AIMS Mathematics, 9(9): 23997–24017.
DOI:10.3934/math.20241167
Received: 19 April 2024
Revised: 28 July 2024
Accepted: 05 August 2024
Published: 13 August 2024

https://www.aimspress.com/journal/Math

Research article

Nash equilibria in risk-sensitive Markov stopping games under communication conditions

Jaicer López-Rivero*, Hugo Cruz-Suárez and Carlos Camilo-Garay

Facultad de Ciencias Fisico Matemáticas, Benemérita Universidad Autónoma de Puebla, Av. San Claudio y Río Verde, Col. San Manuel, Ciudad Universitaria, Puebla, Pue 72570, México

* **Correspondence:** Email: jaicer.lopez@alumno.buap.mx.

Abstract: This paper analyzes the existence of Nash equilibrium in a discrete-time Markov stopping game with two players. At each decision point, Player II is faced with the choice of either ending the game and thus granting Player I a final reward or letting the game continue. In the latter case, Player I performs an action that affects transitions and receives a running reward from Player II. We assume that Player I has a constant and non-zero risk sensitivity coefficient, while Player II strives to minimize the utility of Player I. The effectiveness of decision strategies was measured by the risk-sensitive expected total reward of Player I. Exploiting mild continuity-compactness conditions and communication-ergodicity properties, we found that the value function of the game is described as a single fixed point of the equilibrium operator, determining a Nash equilibrium. In addition, we provide an illustrative example in which our assumptions hold.

Keywords: Markov stopping games; risk-sensitive total expected reward; hitting time; certainty equivalent; birth-and-death chains
Mathematics Subject Classification: 91A05, 91A30, 93C55, 93E20

1. Introduction

This paper explores a particular class of discrete-time zero-sum games involving two players, characterized by countable state space and Markovian transitions. The game dynamics are structured as follows: At each decision time, Player II can either stop the game and pay a terminal reward to Player I or allow the system to continue evolving. In the latter case, Player I applies an action that affects the transitions and entitles him to receive a running reward from Player II. It is assumed that Player I has a constant and non-null risk-sensitivity coefficient and that Player II tries to minimize the utility of Player I. The performance of a pair of decision strategies is measured by the risk-sensitive total expected reward of Player I. Then, while Player I strives to maximize his utility, the objective of

Player II is to minimize it.

The primary objective of this paper is twofold: First, to formulate an equilibrium equation delineating the value function of the game; second, to establish the existence of a Nash equilibrium. To achieve these goals, we rely on standard conditions of continuity and compactness, as outlined in Assumption 2.1. Additionally, we assume that if Player II chooses not to terminate the game, the Markov chain induced by any stationary policy adopted by Player I exhibits communication properties and has a stationary distribution, as stated in Assumption 2.2.

The modeling framework referenced in [1] is designed such that only one participant has the authority to halt the game, while the counterpart influences its progression. Initially conceptualized within a gambling framework, this model introduces an intriguing dynamic between players. The Dynkin game, as explored in [2], represents an adaptation of the optimal stopping dilemma, where two agents monitor a Markov chain with the option to terminate its progression, albeit at a terminal cost. The concept of stopping times plays a pivotal role in the realm of stochastic analysis, offering a rich area for exploration. An extensive exploration of this topic, along with a thorough exposition of the underlying theory, is presented in [3] and [4]. Moreover, the application of these concepts extends into the field of financial mathematics, where they provide valuable insights, as delineated in [5, 6].

Game theory finds widespread applications across diverse fields, as exemplified by works such as [7, 8] and [9, 10]. The genesis of Markov game theory can be traced back to seminal works by Shapley [11] and Zachrisson [12]. Generally, Markov decision processes can be conceptualized as stochastic games with a single player. A detailed exposition of Markov decision processes is available in foundational texts such as [13, 14].

The examination of discrete-time Markov models that incorporate risk-sensitive criteria has roots dating back at least the work of [15], with subsequent research motivated by intersections with mathematical finance [16, 17] and the theory of large deviations [18, 19]. Controlled Markov models featuring finite or countable state spaces, along with risk-sensitive criteria, have been subject to investigation in works such as [20–22]. Furthermore, Markov decision processes, situated within a general state space, have been analyzed in publications such as [23] and [24].

Markov stopping games endowed with the risk-neutral total reward criterion have been studied in [25] and [26], where the existence of a Nash equilibrium was proved assuming that the state space is finite and denumerable, respectively. Both papers assume an absorbing state where rewards are zero, and that this state is attainable under any stationary policy from any initial state. Under the same assumptions, these results were extended to the risk-sensitive case in [27]. In line with the study of Markov stopping games under a risk-sensitive criterion, recent publications such as [28] investigated a model under the condition that rewards are strictly positive, while [29] considered discrete-time stopping games under a discounted cost criterion. Furthermore, in [30], a Markov game with a total expected payoff was studied under an absorption condition on the components of the game. In [31], the risk-neutral case was also investigated with a more general model that does not assume the presence of an absorbing state, but rather considers communication-ergodicity conditions similar to those assumed in this paper.

In summary, our contribution to this study is twofold. First, we extend the findings of [31] to a risk-sensitive scenario. Second, we present a concrete example of a specific game where all the assumptions outlined in our study are met, thereby illustrating the practical applicability and validity of our theoretical results. Specifically, concerning the first point, we address two main problems:

- Establishing an equilibrium equation that characterizes the value function of the game.
- Guaranteeing the existence of a Nash equilibrium.

The manuscript is organized as follows: Section 2 provides a formal description of the components of the Markov stopping game model, as well as the assumptions considered in this model. The strategies of the players are introduced, and the risk-sensitive total reward criterion is formulated. Furthermore, the idea of Nash equilibrium is discussed and an example in which all the assumptions hold is provided. Section 3 is devoted to the presentation of the so-called equilibrium operator, which plays a pivotal role in this work. Theorem 3.2 demonstrates the main property of this operator, the existence and uniqueness of a fixed point. In Section 4, this fixed point is used to define the strategies of the players that form a Nash equilibrium. This is the main result of the paper and is stated in Theorem 4.1. In Section 5, we present a numerical example to illustrate the methodology for determining the Nash equilibrium in a practical context. The paper concludes with brief comments in Section 6.

2. The game model

This section formally describes the Markov stopping game model. However, before proceeding, it is necessary to introduce the basic notation used in the following analysis. Given a topological space X, the Banach space $C(X)$ consists of all continuous functions $R : X \to \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers, with a finite supremum norm $\|R\|$, defined as $\|R\| := \sup_{k\in X} |R(k)|$. Additionally, $\mathbb{N}$ denotes the set of non-negative integers. The indicator function of a set A is denoted by $I[A]$ and, unless otherwise specified, all relations involving conditional expectations are valid with probability 1 concerning the underlying probability measure. Furthermore, the infimum of the empty set is defined as ∞. Finally, the following convention concerning summations will be used

$$\sum_{t=n}^{m} a_t := 0, \quad m < n. \tag{2.1}$$

A Markov stopping game $\mathcal{G} = (S, A, \{A(x)\}_{x\in S}, R, G, P)$ is a mathematical model for a dynamic system whose evolution is influenced by two agents, which we call Players I and II. The components of $\mathcal{G}$ have the following meaning:

- S, called the state space, is a non-empty and denumerable set and is endowed with discrete topology.
- A is the action space, which is a metric space.
- For each $x \in S$, $A(x) \subset A$ is a non-empty class of admissible actions at x for Player I.
- $R \in C(\mathbb{K})$ is the running reward function, where the class $\mathbb{K}$ of admissible pairs is defined by $\mathbb{K} := \{(x, a) : x \in S, a \in A(x)\}$, and $G \in C(S)$ is the terminal reward function.
- $P = [p_{x,y}(a)]$ is the controlled transition law on S given $\mathbb{K}$, so that $p_{x,y}(a) \geq 0$ and $\sum_{y\in S} p_{x,y}(a) = 1$ for each $(x, a) \in \mathbb{K}$.

The model $\mathcal{G}$ is interpreted as follows: At each decision time point $t \in \mathbb{N}$, Players I and II observe the state of the system, say $X_t = x \in S$, and Player II must choose one of two actions: to *stop* the system by paying a terminal reward $G(x)$ to Player I, or to let the system *to continue* its evolution. In this

latter case, Player I applies an action (control) $A_t = a \in A(x)$ using the record of states up to time t and actions previous to t. This intervention has two consequences: Player I receives a reward $R(x, a)$ from Player II, and the system moves to $X_{t+1} = y \in S$ with probability $p_{x,y}(a)$, regardless of previous states and actions. This is known as the Markov property of the decision process. The procedure described above is repeated whenever the system transitions to a new state. The goal of Player I is to choose their strategy decision in order to maximize his utility, while the objective of Player II is to minimize the utility of Player I, which is a characteristic of a zero-sum game. In this game model, we make an important and standard assumption, which follows below.

