Jaicer López-Rivero; Rolando Cavazos-Cadena; Hugo Cruz-Suárez
Risk-sensitive Markov stopping games with an absorbing state

Kybernetika, Vol. 58 (2022), No. 1, 101–122

Persistent URL: `http://dml.cz/dmlcz/149604`

KYBERNETIKA — VOLUME *58* (2022), NUMBER 1, PAGES 101–122

RISK-SENSITIVE MARKOV STOPPING GAMES WITH AN ABSORBING STATE

Jaicer López-Rivero, Rolando Cavazos-Cadena and Hugo Cruz-Suárez

This work is concerned with discrete-time Markov stopping games with two players. At each decision time player II can stop the game paying a terminal reward to player I, or can let the system to continue its evolution. In this latter case player I applies an action affecting the transitions and entitling him to receive a running reward from player II. It is supposed that player I has a no-null and constant 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 (expected) total reward of player I and, besides mild continuity-compactness conditions, the main structural assumption on the model is the existence of an absorbing state which is accessible from any starting point. In this context, it is shown that the value function of the game is characterized by an equilibrium equation, and the existence of a Nash equilibrium is established.

Keywords: monotone operator, fixed point, equilibrium equation, hitting time, bounded rewards, certainty equivalent

Classification: 93E20, 93C55, 60J05

1. INTRODUCTION

In this paper our attention is focused on a class of zero-sum games in discrete time, countable state space, and Markovian transitions. The game is driven by two players, and at each decision time player II always has two options, namely, to stop the game paying a terminal reward to player I, or else, to let the system to continue its evolution, and in this case player I applies an action affecting the system transition and entitling him to receive a running reward from player II. A random reward is evaluated by player I via an exponential utility function with non-null and constant risk-sensitivity, and the performance of a pair of strategies is measured by the risk-sensitive total reward received by player I. Whereas player I chooses his decision strategy trying to maximize his utility, it is supposed that the objective of player II is to minimize the utility of player I. The main structural condition on the game is the existence of an absorbing state which, regardless of the strategies of the players, will be eventually reached from any initial state, see Assumption 2.3 below. Within this framework, the main problems studied below can be stated as follows:

DOI: 10.14736/kyb-2022-1-0101

- To characterize the value function of the game via en equilibrium equation, and
- To determine a Nash equilibrium.

Game theory has interesting applications in various areas, some of which can be found in the articles by Altman, Shwartz [2], Atar and Budhiraja [3], and in the books by Filar and Vrtieze [17] and Kolokoltsov and Malafeyev [21]. While the theory of Markov games can be traced back to the articles by Shapley [28] and Zachrisson [33] and recent advances and applications can be found, for instance, in [17] or [21]. On the other hand, the idea of stopping times is of great relevance in stochastic analysis, and a complete description of the theory can be found in [29] and [25], in addition to which applications to mathematical finance can be found in [8], and in [24]. In general, Markov decision processes can be viewed as single-player stochastic games. A well-established theory of Markov decision chains is available in [27] and [18], and applications can be found, for example, in [6] or [32], where risk sensitive criterion is presented.

The analysis of discrete-time Markov models with the risk-sensitive criteria can be traced back, at least, to [19], where controlled Markov chains on a finite state space were analyzed and, under appropriate communication conditions, a characterization of the optimal *risk-sensitive* average cost was derived in terms of an optimality equation. More recently, work on risk-sensitive criteria has been stimulated by connections with the fields of mathematical finance [5, 32, 26], and large deviations [4, 22]. For uncontrolled and controlled models, the results by Howard and Matheson (1972) [19] have been extended to the case of a general transition structure in [10] and [1], respectively. Controlled Markov models with finite or denumerable state-space endowed with risk-sensitive criteria have been studied, for instance, in [6, 9, 13, 30, 31], whereas Markov decision processes on a general state space are analyzed, for instance, in [14, 15, 16] or [20]. On the other hand, game theory has been intensively studied and applications can be found, for instance, in [2, 17, 21] or [7]. A comprehensive account of the idea of stopping time can be found in [29] and [25]. Finally, the theory on controlled Markov processes used in this note is well established and is presented, for instance, in [27] and [18].

