-Verlag, New York 2003.

[6] K. Hinderer: Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Springer-Verlag, Berlin 1970. DOI:10.1007/978-3-642-46229-0

[7] R. A. Howard and J. Matheson: Risk-sensitive Markov decision processes. Management Sci. *23* (1972), 356–369. DOI:10.1287/mnsc.18.7.356

[8] V. N. Kolokoltsov and O. A. Malafayev: Understanding Game Theory. World Scientific, Singapore 2010. DOI:10.1142/7564

[9] J. Nash: Equilibrium points in n-person games. Proc. National Acad. Sci. United States of America *36* (1950), 48–49. DOI:10.1073/pnas.36.1.48

[10] M. L. Puterman: Markov Decision Processes – Discrete Stochastic Dynamic Programming. Wiley, New York 1994. DOI:10.1002/9780470316887

[11] T. E. S. Raghavan, S. H. Tijs, O. J. and Vrieze: On stochastic games with additive reward and transition structure. J. Optim. Theory Appl. *47* (1985), 451–464. DOI:10.1007/BF00942191

[12] S. Ross: Introduction to Probability Models. Ninth edition. Elsevier 2007.

[13] L. S. Shapley: Stochastic games. Proc. National Academy Sciences of United States of America *39* (1953), 1095–1100. DOI:10.1073/pnas.39.10.1095

[14] A. Shiryaev: Optimal Stopping Rules. Springer, New York 1978.

[15] K. Sladký and V. M. Martínez-Cortés: Risk-sensitive optimality in Markov games. In: Proc. 35th International Conference Mathematical Methods in Economics 2017 (P. Pražák, ed.). Univ. Hradec Králové 2017, pp. 684–689.

[16] L. C. Thomas: Connectedness conditions used in finite state Markov decision processes. J. Math. Anal. Appl. *68* (1979), 548–556. DOI:10.1016/0022-247X(79)90135-5

[17] L. C. Thomas: Connectedness conditions for denumerable state Markov decision processes. In: Recent Developments in Markov Decision Processes (R. Hartley, L.—,C. Thomas and D. J. White, eds.), Academic Press, New York 1980, pp. 181–204.

[18] F. Thuijsman: Optimality and Equilibria in Stochastic Games. Mathematical Centre Tracts, Amsterdam 1992.

[19] J. Van der Wal: Discounted Markov games: successive approximations and stopping times. Int. J. Game Theory *6* (1977), 11–22. DOI:10.1007/BF01770870

[20] J. Van der Wal: Stochastic Dynamic Programming. Mathematical Centre Tracts, Amsterdam 1981.

[21] O. J. Vrieze: Stochastic Games with Finite State and Action Spaces. Mathematical Centre Tracts, Amsterdam 1987.

[22] L. Zachrisson: Markov games. In: Advances in Game Theory (M. Dresher, L. S. Shapley and A. W. Tucker, eds.), Princeston University Press 1964. DOI:10.1515/9781400882014-014

[23] W. H. M. Zijm: Nonnegative Matrices in Dynamic Programming. Mathematisch Centrum, Amsterdam 1983.

Martínez–Cortés Victor Manuel, Universidad Autónoma Metropolitana-Iztapalapa, Avenida San Rafael Atlixco 186, Vicentina, 09340 Iztapalapa, CDMX. Portugal.

e-mail: mat.victor.m.mtz.c@gmail.com