Assumption 2.1. (*i*) *For each $x \in S$, $A(x)$ is a compact subset of A.*
(*ii*) *For every $x, y \in S$, the mappings $a \mapsto R(x, a)$ and $a \mapsto p_{x,y}(a)$ are continuous in $a \in A(x)$.*
(*iii*) *For all $x \in S$ and $a \in A(x)$, $G(x) \geq 0$ and $R(x, a) \geq 0$.*
(*iv*) *There exists a state z such that $R(z, a) > 0$ for every $a \in A(z)$.*

Remark 2.1. *Items* (*i*)–(*iii*) *of Assumption 2.1 are well-established in risk-sensitive Markov decision processes (MDPs) literature and are used for ensuring the existence of optimal policies (see Remark 3.1). Additionally, item* (*iv*) *plays a crucial role in deriving the two main results of this paper, namely the uniqueness of the fixed point of the equilibrium operator and the existence of a Nash equilibrium for the game.*

To define the risk-sensitive total reward, we will introduce the strategies of players and some useful notation. For each $t \in \mathbb{N}$, we define $\mathbb{H}_t$ as the space of possible histories up to time t, where $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K} \times \mathbb{H}_{t-1}$, when $t > 0$. We use $h_t = (x_0, a_0, \ldots, x_i, a_i, \ldots, x_t)$ to represent a generic element of $\mathbb{H}_t$, where $a_i \in A(x_i)$. A policy $\pi = \{\pi_t\}$ is a special sequence of stochastic kernels, that is, for each $t \in \mathbb{N}$ and $h_t \in \mathbb{H}_t$, $\pi_t(\cdot|h_t)$ is a probability measure in A concentrated in $A(x_t)$, and for each Borel subset $B \subset A$ the mapping $h_t \mapsto \pi_t(B|h_t)$, $h_t \in \mathbb{H}_t$ is Borel measurable. The class of all policies constitutes the family of *admissible strategies for Player I* and is denoted by $\mathcal{P}$. When Player I drives the system using π, the control A_t applied at time t belongs to $B \subset A$ with probability $\pi_t(B|h_t)$, where $h_t \in \mathbb{H}_t$ is the observed history of the process up to time t. Given $\pi \in \mathcal{P}$ and the initial state $X_0 = x$, a unique probability measure P_x^π is uniquely determined on the Borel σ-field of the space $\mathbb{H} := \prod_{t=0}^{\infty} \mathbb{K}$ of all possible realizations of the state-action process $\{(X_t, A_t)\}$. The corresponding expectation operator is denoted by E_x^π. Next, define $\mathbb{F} := \prod_{x \in S} A(x)$ and notice that $\mathbb{F}$ is a compact metric space, which consists of all functions $f : S \to A$ such that $f(x) \in A(x)$ for each $x \in S$. A policy π will be called stationary if there exists $f \in \mathbb{F}$ such that the probability measure $\pi_t(\cdot|h_t)$ is always concentrated at $f(x_t)$, and in this case π and f are naturally identified; with this convention, $\mathbb{F} \subset \mathcal{P}$.

Moreover, setting

$$\mathcal{F}_t := \sigma(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t), \tag{2.2}$$

the space $\mathcal{T}$ of *strategies for Player II* consists of all stopping times $\tau : \mathbb{H} \to \mathbb{N} \cup \{\infty\}$ with respect to the filtration $\{\mathcal{F}_t\}$, that is, $[\tau = t] \in \mathcal{F}_t$ for every $t \in \mathbb{N}$. Intuitively, this condition means that the decision of Player II to stop or not to stop at time t should be based only on the information available at time t and not on any information available in the future.

The risk-sensitive total reward received by Player I measures the performance of a pair of strategies $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$. Throughout the remainder, it is assumed that Player I has a fixed and constant risk-sensitivity coefficient $\lambda \in \mathbb{R} - \{0\}$. Since Player I has a constant risk sensitivity coefficient, this means

that a random reward Y is evaluated by the expectation of $U_\lambda(Y)$, where the utility function $U_\lambda : \mathbb{R} \to \mathbb{R}$ is given by

$$U_\lambda(v) := \text{sign}(\lambda)e^{\lambda v}, \quad v \in \mathbb{R}. \tag{2.3}$$

Remark 2.2. *Note that $U_\lambda(\cdot)$ is a strictly increasing function and that*

$$U_\lambda(v + w) = e^{\lambda v}U_\lambda(w), \quad v, w \in \mathbb{R}. \tag{2.4}$$

Besides, when choosing between two random rewards W and Y, Player I will prefer Y if $E[U_\lambda(W)] < E[U_\lambda(Y)]$ and will be indifferent when $E[U_\lambda(W)] = E[U_\lambda(Y)]$. The certainty equivalent of Y (with respect to U_λ) is the constant $\mathcal{E}_\lambda(Y) \in \mathbb{R} \cup \{-\infty, \infty\}$ satisfying $U_\lambda(\mathcal{E}_\lambda(Y)) = E[U_\lambda(Y)]$, so that Player I is indifferent between receiving a random reward Y or the corresponding certainty equivalent $\mathcal{E}_\lambda(Y)$.

Definition 2.1. *Given the initial state $X_0 = x \in S$, suppose that the system has been driven by Players I and II, according to strategies $\pi \in \mathcal{P}$ and $\tau \in \mathcal{T}$, respectively. The total reward obtained by Player I until the system is halted at time τ by Player II is given by*

$$\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty],$$

and the corresponding certainty equivalent is the performance index *$V_\lambda(x; \pi, \tau)$ associated with the pair $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$ at state $x \in S$, defined by*

$$V_\lambda(x; \pi, \tau) := \frac{1}{\lambda} \log\left(E_x^\pi\left[e^{\lambda(\sum_{t=0}^{\tau-1} R(X_t,A_t)+G(X_\tau)I[\tau<\infty])}\right]\right). \tag{2.5}$$

Remark 2.3. *Note that the certainty equivalent of a reward Y can be expressed as $\mathcal{E}_\lambda(Y) = \log(E[e^{\lambda Y}])/\lambda$. This leads to the performance index $V_\lambda(x; \pi, \tau)$, as given in (2.5).*

Now we define the upper and lower value of the game. If Player II uses the strategy τ, the highest value of the certainty equivalent that can be reached by Player I is $\sup_{\pi\in\mathcal{P}} V_\lambda(x; \pi, \tau)$, which is a function of x and τ, say $\varphi(x; \tau)$. The main goal of Player II is to minimize the expected utility of his counterpart. Therefore, they will strive to choose a stopping time $\tilde{\tau}$ such that $\varphi(x; \tilde{\tau})$ is as close to $\inf_{\tau\in\mathcal{T}} \varphi(x; \tau)$ as possible. This last quantity is the *upper-value* function of the game and is explicitly determined by

$$V_\lambda^*(x) := \inf_{\tau\in\mathcal{T}}\left[\sup_{\pi\in\mathcal{P}} V_\lambda(x; \pi, \tau)\right], \quad x \in S. \tag{2.6}$$

Interchanging the order in which the supremum and the infimum are taken, the following lower-value function of the game is obtained:

$$V_{\lambda,*}(x) := \sup_{\pi\in\mathcal{P}}\left[\inf_{\tau\in\mathcal{T}} V_\lambda(x; \pi, \tau)\right], \quad x \in S. \tag{2.7}$$

Since $\sup_{\pi\in\mathcal{P}} V_\lambda(x; \pi, \tau) \geq V_\lambda(x; \pi, \tau) \geq \inf_{\tau\in\mathcal{T}} V_\lambda(x; \pi, \tau)$, these definitions lead to

$$V_\lambda^*(\cdot) \geq V_{\lambda,*}(\cdot). \tag{2.8}$$

If $V_\lambda^*(x) = V_{\lambda,*}(x)$ for all $x \in S$, then the common function is called the value of the game.

The aim of the paper is to establish the existence of a pair of strategies $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$ that form a Nash equilibrium, as defined below.