The approach of this work combines basic ideas about stopping times, Markov chains and dynamic programming, with the analysis of a *monotone operator* introduced in Section 3. It is shown that such an operator has a fixed point which is used to define strategies for players I and II, and then it is shown that those strategies conform a Nash equilibrium. *The organization* of the subsequent material is as follows: In Section 2 the idea of Markov stopping game is formally introduced, the exponential utility with constant risk-sensitivity coefficient and the corresponding risk-sensitive total reward criterion are formulated, and the idea of Nash equilibrium is discussed, whereas in Section 3 the fundamental operator considered in this work is introduced, and its main property, namely, the existence of a fixed point, is stated in Theorem 3.2. Such a result is used to define strategies for players I and II which, as it is shown in Theorem 3.3, constitute a Nash equilibrium. Then, Section 4 is dedicated to establish Theorem 3.2 using an argument that relies on two properties of the basic operator, namely, *monotonicity*, and *continuity* with respect to the topology of pointwise convergence. Next, Section 5 contains two technical results that will be used to verify the existence of a Nash equilibrium in Section 6, and the paper concludes with some brief comments in Section 7.

2. THE MODEL

In this section the dynamic model studied in the paper is formally described but, before going any further, it is convenient to introduce the basic notation used in the forthcoming analysis. Given a topological space $\mathbb{K}$, the Banach space $\mathcal{C}(\mathbb{K})$ consist of all continuous functions $R : \mathbb{K} \to \mathbb{R}$ whose supremum norm $\|R\|$ is finite, where $\|R\| := \sup_{k\in\mathbb{K}} |R(k)|$, whereas $\mathbb{N}$ stands for the set of nonnegative integers. The indicator function of an event A is denoted by $I[A]$ and, even without explicit mention, all relations involving conditional expectations are valid with probability 1 with respect to the underlying probability measure. On the other hand, $a \wedge b$ and $a \vee b$ are used as infix notations for $\min\{a, b\}$ and $\max\{a, b\}$, respectively, where $a, b \in \mathbb{R}$. The minimum of the empty set is ∞ and, finally, the following convention concerning summations will be used:

$$\sum_{t=n}^{m} a_t := 0, \quad m < n. \tag{1}$$

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, who are referred to as players I and II. The components of $\mathcal{G}$ have the following meaning: The (nonempty and) denumerable set S is the state space and is endowed with the discrete topology, the metric space A is the action set and, for each $x \in S$, $A(x) \subset A$ is the nonempty class of admissible actions at x for player I. On the other hand, $R \in \mathcal{C}(\mathbb{K})$ is the running reward function, where the class $\mathbb{K}$ of admissible pairs is defined by $\mathbb{K} := \{(x, a) : a \in A(x), x \in S\}$, and $G \in \mathcal{C}(S)$ is the terminal reward. Finally, $P = [p_{x,y}(a)]$ 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 $(x, a) \in \mathbb{K}$. Model $\mathcal{G}$ is interpreted as follows: At each decision epoch $t \in \mathbb{N}$, players I and II observe the state of the system, say $X_t = x \in S$, and player II must select one of two actions: To *stop* the system paying a terminal reward $G(x)$ to player I, or let the system *to continue* its evolution. In this latter case, using the record of states up to time t and actions previous to t, player I applies an action (control) $A_t = a \in A(x)$, an intervention that has two consequences: player I gets a reward $R(x, a)$ from player II and, regardless of the previous states and actions, the system moves to $X_{t+1} = y \in S$ with probability $p_{x,y}(a)$; this is the Markov property of the decision process. The following conditions will be enforced throughout the remainder.

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 each $x \in S$ and $a \in A(x)$, $G(x) \geq 0$ and $R(x, a) \geq 0$.

Decision Strategies. For each $t \in \mathbb{N}$ the space $\mathbb{H}_t$ of possible histories up to time t is defined by $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K}^t \times S$ when $t > 0$, whereas $h_t = (x_0, a_0, \ldots, x_i, a_i, \ldots, x_t)$ stands for 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: For each $t \in \mathbb{N}$ and $h_t \in \mathbb{H}_t$, $\pi_t(\cdot|h_t)$ is a probability measure on A concentrated on $A(x_t)$, 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 $\{(X_t, A_t)\}$ [18, 27], and 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 is *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, $\mathbb{F} \subset \mathcal{P}$. On the other hand, setting

$$\mathcal{F}_t := \sigma(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t), \tag{2}$$

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 $t \in \mathbb{N}$.