Definition 2.2. *A Nash equilibrium is a pair* $(\pi^*, \tau^*) \in \mathcal{P} \times \mathcal{T}$ *such that, for every state* $x \in S$

$$V_\lambda(x; \pi, \tau^*) \leq V_\lambda(x; \pi^*, \tau^*) \leq V_\lambda(x; \pi^*, \tau), \quad \pi \in \mathcal{P}, \quad \tau \in \mathcal{T}. \tag{2.9}$$

If the strategies π^* and τ^* actually used by Players I and II form a Nash equilibrium, Player I has no incentive to switch to a different strategy if Player II continues using strategy τ^*, as shown in the first inequality in the above display. Likewise, the latter inequality implies that if Player I continues using π^*, then Player II has no motivation to change the strategy used, τ^*. Note also that if (π^*, τ^*) is a Nash equilibrium, then (2.9) implies that

$$V_\lambda^*(\cdot) \leq \sup_\pi V_\lambda(\cdot; \pi, \tau^*) \leq V_\lambda(\cdot; \pi^*, \tau^*) \leq \inf_\tau V_\lambda(x; \pi^*, \tau) \leq V_{\lambda,*}(\cdot),$$

where the left- and right-most inequalities are due to (2.6) and (2.7), respectively, so that via (2.8), it follows that the upper and lower value functions are equal and coincide with $V_\lambda(\cdot; \pi^*, \tau^*)$.

Assumption 2.2. *For each* $f \in \mathbb{F}$*, the following conditions hold:*

(*i*) *The Markov chain induced by* f *is communicating. This means that given* $x, y \in S$*, there exists a positive integer* n *and states* $x_0 = x, x_1, \ldots, x_n = y \in S$ *such that*

$$p_{x_{i-1},x_i}(f(x_{i-1})) > 0, \quad i = 1, 2, \ldots n.$$

(*ii*) *There is a probability distribution* $\rho_f(\cdot)$ *on* S *such that*

$$\rho_f(y) = \sum_{x \in S} \rho_f(x) p_{x,y}(f(x)), \quad y \in S.$$

Remark 2.4. *a) Assumption 2.2*(*i*) *is well known in the literature on risk-sensitive MDPs (see, for instance, [31] and [32]). In particular, this assumption is used to guarantee the uniqueness of solutions of the optimality equation associated with the average cost criterion, provided the equation admits a bounded solution [32]. In this manuscript, Assumption 2.2*(*i*) *is applied in the proofs of Theorems 3.2 and 4.1 to guarantee the finiteness of hitting times.*

b) Assumption 2.2(*ii*) *requires an invariant distribution. This condition has been employed in Markov stopping games for the neutral MDPs case (see [31]) to establish the existence of a Nash equilibrium. In this document, we adopt this condition for the same purpose. The existence of such an invariant distribution can be guaranteed via the Perron-Frobenius theorem in the finite case or by applying a Harris condition in the countable case (see [33]). Moreover, observe that the invariant distribution* ρ_f *of the Markov chain induced by any* $f \in \mathbb{F}$ *must satisfy that*

$$\rho_f(x) > 0, \quad x \in S. \tag{2.10}$$

In the following example, we verify that Assumptions 2.1 and 2.2 hold in a specific Markov stopping game.

Example 2.1. *Let* N *be a fixed positive integer and consider a Markov stopping game* $\mathcal{G}$ *with the following components:*

- *State space* $S = \mathbb{N}$.

- *Action space* $A = \{b_1, b_2, \ldots, b_N\}$, *where* $0 < b_1 < b_2 < \cdots < b_N < 1$ *and* $b_N + b_1 < 1$.
- $A(x) = A$, *for each* $x \in S$.
- *The running reward and terminal reward functions are given by*

$$R(x,a) = \begin{cases} 0 & \text{if } x \geq N \\ \dfrac{1}{ax+1} & \text{if } x < N \end{cases}, \text{ for all } (x,a) \in \mathbb{K} \text{ and } G(x) = \frac{N}{x+1}, \text{ for all } x \in S.$$

- *The controlled transition law is described as follows:*

$$p_{0,1}(a) = 1,$$
$$p_{x,x+1}(a) = a,$$
$$p_{x,x-1}(a) = 1 - a,$$

for each $x \neq 0$ *and* $a \in A$.

For this game, observe that Assumption 2.1 is fulfilled. Based on the controlled transition law of the game, each $f \in \mathbb{F}$ *induces an irreducible birth-and-death chain. This follows from the fact that* $b_i > 0$ *for all* $i \in \{1, 2, \ldots, N\}$, *thus Assumption 2.2*$(i)$ *holds. Moreover, consider the following inequalities:*

$$\sum_{x=1}^{\infty} \frac{f(1)\cdots f(x-1)}{(1-f(1))\cdots(1-f(x))} \leq \frac{1}{b_N} \sum_{x=1}^{\infty} \left(\frac{b_N}{1-b_1}\right)^x < \infty,$$

where the first inequality holds because $b_1 \leq f(x) \leq b_N$, $\forall x \in S$, *and the series is convergent since* $b_N + b_1 < 1$. *Therefore, the chain induced for each* $f \in \mathbb{F}$ *is positive recurrent (see Chapter 2, Example 5 in [33]). Consequently, Assumption 2.2 is satisfied, as an irreducible, positive recurrent Markov chain has a unique stationary distribution [33].*

3. Equilibrium operator

In this section, we will prove the existence and uniqueness of the fixed point of an operator, which will be called the equilibrium operator. To this end, we introduce a subset of $C(S)$ and the operator T_λ.

Definition 3.1. (i) *The space* $[\![0, G]\!] \subset C(S)$ *is defined as:*

$$[\![0, G]\!] := \{h \in C(S) \mid 0 \leq h(x) \leq G(x)\}.$$

(ii) *The operator* $T_\lambda : [\![0, G]\!] \to [\![0, G]\!]$ *is determined as follows: For each* $W \in [\![0, G]\!]$ *and* $x \in S$,

$$U_\lambda(T_\lambda[W](x)) := \min\left\{U_\lambda(G(x)),\ \sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a) + W(y))\right]\right\}. \tag{3.1}$$

Given that $U_\lambda(\cdot)$ is increasing and R and G are non-negative, it can be verified that T_λ transforms $[\![0, G]\!]$ into itself. Additionally, T_λ is an increasing monotone operator, meaning that for $V, W \in [\![0, G]\!]$,

$$V \leq W \Rightarrow T_\lambda[V] \leq T_\lambda[W]. \tag{3.2}$$

In the rest of the manuscript, W_λ represents a fixed point of T_λ, i.e., $W_\lambda \in [\![0, G]\!]$ and also $T_\lambda[W_\lambda] = W_\lambda$. Equality (3.1) states that the latter expression can be expressed as follows: For each $x \in S$,

$$U_\lambda(W_\lambda(x)) = \min\left\{U_\lambda(G(x)),\ \sup_{a\in A(x)}\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a) + W_\lambda(y))\right\}. \tag{3.3}$$

Remark 3.1. *Given that G is bounded, the inclusion of $W_\lambda \in [\![0, G]\!]$ and the Assumption 2.1 imply that there exists a policy $f \in \mathbb{F}$ such that, for all $x \in S$,*

$$\sum_{y\in S} p_{x,y}(f(x))U_\lambda(R(x, f(x)) + W_\lambda(y)) = \sup_{a\in A(x)} \left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x, a) + W_\lambda(y))\right]. \tag{3.4}$$

Before presenting the main result of this section, we provide a lemma and a theorem that will be instrumental in the subsequent proof. It is important to note that, throughout the remainder of this document, Assumptions 2.1 and 2.2 are implicitly assumed.

Lemma 3.1. *Define S_{W_λ} as follows:*

$$S_{W_\lambda} := \{x \in S \mid W_\lambda(x) = G(x)\}.$$

Then, $S_{W_\lambda} \neq \emptyset$.