Exponential Utility. Throughout the remainder it is supposed that player I has a constant risk-sensitivity coefficient $\lambda \neq 0$, so that a random reward Y is assessed via the expectation of $U_\lambda(Y)$, where the utility function $U_\lambda : \mathbb{R} \to \mathbb{R}$ is given by

$$U_\lambda(x) := \operatorname{sign}(\lambda) e^{\lambda x}, \quad x \in S; \tag{3}$$

notice that $U_\lambda(\cdot)$ is a strictly increasing function and that

$$U_\lambda(x + y) = e^{\lambda x} U_\lambda(y), \quad x, y \in \mathbb{R}. \tag{4}$$

When choosing between two random rewards W and Y, player I prefers Y if $E[U_\lambda(W)] < E[U_\lambda(Y)]$, and is indifferent between them 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 $\mathcal{E}_\lambda(Y)$. Observe that

$$\mathcal{E}_\lambda(Y) := \log(E[e^{\lambda Y}])/\lambda. \tag{5}$$

Performance Criterion. Given the initial state $X_0 = x \in S$, suppose that players I and II drive the system using strategies $\pi \in \mathcal{P}$ and $\tau \in \mathcal{T}$, respectively. The total (random) 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 $x \in S$:

$$V_\lambda(x; \pi, \tau) := \frac{1}{\lambda} \log \left(E_x^\pi \left[e^{\lambda (\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau) I[\tau < \infty])} \right] \right); \tag{6}$$

see (5) and observe that, since R and G are nonnegative,

$$V_\lambda(x;\pi,\tau)\geq 0. \tag{7}$$

When player II employs the strategy τ, the largest value of the certainty equivalent that can be achieved by player I is $\sup_{\pi\in\mathcal{P}} V_\lambda(x;\pi,\tau)$, which is a function of x and τ, say $\varphi(x;\tau)$. It is supposed that the *main objective of player* II *is to minimize the expected utility* of player I, so that player II will try hard to employ a stopping time $\tilde{\tau}$ such that $\varphi(x;\tilde{\tau})$ is as close as possible to $\inf_{\tau\in\mathcal{T}}\varphi(x;\tau)$. This last quantity is the (*upper*-)*value* function of the game and is explicitly determined by

$$V_\lambda^*(x):=\inf_{\tau\in\mathcal{T}}\left[\sup_{\pi\in\mathcal{P}} V_\lambda(x;\pi,\tau)\right],\quad x\in S. \tag{8}$$

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

$$V_{\lambda,*}(x):=\sup_{\pi\in\mathcal{P}}\left[\inf_{\tau\in\mathcal{T}} V_\lambda(x;\pi,\tau)\right],\quad x\in S. \tag{9}$$

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 immediately lead to

$$V_\lambda^*(\cdot)\geq V_{\lambda,*}(\cdot). \tag{10}$$

Equilibrium Strategies. The remainder of the paper analyzes the existence of a Nash equilibrium, an idea that is introduced below.

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

$$V_\lambda(x;\pi,\tau^*)\leq V_\lambda(x;\pi^*,\tau^*)\leq V_\lambda(x;\pi^*,\tau),\quad \pi\in\mathcal{P},\quad \tau\in\mathcal{T}. \tag{11}$$

When the strategies π^* and τ^* actually used by players I and II form a Nash equilibrium, it follows from the first inequality in the above display that, if player II keeps on using strategy τ^*, then player I does not have any incentive to switch to other policy. Similarly, the second inequality in (11) implies that, if player I keeps on using π^*, then player II does not have any motivation to change the strategy τ^* in use. Also, note that if (π^*,τ^*) is a Nash equilibrium, then (11) 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 (8) and (9), respectively, so that via (10), it follows that the upper and lower value functions are equal and coincide with $V_\lambda(\cdot;\pi,\tau^*)$.

In [11] the existence of a Nash equilibrium was established for Markov stopping games with the *risk-neutral* discounted criterion. As it was pointed out in [23] the discounted index is a particular case of the total reward criterion applied to models with an absorbing state z satisfying the following two properties: (i) The running and terminal reward are null at z, and (ii) Under any stationary policy, state z is accessible from any initial state. These conditions will be assumed throughout the sequel, and there formally stated as follows.

Assumption 2.3. There exists a state $z \in S$ satisfying conditions (i) and (ii) below.
(i) For every $x \in S$ and $f \in \mathbb{F}$,

$$P_x^f[\tau_z < \infty] = 1, \tag{12}$$

where

$$\tau_z := \min\{n \in \mathbb{N} : X_n = z\}. \tag{13}$$

(ii) $G(z) = 0 = R(z, a)$ and $p_{z,z}(a) = 1, a \in A(z)$.