Proof. The proof proceeds by contradiction. Suppose that $S_{W_\lambda} = \emptyset$, i.e., $G(x) \neq W_\lambda(x) = T_\lambda[W_\lambda](x)$ for all $x \in S$. In consequence,

$$U_\lambda(W_\lambda(x)) = \sup_{a\in A(x)} \left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x, a) + W_\lambda(y))\right] = \sum_{y\in S} p_{x,y}(f(x))U_\lambda(R(x, f(x)) + W_\lambda(y)),$$

due to (3.3) and (3.4). This last expression can be rewritten as

$$U_\lambda(W_\lambda(x) - R(x, f(x))) = \sum_{y\in S} p_{x,y}(f(x))U_\lambda(W_\lambda(y)), \tag{3.5}$$

as a consequence of (2.4). Then, since the function $U_\lambda(\cdot)$ is strictly increasing and R is non-negative, we have that

$$U_\lambda(W_\lambda(x)) \geq U_\lambda(W_\lambda(x) - R(x, f(x))),$$

which leads us to the following inequality

$$U_\lambda(W_\lambda(x)) \geq \sum_{y\in S} p_{x,y}(f(x))U_\lambda(W_\lambda(y)).$$

From the previous inequality, we get the following expression:

$$U_\lambda(W_\lambda(x)) + \delta(x) = \sum_{y\in S} p_{x,y}(f(x))U_\lambda(W_\lambda(y)), \tag{3.6}$$

where

$$\delta(x) := \sum_{y\in S} p_{x,y}(f(x))U_\lambda(W_\lambda(y)) - U_\lambda(W_\lambda(x)) \leq 0, \ x \in S.$$

Assumption 2.2(*ii*) guarantees the existence of $\rho_f(\cdot)$, the invariant distribution of the Markov chain induced by f, and it follows that

$$\sum_{x\in S} \rho_f(x) \left[U_\lambda(W_\lambda(x)) + \delta(x)\right] = \sum_{x\in S} \rho_f(x) \left[\sum_{y\in S} p_{x,y}(f(x))U_\lambda(W_\lambda(y))\right]$$

$$= \sum_{y\in S}\left[\sum_{x\in S}\rho_f(x)p_{x,y}(f(x))\right]U_\lambda(W_\lambda(y))$$
$$= \sum_{y\in S}\rho_f(y)U_\lambda(W_\lambda(y)),$$

from which we obtain that

$$\sum_{x\in S}\rho_f(x)\delta(x) = 0\,.$$

Since $\delta(\cdot) \le 0$, this last equality and (2.10) result in $\delta(\cdot) = 0$, so (3.5) and (3.6) imply that

$$U_\lambda(W_\lambda(x) - R(x, f(x))) = U_\lambda(W_\lambda(x)).$$

Since $U_\lambda(\cdot)$ is strictly increasing, we have that $R(x, f(x)) = 0$, for all $x \in S$, contrary to the Assumption 2.1(*iv*). □

Operator T_λ is a continuous operator concerning the pointwise convergence topology in $[\![0, G]\!]$ space, which was proved in [27] and is stated in the following theorem.

Theorem 3.1. *Suppose that the sequence* $\{W_n\} \subset [\![0, G]\!]$ *converges pointwise to a function* $V : S \to \mathbb{R}$*, that is,*

$$\lim_{n\to\infty} W_n(x) = V(x), \quad x \in S.$$

In this case

$$V \in [\![0, G]\!] \qquad \textit{and} \quad \lim_{n\to\infty} T_\lambda[W_n](x) = T_\lambda[V](x), \quad x \in S.$$

The theorem that establishes the existence of a unique fixed point of the operator T_λ is stated below.

Theorem 3.2. *Under Assumptions 2.1 and 2.2, there is only one fixed point of the operator* T_λ*, i.e., there is only one function* $W^*_\lambda \in [\![0, G]\!]$ *such that*

$$W^*_\lambda = T_\lambda[W^*_\lambda]\,. \tag{3.7}$$

Proof. First, consider the following sequence: $W_{0,\lambda_0} := 0, W_{0,\lambda_1} := G$ and $W_{n,\lambda_0} := T^n_\lambda[0], W_{n,\lambda_1} := T^n_\lambda[G]$ for $n \in \mathbb{N} \setminus \{0\}$. Then

$$W_{n+1,\lambda_0} := T_\lambda[W_{n,\lambda_0}] \quad \text{and} \quad W_{n+1,\lambda_1} := T_\lambda[W_{n,\lambda_1}], \quad n \in \mathbb{N}. \tag{3.8}$$

Since $W_{0,\lambda_0}, W_{1,\lambda_0} = T_\lambda[0] \in [\![0, G]\!]$ and $W_{0,\lambda_1}, W_{1,\lambda_1} = T_\lambda[G] \in [\![0, G]\!]$, it follows that $W_{0,\lambda_0} \le W_{1,\lambda_0}$ and $W_{1,\lambda_1} \le W_{0,\lambda_1}$. Combining this with an induction argument and the property (3.2), it follows that

$$0 \le W_{n,\lambda_0} \le W_{n+1,\lambda_0} \le G \quad \text{and} \quad 0 \le W_{n+1,\lambda_1} \le W_{n,\lambda_1} \le G, \quad n \in \mathbb{N},$$

where the extreme inequalities are due to the functions W_{n,λ_0} and W_{n,λ_1} belonging to $[\![0, G]\!]$ for all $n \in \mathbb{N}$. It follows that the sequences $\{W_{n,\lambda_0}(y)\}_{n\in\mathbb{N}}$ and $\{W_{n,\lambda_1}(y)\}_{n\in\mathbb{N}}$ are monotone and bounded, so that

$$\lim_{n\to\infty} T^n_\lambda[0](y) := W_{\lambda_0}(y) \quad \text{and} \quad \lim_{n\to\infty} T^n_\lambda[G](y) := W_{\lambda_1}(y)$$

exist for all $y \in S$. Theorem 3.1 allows us to state that

$$W_{\lambda_0}, W_{\lambda_1} \in [\![0, G]\!], \tag{3.9}$$

and also

$$\lim_{n\to\infty} T_\lambda[W_{n,\lambda_0}](x) = T_\lambda[W_{\lambda_0}](x) \quad \text{and} \quad \lim_{n\to\infty} T_\lambda[W_{n,\lambda_1}](x) = T_\lambda[W_{\lambda_1}](x), \tag{3.10}$$

for all $x \in S$. Thus, by taking the limit as n approaches infinity on both sides of the equalities in (3.8) we can see that W_{λ_0} and W_{λ_1} are fixed points of the T_λ operator, as $W_{\lambda_0} = T_\lambda[W_{\lambda_0}]$ and $W_{\lambda_1} = T_\lambda[W_{\lambda_1}]$.

On the other hand, it should be noted that $W_\lambda = T_\lambda^n[W_\lambda], n \in \mathbb{N}$. By combining the inequalities $0 \leq W_\lambda \leq G$ with the property (3.2) of the operator T_λ, it can be deduced that $T_\lambda^n[0] \leq T_\lambda^n[W_\lambda] \leq T_\lambda^n[G]$ for all $n \in \mathbb{N}$. This relation, along with the observations made above and (3.10), leads to the conclusion that

$$W_{\lambda_0} \leq W_\lambda \leq W_{\lambda_1}. \tag{3.11}$$

To demonstrate the uniqueness of the fixed point of operator T_λ, it suffices to confirm that

$$W_{\lambda_0} \geq W_{\lambda_1}. \tag{3.12}$$

Note that if there is an $\hat{x} \in S$ such that $W_{\lambda_0}(\hat{x}) = G(\hat{x})$, then by (3.9) and (3.11) it follows that

$$W_{\lambda_0}(\hat{x}) = W_{\lambda_1}(\hat{x}) = G(\hat{x}). \tag{3.13}$$

Let $x \in S$, then we have that

$$\begin{aligned}
U_\lambda(W_{\lambda_0}(x)) &= U_\lambda(T_\lambda[W_{\lambda_0}](x)) \\
&= \min\left\{U_\lambda(G(x)), \sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a) + W_{\lambda_0}(y))\right]\right\} \\
&\leq \min\left\{U_\lambda(G(x)), \sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a) + W_{\lambda_1}(y))\right]\right\} \\
&\quad + \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S} p_{x,y}(a)|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|\right] \\
&= U_\lambda(T_\lambda[W_{\lambda_1}](x)) + \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S} p_{x,y}(a)|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|\right] \\
&\leq U_\lambda(W_{\lambda_1}(x)) + e^{|\lambda|\|R\|}\sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|\right].
\end{aligned}$$

Since $U_\lambda(W_{\lambda_0}) - U_\lambda(W_{\lambda_1})$ is bounded, it follows from Assumption 2.1 that there exists $\tilde{f} \in \mathbb{F}$ such that

$$\sum_{y\in S} p_{x,y}(\tilde{f}(x))|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))| = \sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|\right], \quad x \in S,$$

which implies that

$$U_\lambda(W_{\lambda_0}(x)) - U_\lambda(W_{\lambda_1}(x)) \leq e^{|\lambda|\|R\|} \sum_{y\in S} p_{x,y}(\tilde{f}(x))|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|.$$

The inequality

$$U_\lambda(W_{\lambda_1}(x)) - U_\lambda(W_{\lambda_0}(x)) \leq e^{|\lambda|\|R\|} \sum_{y\in S} p_{x,y}(\tilde{f}(x))|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|$$

is obtained by exchanging the roles of W_{λ_0} and W_{λ_1}, therefore

$$|U_\lambda(W_{\lambda_0}(x)) - U_\lambda(W_{\lambda_1}(x))| \leq e^{|\lambda|\|R\|} \sum_{y\in S} p_{x,y}(\tilde{f}(x))|U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y))|.$$