Notice that τ_z is a stopping time with respect to the filtration $\{\mathcal{F}_n\}$ given in (2), and that

$$X_{\tau_z} = z \text{ on the event } [\tau_z < \infty]. \tag{14}$$

Under Assumptions 2.1 and 2.3, Markov stopping games endowed with the risk-neutral total reward criterion were studied in [23] and [12], where the existence of Nash equilibrium was proved assuming that the state space is finite and denumerable, respectively. Within the framework determined by Assumptions 2.1 and 2.3, the main objective of the paper is to show that there exists a Nash equilibrium with respect to the *risk sensitive* total reward index (6).

3. MAIN THEOREM

In this section the main result of the paper on the existence of a Nash equilibrium is stated. First, a subset of $\mathcal{C}(S)$ and an operator on that set are introduced.

Definition 3.1. (i) The space $[\![0, G]\!] \subset \mathcal{C}(S)$ is defined by

$$[\![0, G]\!] := \{h \in \mathcal{C}(S) \,:\, 0 \leq h(x) \leq G(x)\}. \tag{15}$$

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

$$U_\lambda(T_\lambda[W](x)) := \min\left\{U_\lambda(G(x)), \sup_{a \in A(x)} \sum_{y \in S} p_{x,y}(a) U_\lambda(R(x, a) + W(y))\right\}. \tag{16}$$

Using that $U_\lambda(\cdot)$ is increasing and that R and G are nonnegative, it is not difficult to verify that T_λ actually transforms $[\![0, G]\!]$ into itself, as well as the following two properties:

$$T_\lambda[W](z) = W(z) = 0, \quad W \in [\![0, G]\!], \tag{17}$$

and

$$W, W_1 \in [\![0, G]\!] \text{ and } W \leq W_1 \implies T_\lambda[W] \leq T_\lambda[W_1]. \tag{18}$$

The existence of a Nash equilibrium for the criterion (6), stated below as Theorem 3.2, relies heavily on the following result.

Theorem 3.2. Under Assumptions 2.1 and 2.3 the operator T_λ has a fixed point, that is, there exists a function $W_\lambda^* \in [\![0, G]\!]$ satisfying

$$W_\lambda^* = T_\lambda[W_\lambda^*]. \tag{19}$$

Throughout the remainder W_λ^* stands for a given fixed point of T_λ. Via Definition 3.1, (19) can be equivalently written as follows: For every $x \in S$,

$$U_\lambda(W_\lambda^*(x)) = \min\left\{ U_\lambda(G(x)), \sup_{a\in A(x)} \sum_{y\in S} p_{x,y}(a) U_\lambda(R(x,a) + W_\lambda^*(y)) \right\}. \tag{20}$$

Additionally, using the G is bounded, the inclusion $W_\lambda^* \in [\![0, G]\!]$ and Assumption 2.1 together imply that there exists a policy $f^* \in \mathbb{F}$ such that, for every $x \in S$,

$$\begin{aligned}&\sum_{y\in S} p_{x,y}(f^*(x)) U_\lambda(R(x, f^*(x)) + W_\lambda^*(y)) \\ &\qquad = \sup_{a\in A(x)} \left[\sum_{y\in S} p_{x,y}(a) U_\lambda(R(x,a) + W_\lambda^*(y)) \right].\end{aligned} \tag{21}$$

Also, observing that $W_\lambda^* \geq 0$, Assumption 2.3(ii) and (20) together imply that $U_\lambda(W_\lambda^*(z)) = U_\lambda(G(z)) = U_\lambda(0)$, and then

$$W_\lambda^*(z) = 0 = G(z). \tag{22}$$

Next, define the subset S^* of the state space by

$$S^* := \{x \in S : W_\lambda^*(x) = G(x)\}, \tag{23}$$

and let τ^* be the time of the first visit to S^*, that is,

$$\tau^* := \min\{n \in \mathbb{N} : X_n \in S^*\}, \tag{24}$$

so that τ^* is a stopping time with respect to the filtration $\{\mathcal{F}_t\}$ in (2), that is, τ^* belongs to the space $\mathcal{T}$ of admissible strategies for player II; observe that $z \in S^*$, by (22) and (23), and then

$$\tau^* \leq \tau_z. \tag{25}$$

With this notation, the main conclusion of this paper can be stated as follows.

Theorem 3.3. Under Assumptions 2.1 and 2.3, the following assertions (i)–(ii) hold.
(i) For every $x \in S$,

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

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

Theorem 3.2 will be proved in the following section, and then, after establishing the necessary technical tools in Section 5, Theorem 3.3 will be verified in Section 6. Throughout the reminder Assumptions 2.1 and 2.3 are enforced.

4. EXISTENCE OF A FIXED POINT

In this section Theorems 3.2 will be verified. The backbone of the argument is the following result, establishing that T_λ is a continuous operator with respect to the topology of pointwise convergence on the space $[\![0, G]\!]$.

Theorem 4.1. Suppose that the sequence $\{W_n\} \subset [\![0, G]\!]$ converges pointwise to a function $V : S \to \mathbb{R}$, that is,

$$\lim_{n\to\infty} W_n(x) = V(x), \quad x \in S. \tag{26}$$

In this case

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

The proof of this theorem relies on the following lemma.

Lemma 4.2. (i) Consider a family $\{S_k\}$ of *finite* subsets of S such that

$$S = \bigcup_{k=0}^{\infty} S_k, \quad S_k \subset S_{k+1}, \quad k \in \mathbb{N}, \tag{27}$$

and for each $x \in S$ and $k \in \mathbb{N}$ define

$$\delta_k(x) := \sup_{a\in A(x)} \left[1 - \sum_{y\in S_k} p_{x,y}(a)\right] = \sup_{a\in A(x)} \sum_{y\in S\setminus S_k} p_{x,y}(a). \tag{28}$$

In this case,

$$\lim_{k\to\infty} \delta_k(x) = 0, \quad x \in S.$$

(ii) Suppose that $\{\tilde{W}_n\} \subset \mathcal{C}(S)$ is such that

$$c := \sup_{n\in\mathbb{N}} \|\tilde{W}_n\| < \infty \quad \text{and} \quad \lim_{n\to\infty} \tilde{W}_n(y) = 0, \quad y \in S. \tag{29}$$

In this case, for every $x \in S$

$$\sup_{a\in A(x)} e^{\lambda R(x,a)} \sum_{y\in S} p_{x,y}(a) |\tilde{W}_n(y)| \to 0 \quad \text{as} \quad n \to \infty.$$

Proof. (i) Since the sets S_k are finite, Assumption 2.1 yields that for each $k \in \mathbb{N}$ the mapping $a \mapsto \sum_{y\in S_k} p_{x,y}(a)$ is continuous on the compact space $A(x)$, whereas using the conditions in (27) it follows that

$$\sum_{y\in S_k} p_{x,y}(a) \nearrow \sum_{y\in S} p_{x,y}(a) = 1 \text{ as } k \nearrow \infty,$$

so that Dini's theorem implies that the convergence is uniform on the space $A(x)$, that is, $\sup_{a\in A(x)}\left[1-\sum_{y\in S_k}p_{x,y}(a)\right]\to 0$ as $k\to\infty$.

(ii) Let $x\in S$ be fixed and observe that for every $k\in\mathbb{N}$

$$
\begin{aligned}
&\sup_{a\in A(x)} e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)|\tilde W_n(y)|\\
&\quad\le \sup_{a\in A(x)} e^{\lambda R(x,a)}\sum_{y\in S_k}p_{x,y}(a)|\tilde W_n(y)| + \sup_{a\in A(x)} e^{\lambda R(x,a)}\sum_{y\in S\setminus S_k}p_{x,y}(a)|\tilde W_n(y)|\\
&\quad\le e^{|\lambda|\,\|R\|}\left(\max_{y\in S_k}|\tilde W_n(y)| + c\sup_{a\in A(x)}\sum_{y\in S\setminus S_k}p_{x,y}(a)\right)\\
&\quad= e^{|\lambda|\,\|R\|}\left(\max_{y\in S_k}|\tilde W_n(y)| + c\delta_k(x)\right)
\end{aligned}
$$

where (29) was used to set the second inequality, and the equality is due to (28). Recalling that the sets S_k are finite, the convergence in (29) yields

$$
\limsup_{n\to\infty}\left|\sup_{a\in A(x)} e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)|\tilde W_n(y)|\right| \le e^{|\lambda|\,\|R\|}c\,\delta_k(x), \quad x\in S,
$$

and then, since $k\in\mathbb{N}$ is arbitrary, the conclusion follows from part (i). □