Then, since $W_{\lambda_0} \leq W_{\lambda_1}$, we have

$$\begin{aligned} U_\lambda(W_{\lambda_0}(x)) - U_\lambda(W_{\lambda_1}(x)) &\geq e^{|\lambda|\|R\|} \sum_{y\in S} p_{x,y}(\tilde{f}(x))U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y)) \\ &\geq \sum_{y\in S} p_{x,y}(\tilde{f}(x))U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y)). \end{aligned}$$

This relation, together with the Markov property, implies that for all $x \in S$ and $n \in \mathbb{N}$,

$$\begin{aligned} U_\lambda(W_{\lambda_0}(X_n)) - U_\lambda(W_{\lambda_1}(X_n)) &\geq \sum_{y\in S} p_{X_n,y}(\tilde{f}(X_n))U_\lambda(W_{\lambda_0}(y)) - U_\lambda(W_{\lambda_1}(y)) \\ &= E_x^{\tilde{f}}\left[U_\lambda(W_{\lambda_0}(X_{n+1})) - U_\lambda(W_{\lambda_1}(X_{n+1}))|\mathcal{F}_n\right]. \end{aligned}$$

Therefore, we can conclude that $\{U_\lambda(W_{\lambda_0}(X_n)) - U_\lambda(W_{\lambda_1}(X_n)), \mathcal{F}_n\}$ is a supermartingale with respect to $P_x^{\tilde{f}}$. Let τ_0 be the time of the first visit to $S_{W_{\lambda_0}}$, i.e.,

$$\tau_0 = \min\{n \in \mathbb{N} \mid X_n \in S_{W_{\lambda_0}}\},$$

so that τ_0 is a stopping time with respect to the filtration $\{\mathcal{F}_t\}$ defined in (2.2), i.e., $[\tau_0 = k] \in \mathcal{F}_k$ for all $k \in \mathbb{N}$. On the other hand, we have that

$$P_x^{\tilde{f}}[\tau_0 < \infty] = 1,$$

by the Assumption 2.2 and Lemma 3.1. Then, using the fact that the function $U_\lambda(W_{\lambda_0}(\cdot)) - U_\lambda(W_{\lambda_1}(\cdot))$ is bounded, the optional sampling theorem leads to

$$U_\lambda(W_{\lambda_0}(x)) - U_\lambda(W_{\lambda_1}(x)) \geq E_x^{\tilde{f}}\left[U_\lambda(W_{\lambda_0}(X_{\tau_0})) - U_\lambda(W_{\lambda_1}(X_{\tau_0}))\right],\ x \in S.$$

Finally, given that $X_{\tau_0} \in S_{W_{\lambda_0}}$ on the event $[\tau_0 < \infty]$, it follows that

$$U_\lambda(W_{\lambda_0}(x)) - U_\lambda(W_{\lambda_1}(x)) \geq 0\,,\ x \in S,$$

for (3.13). So (3.12) is obtained by using that $U_\lambda(\cdot)$ is strictly increasing. □

Throughout the rest of this manuscript, W_λ^* will denote the unique fixed point of T_λ.

4. Main result

In this section, we present the main result of the paper, which establishes the existence of a Nash equilibrium for the game under the total risk sensitivity criterion. We begin by defining the pair of strategies for Players I and II that form a Nash equilibrium for the game.

The strategies for Players I and II that constitute a Nash equilibrium are defined using the unique fixed point W_λ^*. To do this, we define the subset S^* of the state space as

$$S^* := \{x \in S \mid W_\lambda^*(x) = G(x)\}, \tag{4.1}$$

and let τ^* be the time of the first visit to S^*. In other words,

$$\tau^* := \min\{n \in \mathbb{N} \mid X_n \in S^*\}. \tag{4.2}$$

Therefore, τ^* is a stopping time with respect to the filtration $\{\mathcal{F}_t\}$. This means that τ^* belongs to the space $\mathcal{T}$ of admissible strategies for Player II. Based on the Remark 3.1, it can be concluded that there exists a policy $f^* \in \mathbb{F}$ such that, for all $x \in S$,

$$\sum_{y\in S} p_{x,y}(f^*(x))U_\lambda(R(x, f^*(x)) + W_\lambda^*(y)) = \sup_{a\in A(x)}\left[\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a) + W_\lambda^*(y))\right], \tag{4.3}$$

which is a key aspect in the choice of strategy of Player I. The next step is to demonstrate that the pair $(f^*, \tau^*) \in \mathcal{P} \times \mathcal{T}$ constitutes a Nash equilibrium.

Remark 4.1. *Lemma 3.1 ensures that $S^* \neq \emptyset$. Additionally, since the Markov chain associated with each $f \in \mathbb{F}$ is communicating and has an invariant distribution, we have that the set S^* is accessible from any initial state under any stationary policy, i.e.,*

$$P_x^f[\tau^* < \infty] = 1, \quad x \in S, \quad f \in \mathbb{F}. \tag{4.4}$$

Moreover,

$$V_\lambda(x, f, \tau^*) < \infty, \quad x \in S, \quad f \in \mathbb{F}. \tag{4.5}$$

The property (4.4) can be extended to the class of all policies for Player I, mirroring the approach undertaken in [27] (see Lemma 5.1). Therefore, we have

$$P_x^\pi[\tau^* < \infty] = 1, \quad x \in S, \quad \pi \in \mathcal{P}. \tag{4.6}$$

The following lemma provides an auxiliary result for proving the main theorem. A detailed proof can be found in [27].

Lemma 4.1. (*i*) *Given $x \in S$, let $f \in \mathbb{F}$ and $\tau \in \mathcal{T}$ be such that*

$$P_x^f[\tau < \infty] = 1 \quad \textit{and} \quad V_\lambda(x; f, \tau) < \infty.$$

In this case

$$\lim_{n\to\infty} E_x^f\left[\left|U_\lambda\left(\sum_{k=0}^{n} R(X_t, A_t)\right)\right| I[\tau > n+1]\right] = 0.$$

(*ii*) *For every* $n \in \mathbb{N}$, $x \in S$ *and* $\tau \in \mathcal{T}$,

$$\begin{aligned} U_\lambda(W_\lambda^*(x)) &\leq \sum_{k=0}^{n} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau = k]\right] \\ &+ E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau \geq n+1]\right]. \end{aligned} \tag{4.7}$$

(*iii*) *If* $P_x^\pi[\tau^* < \infty] = 1$, *it follows that for every* $x \in S$,

$$V_\lambda(x; \pi, \tau^*) \leq W_\lambda^*(x), \quad \pi \in \mathcal{P}.$$

The following theorem is the main result of the paper.

Theorem 4.1. *Under Assumptions 2.1 and 2.2, the following statements* (*i*) *and* (*ii*) *hold.*

(*i*) *For every* $x \in S$,

$$V_\lambda(x; f^*, \tau^*) = W_\lambda^*(x).$$

(*ii*) *The pair* $(f^*, \tau^*) \in \mathcal{P} \times \mathcal{T}$ *is a Nash equilibrium.*

Proof. (*i*) Assume that $x \in S^*$, so that (4.1) and (4.2) lead to

$$W_\lambda^*(x) = G(x) \quad \text{and} \quad P_x^{f^*}[\tau^* = 0] = 1,$$

while (2.1) and (2.5) lead to $V_\lambda(x; f^*, \tau^*) = G(x)$. Hence, it follows that the value function of the game coincides with the fixed point W_λ^*.

Now, we will demonstrate that the following equality holds for all $n \in \mathbb{N} \setminus \{0\}$ and $x \in S \setminus S^*$:

$$\begin{aligned} U_\lambda(W_\lambda^*(x)) = \sum_{k=1}^{n} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau^* = k]\right] \\ + E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n]\right]. \end{aligned} \tag{4.8}$$

The proof is by induction. First, we observe that $U_\lambda(W_\lambda^*(x)) < U_\lambda(G(x))$ if $x \notin S^*$, by (3.7) and (4.1), and then it holds that

$$\begin{aligned} U_\lambda(W_\lambda^*(x)) &= \sup_{a \in A(x)} \sum_{y \in S} p_{x,y}(a) U_\lambda\left(R(x, a) + W_\lambda^*(y)\right) \\ &= \sum_{y \in S} p_{x,y}(f^*(x)) U_\lambda\left(R(x, f^*(x)) + W_\lambda^*(y)\right) \\ &= E_x^{f^*}\left[U_\lambda\left(R(X_0, A_0) + W_\lambda^*(X_1)\right)\right], \quad x \in S \setminus S^*. \end{aligned} \tag{4.9}$$