Proof. (Proof of Theorem 4.1) Note that (15) and (26) together imply that $V\in[\![0,G]\!]$. Next, using (4) observe that

$$
\begin{aligned}
&\sup_{a\in A(x)}\sum_{y\in S}p_{x,y}(a)U_\lambda(R(x,a)+W_n(y))\\
&\quad= \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)U_\lambda(W_n(y))\right]\\
&\quad= \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)U_\lambda(V(y))\right.\\
&\qquad\qquad\left. + e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)[U_\lambda(W_n(y))-U_\lambda(V(y))]\right]\\
&\quad\le \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)U_\lambda(V(y))\right]\\
&\qquad\qquad + \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S}p_{x,y}(a)|U_\lambda(W_n(y))-U_\lambda(V(y))|\right],
\end{aligned}
$$

and an additional application of (4) leads to

$$\begin{aligned}\sup_{a\in A(x)}\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a)+W_n(y))& \\ \leq \sup_{a\in A(x)}\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a)+V(y))+\Delta_n(x),& \qquad (30)\end{aligned}$$

where

$$\Delta_n(x) := \sup_{a\in A(x)}\left[e^{\lambda R(x,a)}\sum_{y\in S} p_{x,y}(a)|U_\lambda(W_n(y))-U_\lambda(V(y))|\right], \qquad (31)$$

whereas the inequality

$$\begin{aligned}\sup_{a\in A(x)}\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a)+V(y))& \\ \leq \sup_{a\in A(x)}\sum_{y\in S} p_{x,y}(a)U_\lambda(R(x,a)+W_n(y))+\Delta_n(x),&\end{aligned}$$

can be established along similar lines. Combining the definition of T_λ in (16) with (30) and the previous display, it follows that $U_\lambda(T_\lambda[W_n](x)) \leq U_\lambda(T_\lambda[V](x)) + \Delta_n(x)$ and $U_\lambda(T_\lambda[V](x)) \leq U_\lambda(T_\lambda[W_n](x)) + \Delta_n(x)$, so that

$$|U_\lambda(T_\lambda[W_n](x)) - U_\lambda(T_\lambda[V](x))| \leq \Delta_n(x). \qquad (32)$$

Observe now that (3) and (26) together imply that

$$\lim_{n\to\infty}[U_\lambda(W_n(y)) - U_\lambda(V(y))] = 0, \quad y \in S.$$

Additionally, using that $\|W\| \leq \|G\|(<\infty)$ if $W \in [\![0,G]\!]$, the inclusions $W_n, V \in [\![0,G]\!]$ and (3) together yield that $\|U_\lambda(W_n(\cdot))\|, \|U_\lambda(V(\cdot))\| \leq e^{|\lambda|\,\|G\|}$, and then

$$\|U_\lambda(W_n(\cdot)) - U_\lambda(V(\cdot))\| \leq 2e^{|\lambda|\|G\|}.$$

Using Lemma 4.2 (ii) with $U_\lambda(W_n) - U_\lambda(V)$ instead of $\tilde{W}_n$, the two last displays and (31) together imply that $\lim_{n\to\infty}\Delta_n(\cdot) = 0$, a convergence that via (32) leads to $\lim_{n\to\infty} U_\lambda(T_\lambda[W_n](x)) = U_\lambda(T_\lambda[V](x))$ for each $x \in S$; since $U_\lambda(\cdot)$ is strictly increasing and continuous, it follows that $T_\lambda[W_n](x) \to T_\lambda[V](x)$ as $n \to \infty$ for every state x. □

Now, Theorem 4.1 will be used to derive the existence of a fixed point of the operator T_λ.

Proof. (Proof of Theorem 3.2) Set $W_{0,\lambda} := 0$ and $W_{n,\lambda} := T_\lambda^n[0]$ for $n \in \mathbb{N}\setminus\{0\}$, and observe that

$$W_{n+1,\lambda} = T_\lambda[W_{n,\lambda}], \quad n \in \mathbb{N}. \qquad (33)$$

Since $W_{0,\lambda} = 0 \in [\![0, G]\!]$ and $W_{1,\lambda} = T[0] \in [\![0, G]\!]$ it follows that $W_{1,\lambda} \geq W_{0,\lambda}$, and then an induction argument combining the above display and monotonicity property (18) immediately yields that

$$0 \leq W_{n,\lambda} \leq W_{n+1,\lambda} \leq G, \quad n \in \mathbb{N},$$

where the extreme inequalities are due to the fact that the functions $W_{k,\lambda}$ belong to $[\![0, G]\!]$. It follows that $\{W_{n,\lambda}(y)\}_{n\in\mathbb{N}}$ is always an increasing and bounded sequence, so that

$$\lim_{n\to\infty} W_{n,\lambda}(y) =: W^*_\lambda(y)$$

exists for every $y \in S$. From this point, Theorem 4.1 yields that $W^*_\lambda \in [\![0, G]\!]$ and

$$\lim_{n\to\infty} T_\lambda[W_{n,\lambda}](x) = T_\lambda[W^*_\lambda](x), \quad x \in S.$$

Thus, taking the limit as n goes to ∞ in both sides of (33), the two previous displays together imply that $W^*_\lambda = T_\lambda[W^*_\lambda]$, showing that W^*_λ is a fixed point of T_λ. □

5. TECHNICAL TOOLS

This section presents auxiliary results that will be used in the proof of Theorem 3.3. The following lemma extends property (12) to the class of all policies.

Lemma 5.1. For each $x \in S$, and $n \in \mathbb{N}$, define

$$M_n(x) := \sup_{\pi\in\mathcal{P}} P^\pi_x[\tau_z > n] \in [0, 1]. \tag{34}$$

With this notation,

(i) $\lim_{n\to\infty} M_n(x) = 0, \;\; x \in S.$

(ii) $P^\pi_x[\tau_z < \infty] = 1$ for every $x \in S$ and $\pi \in \mathcal{P}$.

Proof. Note that the inclusion $[\tau_z > n + 1] \subset [\tau_z > n]$ and (34) together lead to

$$M_{n+1} \leq M_n, \quad n \in \mathbb{N}, \tag{35}$$

and then

$$M(x) := \lim_{n\to\infty} M_n(x) \in [0, 1] \tag{36}$$

exists for every $x \in S$; since $P^\pi_z[\tau_z = 0] = 1$ for every $\pi \in \mathcal{P}$, by (13), it follows that $M_n(z) = 0$ for every positive n, so that

$$M(z) = 0.$$

Given $(x, \tilde{a}) \in \mathbb{K}$ and a policy $\pi \in \mathcal{P}$, define the new policy $\pi_{x,\tilde{a}} = \{\pi_{x,\tilde{a},n}\}$ as follows: for every $t \in \mathbb{N}$ and $h_t \in \mathbb{H}_t$, $\pi_{x,\tilde{a},t}(\cdot|h_t) = \pi_{t+1}(\cdot|x, \tilde{a}, h_t)$. Next, using (13), notice that

$[\tau_z > n+1] = [X_k \neq z, 0 \leq k \leq n+1]$ and note that an application of the Markov property yields that for every $\pi \in \mathcal{P}$, $n \in \mathbb{N}$ and $(x, \tilde{a}) \in \mathbb{K}$ with $x \neq z$

$$\begin{aligned} P_x^\pi[\tau_z > n+1 | A_0 = \tilde{a}] &= \sum_{y \in S \setminus \{z\}} p_{x,y}(\tilde{a}) P_y^{\pi_{x,\tilde{a}}}[\tau_z > n] \\ &\leq \sum_{y \in S \setminus \{z\}} p_{x,y}(\tilde{a}) M_n(y) \leq \sup_{a \in A(x)} \sum_{y \in S \setminus \{z\}} p_{x,y}(a) M(y), \end{aligned}$$

where the inequalities are due to (34)–(36). Therefore,

$$P_x^\pi[\tau_z > n+1] \leq \sup_{a \in A(x)} \sum_{y \in S \setminus \{z\}} p_{x,y}(a) M(y), \quad x \neq z.$$

Since the left hand side of this inequality is null when $x = z$, via (34) and (36) it follows that

$$M(x) \leq \sup_{a \in A(x)} \sum_{y \in S \setminus \{z\}} p_{x,y}(a) M(y), \quad x \in S.$$

Now, using that $M(\cdot)$ is bounded, observe that Assumption 2.1 implies that there exists a policy $\hat{f} \in \mathbb{F}$ such that $\sup_{a \in A(x)} \sum_{y \in S \setminus \{z\}} p_{x,y}(a) M(y) = \sum_{y \in S \setminus \{z\}} p_{x,y}(\hat{f}(x)) M(y)$ for every state x, and then

$$M(x) \leq \sum_{y \in S \setminus \{z\}} p_{x,y}(\hat{f}(x)) M(y) = \sum_{y \in S} p_{x,y}(\hat{f}(x)) M(y), \quad x \in S;$$

see (5) for the equality. Combining this relation with the Markov property, it follows that for every initial state $x \in S$ and $n \in \mathbb{N}$,