Since $P_x^{f^*}[\tau^* > 0] = 1$, by (4.2), it follows that

$$U_\lambda(W_\lambda^*(x)) = E_x^{f^*}\left[U_\lambda\left(R(X_0, A_0) + W_\lambda^*(X_1)\right) I[\tau^* = 1]\right] + E_x^{f^*}\left[U_\lambda\left(R(X_0, A_0) + W_\lambda^*(X_1)\right) I[\tau^* > 1]\right],$$

an expression equivalent to (4.8) with $n = 1$. On the other hand, using the fact that $X_t \notin S^*$ for $0 \leq t < \tau^*$, by (4.2), the equality in (4.9), and the Markov property, it follows that for every $n \in \mathbb{N}$, the following relation holds almost surely with respect to P_x^π

$$U_\lambda(W_\lambda^*(X_n)) = E_x^{f^*}\left[U_\lambda\left(R(X_n, A_n) + W_\lambda^*(X_{n+1})\right)\middle|\mathcal{F}_n, A_n\right] \text{ on } [\tau^* > n].$$

Multiplying both sides of this inequality by $e^{\lambda\sum_{t=0}^{n-1} R(X_t,A_t)} I[\tau^* > n]$, which is an $\mathcal{F}_n$-measurable random variable, an application of (2.4) leads us to

$$U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n] = E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau^* > n]\middle|\mathcal{F}_n, A_n\right].$$

Now taking the expectation with respect to $P_x^{f^*}$ and using the equality $I[\tau^* > n] = I[\tau^* = n+1] + I[\tau^* > n+1]$, we have that

$$\begin{aligned} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n]\right] &= E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau^* = n+1]\right] \\ &+ E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau^* > n+1]\right]. \end{aligned}$$

Then, combining this equality with the induction hypothesis, it follows that (4.8) is valid with $n+1$ instead of n. Moreover, leveraging the property that $U_\lambda(\cdot)$ maintains a constant sign, the monotone convergence theorem yields the following result:

$$\begin{aligned} \lim_{n\to\infty}\sum_{k=1}^{n} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau^* = k]\right] &= \sum_{k=1}^{\infty} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau^* = k]\right] \\ &= E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{\tau^*-1} R(X_t, A_t) + W_\lambda^*(X_{\tau^*})\right) I[\tau^* < \infty]\right] \\ &= U_\lambda(V_\lambda(x; f^*, \tau^*)), \end{aligned}$$

where the last equality follows from the combination of (2.5) and (4.4). Additionally, it follows from Lemma 4.1(i) that (4.4) and (4.5) imply that

$$\lim_{n\to\infty} E_x^{f^*}\left[\left|U_\lambda\left(\sum_{k=0}^{n} R(X_t, A_t)\right)\right| I[\tau^* > n+1]\right] = 0.$$

Taking the limit as n goes to infinity on the right side of (4.8), the last two convergences together imply that $U_\lambda(W_\lambda^*(x)) = U_\lambda(V_\lambda(x; f^*, \tau^*))$. Starting from this equality and using the fact that U_λ is strictly increasing, we get $V_\lambda(x; f^*, \tau^*) = W_\lambda^*(x)$, with $x \in S \setminus \{S^*\}$.

(ii) Since the value function of the game coincides with the fixed point W_λ^* of the operator T_λ, to prove that the pair (f^*, τ^*) is a Nash equilibrium we have to prove the following inequalities:

$$V_\lambda(x; \pi, \tau^*) \leq W_\lambda^*(x) \leq V_\lambda(x; f^*, \tau), \quad \pi \in \mathcal{P}, \quad \tau \in \mathcal{T},$$

according to Definition 2.2. Using Lemma 4.1(*iii*) and (4.6), we can confirm that the first inequality is indeed satisfied. To prove the second inequality:

$$W_\lambda^*(x) \le V_\lambda(x; f^*, \tau), \tag{4.10}$$

we consider the following cases for the pair (x, τ), where x is an arbitrary element in S:

- **Case 1.**

$$P_x^{f^*}[\tau < \infty] = 1. \tag{4.11}$$

In the following argument, we assume that

$$V_\lambda(\cdot; f^*, \tau) < \infty, \tag{4.12}$$

since (4.10) certainly holds if $V_\lambda(\cdot; f^*, \tau) = \infty$. Note that (2.4) and the inclusion $W_\lambda^* \in [\![0, G]\!]$ together yield that

$$\left| U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right)\right| = \left| e^{\lambda W_\lambda^*(X_{n+1})} U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t)\right)\right| \le e^{|\lambda| \|G\|} \left| U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t)\right)\right|.$$

Also, by Lemma 4.1(*i*), (4.11) and (4.12) together imply that

$$\lim_{n\to\infty} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t)\right) I[\tau > n+1]\right] = 0,$$

and combining this convergence with the previous display, it follows that

$$E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau > n+1]\right] \to 0 \ \text{ as } n \to \infty.$$

On the other hand, since $U_\lambda(\cdot)$ has a constant sign, the monotone convergence theorem immediately yields that

$$\begin{aligned}
\lim_{n\to\infty} \sum_{k=0}^{n} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau = k]\right] &= \sum_{k=0}^{\infty} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau = k]\right] \\
&= E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + W_\lambda^*(X_\tau)\right) I[\tau < \infty]\right] \\
&\le E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)\right) I[\tau < \infty]\right] \\
&= U_\lambda(V_\lambda(x, f^*, \tau)),
\end{aligned}$$

where the inequality is due to the inclusion $W_\lambda^* \in [\![0, G]\!]$ and the monotonicity of $U_\lambda(\cdot)$, and using (4.11), the last equality is due to (2.3) and (2.5). From Lemma 4.1(*ii*), taking the limit as n goes to ∞ in the right-hand side of (4.7), the two previous displays yield that $U_\lambda(W_\lambda^*(x)) \le U_\lambda(V_\lambda(x, f^*, \tau))$ and then (4.10) follows from the fact that $U_\lambda(\cdot)$ is strictly increasing.

- **Case 2.**

$$P_x^{f^*}[\tau = \infty] > 0.$$

Let z be the state as in Assumption 2.1(iv), and note that the communication property of Assumption 2.2 leads to

$$P_x^{f^*}[X_n = z \ \ i.o.] = 1,$$

where $i.o.$ means infinitely often. Now, given that R is non-negative and that $R(z, f^*(z)) > 0$, it follows that

$$P_x^{f^*}\left[\sum_{n=0}^{\infty} R(X_n, A_n) = \infty\right] = 1,$$

and since the event $[\tau = \infty]$ has positive probability, it follows that

$$\begin{aligned} V_\lambda(x; f^*, \tau) &= \frac{1}{\lambda}\log\left(E_x^{f^*}\left[e^{\lambda\left(\sum_{t=0}^{\tau-1} R(X_t,A_t)+G(X_\tau)I[\tau<\infty]\right)}\right]\right) \\ &= \frac{1}{\lambda}\log\left(E_x^{f^*}\left[e^{\lambda\left(\left(\sum_{t=0}^{\tau-1} R(X_t,A_t)+G(X_\tau)\right)I[\tau<\infty]+\sum_{t=0}^{\infty} R(X_t,A_t)I[\tau=\infty]\right)}\right]\right) \\ &\geq \frac{1}{\lambda}\log\left(E_x^{f^*}\left[e^{\lambda\sum_{t=0}^{\infty} R(X_t,A_t)I[\tau=\infty]}\right]\right) \\ &= \infty. \end{aligned}$$

Then, the inequality in (4.10) holds in this case as well. The pair $(f^*, \tau^*) \in \mathcal{P} \times \mathcal{T}$ is therefore a Nash equilibrium. □

5. A numerical example

In this section, we present a numerical example that illustrates a method for identifying the fixed point of the operator T_λ and, subsequently, the strategy that constitutes a Nash equilibrium. For this purpose, we consider Example 2.1 and introduce Algorithm 1, which details the steps for calculating the fixed point W_λ^*.

We implemented Algorithm 1 in MATLAB, and the numerical results of the experiment are presented in Tables 1 and 2.

It is evident that both the number of iterations and the size of the set S^* increase with N. Additionally, there is a notable discrepancy between the number of λ, and the size of S^* varies significantly with changes in λ (see Figure 1). The results remain consistent as the size of $\hat{S}$ increases with fixed values of N and λ. Regarding the strategy f^*, it was observed that for positive values of λ, f^* generally takes on two values: the minimum and maximum of the action space. In contrast, for negative values of λ, f^* practically remains constant, adopting the minimum value of the action space.

Algorithm 1 Method for finding the Nash equilibrium in Example 1.

Require: $\lambda \neq 0$, $\{b_1, b_2, \ldots, b_N\}$, $S = \{1, 2, \ldots, \hat{S}\}$, with $\hat{S} \in \mathbb{N}$, $G(x)$, $R(x, a)$, ϵ.
Ensure: Iter, W_λ^*, f^*, S^*.

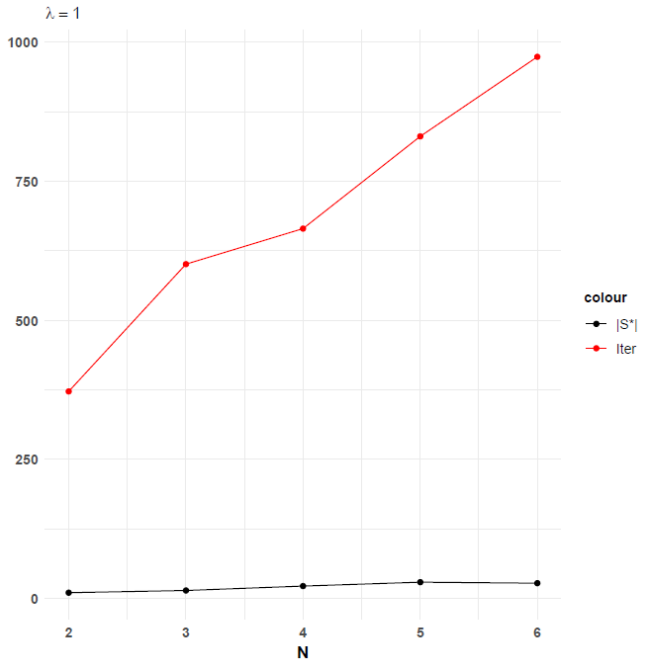
$W \leftarrow \mathbb{O}$, $\hat{W} \leftarrow \mathbb{1}$, $s \leftarrow \mathbb{O}$ (where $\mathbb{O}$ and $\mathbb{1}$ denote arrays of all zeros and all ones, respectively),
Iter $\leftarrow 0$, norm $\leftarrow \|\hat{W} - W\|$, $m \leftarrow 0$.
while norm $> \epsilon$ **do**
  for $l = 1 : N$ **do**
    $s(l) = U_\lambda(R(0, l) + W(1))$.
  end for
  $m = \min\{U_\lambda(G(0)), \max(s)\}$.
  $\hat{W}(0) = \log(m/sign(\lambda))/\lambda$.
  for $k = 1 : \hat{S} - 1$ **do**
    for $l = 1 : N$ **do**
      $s(l) = b(l) \cdot U_\lambda(R(k, l) + W(k + 1)) + (1 - b(l)) \cdot U_\lambda(R(k, l) + W(k - 1))$.
    end for
    $m = \min\{U_\lambda(G(k)), \max(s)\}$.
    $\hat{W}(k) = \log(m/sign(\lambda))/\lambda$.
  end for
  for $l = 1 : N$ **do**
    $s(l) = U_\lambda(R(\hat{S}, l) + W(\hat{S} - 1))$.
  end for
  $m = \min\{U_\lambda(G(\hat{S})), \max(s)\}$.
  $\hat{W}(\hat{S}) = \log(m/sign(\lambda))/\lambda$.
  norm $= \|\hat{W} - W\|$.
  $W \leftarrow \hat{W}$.
  Iter $\leftarrow$ Iter+1.
end while
$W_\lambda^* = \hat{W}$.
Compute f^* and S^* according to (4.3) and (4.1), respectively.

Table 1. Numerical performance of Algorithm 1 for different values of N, with $\hat{S}$ and λ fixed.

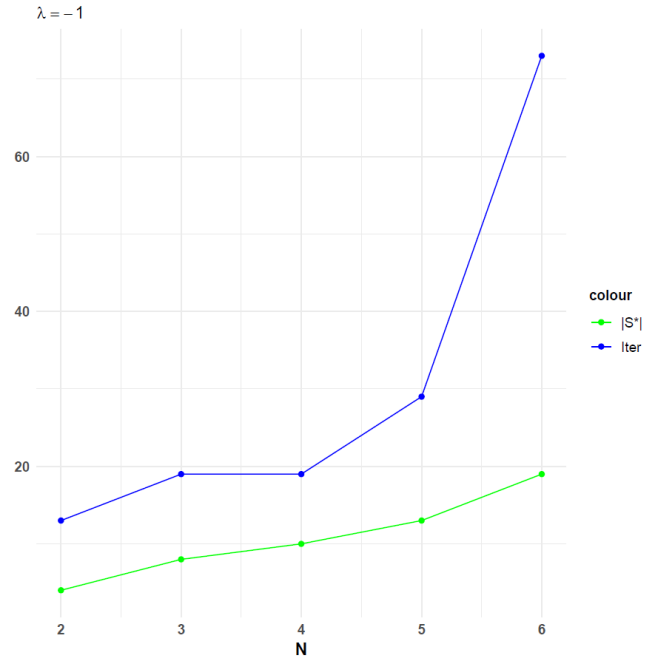
	N	2	3	4	5	6
$\hat{S}$=100000	Iter	372	601	665	831	974
λ=1	$\lvert S^*\rvert$	10	14	22	29	27
$\hat{S}$=100000	Iter	13	19	19	29	73
λ=-1	$\lvert S^*\rvert$	4	8	10	13	19

Table 2. Numerical performance of Algorithm 1 for different values of λ, with $\hat{S}$ and N fixed.

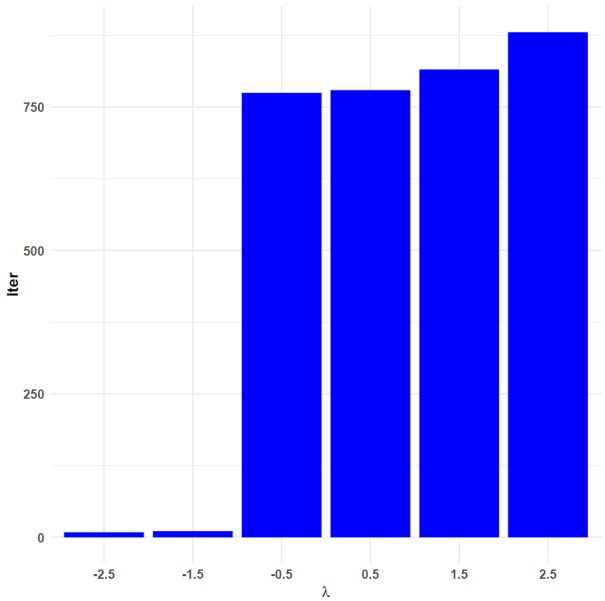
	λ	1/2	-1/2	3/2	-3/2	5/2	-5/2
$\hat{S}$=100000	Iter	780	775	816	11	881	9
N=4	$\lvert S^*\rvert$	10	10	28	5	54	7



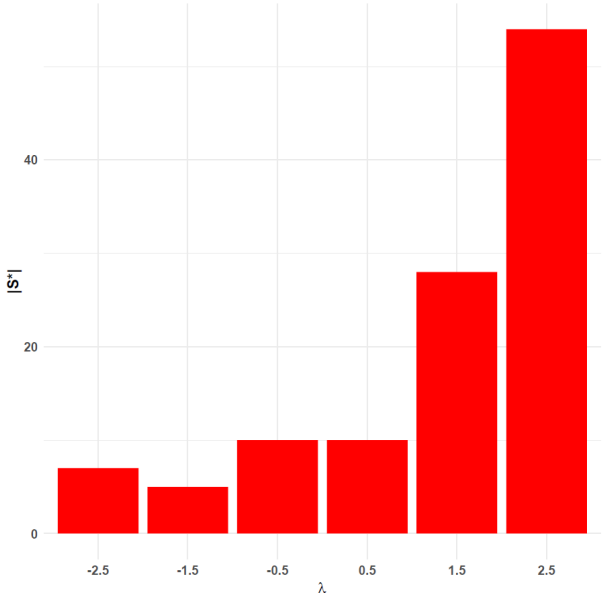
(a) The relationship between the number of iterations and the set size S^* as a function of N, with $\lambda = 1$.



(b) The relationship between the number of iterations and the set size S^* as a function of N, with $\lambda = -1$.



(c) Number of iterations for each value of λ.



(d) Size of the set S^* for each value of λ.

Figure 1. Numerical results from the implementation of Algorithm 1 in Example 1.

6. Conclusions

In this note, Markov stopping games with bounded rewards and risk-sensitive total reward criteria were studied. The communication-ergodicity properties allowed us first to prove that the set S^* is non-empty, a crucial outcome in demonstrating the uniqueness of the fixed point W_λ^*. Subsequently, the strategy of the players that constitute a Nash equilibrium was derived from this fixed point. Additionally, it was demonstrated that the value function of the game coincides with the fixed point W_λ^*, as indicated in Theorem 4.1, which represents the main result of this paper. Furthermore, we utilized certain properties of the operator T_λ and inequalities involving W_λ^*, which were established in [27]. It is important to note that an extension to more general cases with respect to the state space is a complicated task, since as is known from the literature on risk-sensitive PDMs, it is not always possible. An example of this can be illustrated in the following situation. In 1972 [15], Howard and Matheson showed that optimal risk-sensitive average cost is determined via an optimality equation in finite and communicating models. Forty years later, it was shown in [34] that the seminal result by Howard and Matheson cannot be extended to the case of a denumerable state space. Therefore, an interesting future problem is to investigate the feasible extension of the results presented in this manuscript to more general spaces, such as Borel spaces, and to consider the option of incorporating possible unbounded rewards.

Author contributions

Jaicer López-Rivero, Hugo Cruz-Suárez and Carlos Camilo-Garay jointly contributed to the conceptualization and methodology. Jaicer López-Rivero drafted the original manuscript with contributions from Hugo Cruz-Suárez and Carlos Camilo-Garay. Jaicer López-Rivero, Hugo Cruz-Suárez and Carlos Camilo-Garay were involved in the review and editing of the manuscript. All authors have accepted full responsibility for the content of this manuscript, consented to its submission to the journal, reviewed all the results, and approved the final version.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are deeply grateful to the reviewers and the Editor for their careful reading of the original manuscript and for the advice to improve the paper. This works was partially supported by CONAHCyT-México under Grant No. CF-2023-I-1362. The first author would like to express gratitude to CONAHCyT-México for providing financial support through a fellowship.

Conflict of interest

The authors declare that they have no potential conflicts of interest.

References

1. A. Maitra, W. Sudderth, The gambler and the stopper, *Lecture Notes-Monograph Series*, **30** (1996), 191–208.
2. E. Dynkin, Game variant of a problem on optimal stopping, *Soviet Math. Dokl.*, **10** (1969), 270–274.
3. G. Peskir, A. Shiryaev, *Optimal stopping and free-boundary problems*, Basel: Birkhäuser, 2006. http://dx.doi.org/10.1007/978-3-7643-7390-0
4. A. Shiryaev, *Optimal stopping rules*, Berlin: Springer Science & Business Media, 2007. http://dx.doi.org/10.1007/978-3-540-74011-7
5. T. Bielecki, D. Hernández-Hernández, S. Pliska, Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management, *Math. Meth. Oper. Res.*, **50** (1999), 167–188. http://dx.doi.org/10.1007/s001860050094
6. G. Peskir, On the American option problem, *Math. Finance*, **15** (2005), 169–181. http://dx.doi.org/10.1111/j.0960-1627.2005.00214.x
7. E. Altman, A. Shwartz, Constrained Markov games: Nash equilibria, *Proceedings of Advances in Dynamic Games and Applications*, 2000, 213–221. http://dx.doi.org/10.1007/978-1-4612-1336-9_11
8. R. Atar, A. Budhiraja, A stochastic differential game for the inhomogeneous Laplace equation, *Ann. Prob.*, **38** (2010), 498–531. http://dx.doi.org/10.1214/09-AOP494
9. J. Filar, K. Vrieze, *Competitive Markov decision processes*, New York: Springer Science & Business Media, 2012. http://dx.doi.org/10.1007/978-1-4612-4054-9
10. V. Kolokoltsov, O. Malafeyev, *Understanding game theory: introduction to the analysis of many agent systems with competition and cooperation*, Hackensack: World Scientific, 2020.
11. L. Shapley, Stochastic games, *PNAS*, **39** (1953), 1095–1100. http://dx.doi.org/10.1073/pnas.39.10.1095
12. L. Zachrisson, Markov games, In: *Advances in game theory*, Princeton: Princeton University Press, 1964, 211–254. http://dx.doi.org/10.1515/9781400882014-014
13. O. Hernández-Lerma, *Adaptive Markov control processes*, New York: Springer Science & Business Media, 2012. http://dx.doi.org/10.1007/978-1-4419-8714-3
14. M. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, Hoboken: John Wiley & Sons, 2014. http://dx.doi.org/10.1002/9780470316887
15. R. Howard, J. Matheson, Risk-sensitive Markov decision processes, *Manage. Sci.*, **18** (1972), 356–369. http://dx.doi.org/10.1287/mnsc.18.7.356
16. N. Bäuerle, U. Rieder, *Markov decision processes with applications to finance*, Heidelberg: Springer Science & Business Media, 2011. http://dx.doi.org/10.1007/978-3-642-18324-9
17. L. Stettner, Risk sensitive portfolio optimization, *Math. Meth. Oper. Res.*, **50** (1999), 463–474. http://dx.doi.org/10.1007/s001860050081
18. S. Balaji, S. Meyn, Multiplicative ergodicity and large deviations for an irreducible Markov chain, *Stoch. Proc. Appl.*, **90** (2000), 123–144. http://dx.doi.org/10.1016/S0304-4149(00)00032-6
19. I. Kontoyiannis, S. Meyn, Spectral theory and limit theorems for geometrically ergodic Markov processes, *Ann. Appl. Probab.*, **13** (2003), 304–362. http://dx.doi.org/10.1214/aoap/1042765670

20. N. Bäuerle, U. Rieder, More risk-sensitive Markov decision processes, *Math. Oper. Res.*, **39** (2014), 105–120. http://dx.doi.org/10.1287/moor.2013.0601
21. V. Borkar, S. Meyn, Risk-sensitive optimal control for Markov decision processes with monotone cost, *Math. Oper. Res.*, **27** (2002), 192–209. http://dx.doi.org/10.1287/moor.27.1.192.334
22. K. Sladkỳ, Risk-sensitive average optimality in Markov decision processes, *Kybernetika*, **54** (2018), 1218–1230. http://dx.doi.org/10.14736/kyb-2018-6-1218
23. G. Di Masi, Ł. Stettner, Infinite horizon risk sensitive control of discrete time Markov processes under minorization property, *SIAM J. Control Optim.*, **46** (2007), 231–252. http://dx.doi.org/10.1137/040618631
24. A. Jaśkiewicz, Average optimality for risk-sensitive control with general state space, *Ann. Appl. Probab.*, **17** (2007), 654–675. http://dx.doi.org/10.1214/105051606000000790
25. R. Cavazos-Cadena, L. Rodríguez-Gutiérrez, D. Sánchez-Guillermo, Markov stopping games with an absorbing state and total reward criterion, *Kybernetika*, **57** (2021), 474–492. http://dx.doi.org/10.14736/kyb-2021-3-0474
26. V. Martínez-Cortés, Bi-personal stochastic transient Markov games with stopping times and total reward criterion, *Kybernetika*, **57** (2021), 1–14. http://dx.doi.org/10.14736/kyb-2021-1-0001
27. J. López-Rivero, R. Cavazos-Cadena, H. Cruz-Suárez, Risk-sensitive Markov stopping games with an absorbing state, *Kybernetika*, **58** (2022), 101–122. http://dx.doi.org/10.14736/kyb-2022-1-0101
28. M. Torres-Gomar, R. Cavazos-Cadena, H. Cruz-Suárez, Denumerable Markov stopping games with risk-sensitive total reward criterion, *Kybernetika*, **60** (2024), 1–18. http://dx.doi.org/10.14736/kyb-2024-1-0001
29. W. Zhang, C. Liu, Discrete-time stopping games with risk-sensitive discounted cost criterion, *Math. Meth. Oper. Res.*, in press. http://dx.doi.org/10.1007/s00186-024-00864-1
30. F. Dufour, T. Prieto-Rumeau, Nash equilibria for total expected reward absorbing Markov games: the constrained and unconstrained cases, *Appl. Math. Optim.*, **89** (2024), 34. http://dx.doi.org/10.1007/s00245-023-10095-1
31. R. Cavazos-Cadena, M. Cantú-Sifuentes, I. Cerda-Delgado, Nash equilibria in a class of Markov stopping games with total reward criterion, *Math. Meth. Oper. Res.*, **94** (2021), 319–340. http://dx.doi.org/10.1007/s00186-021-00759-5
32. J. Saucedo-Zul, R. Cavazos-Cadena, H. Cruz-Suárez, A discounted approach in communicating average Markov decision chains under risk-aversion, *J. Optim. Theory Appl.*, **187** (2020), 585–606. http://dx.doi.org/10.1007/s10957-020-01758-y
33. P. Hoel, S. Port, C. Stone, *Introduction to stochastic processes*, Long Grove: Waveland Press, 1986.
34. R. Cavazos-Cadena, Characterization of the optimal risk-sensitive average cost in denumerable Markov decision chains, *Math. Oper. Res.*, **43** (2018), 1025–1050. http://dx.doi.org/10.1287/moor.2017.0893