$$M(X_n) \leq E_x^{\hat{f}}[M(X_{n+1}) | X_n] = E_x^{\hat{f}}[M(X_{n+1}) | \mathcal{F}_n], \quad P_x^{\hat{f}}\text{-a. s.},$$

so that $\{(M(X_n), \mathcal{F}_n)\}$ is a submartingale with respect to $P_x^{\hat{f}}$. Since $M(\cdot)$ is bounded, the optional sampling theorem yields that, for every $x \in S$ and $n \in \mathbb{N}$,

$$M(x) \leq E_x^{\hat{f}}[M(X_{\tau_z \wedge n})] = E_x^{\hat{f}}[M(X_n)\ I[\tau_z > n]] \leq P_x^{\hat{f}}[\tau_z > n],$$

where, recalling that $M(z) = 0$, the equality was obtained form (14), and the inclusion in (36) was used in the last step. Since $\lim_{n\to\infty} P_x^{\hat{f}}[\tau_z > n] = P_x^{\hat{f}}[\tau_z = \infty] = 0$, by Assumption 2.3(i), the above display yields that $M(\cdot) = 0$, establishing part (i). To establish assertion (ii), combine (34) with the part (i) to obtain $P_x^\pi[\tau_z = \infty] = \lim_{n\to\infty} P_x^\pi[\tau_z > n] \leq \lim_{n\to\infty} M_n(x) = M(x) = 0$ for every $x \in S$ and $\pi \in \mathcal{P}$.

□

The following lemma shows that the space of strategies of player II can be reduced to the class of *finite* stopping times, a result that will be used in the proof of Theorem 3.3.

Lemma 5.2. For every $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$,

$$V_\lambda(\cdot, \pi, \tau) = V_\lambda(\cdot, \pi, \tau \wedge \tau_z). \tag{37}$$

Proof. Let $x \in S$ and $(\pi, \tau) \in \mathcal{P} \times \mathcal{T}$ be arbitrary. Using that $P_x^\pi[\tau_z < \infty] = 1$, by Lemma 5.1, Assumptions 2.1(ii) and 2.3 together with (13) yield that

$$\text{On } [\tau_z < \infty], \quad X_{\tau_z} = z \quad \text{and} \quad R(X_n, A_n) = G(X_n) = 0, \quad n \ge \tau_z. \tag{38}$$

Next, observe the following facts (a)–(c):
(a) On the event $[\tau = \infty] \cap [\tau_z < \infty]$ the equality $\tau \wedge \tau_z = \tau_z$ holds, and the above display yields that $R(X_t, A_t) = 0$ for $t \ge \tau \wedge \tau_z$ and $G(X_{\tau\wedge\tau_z}) = 0$, so that $\sum_{t=0}^{\tau-1} R(X_t, A_t) = \sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t)$ and $G(X_\tau)I[\tau < \infty] = 0 = G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]$. Thus,

$$\begin{aligned}&\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty]\\ &\quad= \sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty] \quad \text{on } [\tau = \infty, \tau_z < \infty];\end{aligned}$$

since $P_x^\pi[\tau_z < \infty] = 1$, by Lemma 5.1(ii), it follows that

$$\begin{aligned}&E_x^\pi\left[I[\tau = \infty]U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty]\right)\right]\\ &\quad= E_x^\pi\left[I[\tau = \infty]U_\lambda\left(\sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]\right)\right].\end{aligned}$$

(b) On the event $[\tau_z \le \tau < \infty]$, $\tau_z = \tau \wedge \tau_z$ and via (38) it follows that $G(X_\tau)I[\tau < \infty] = G(X_\tau) = 0 = G(X_{\tau_z}) = G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]$ as well as $\sum_{t=0}^{\tau-1} R(X_t, A_t) = \sum_{t=0}^{\tau_z-1} R(X_t, A_t) = \sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t)$, so that

$$\begin{aligned}&E_x^\pi\left[I[\tau_z \le \tau < \infty]\, U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty]\right)\right]\\ &\quad= E_x^\pi\left[I[\tau_z \le \tau < \infty]\, U_\lambda\left(\sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]\right)\right].\end{aligned}$$

(c) On the event $[\tau < \infty, \tau < \tau_z]$, $\tau = \tau \wedge \tau_z$, so that $\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty] = \sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]$, and then

$$\begin{aligned}&E_x^\pi\left[I[\tau < \infty, \tau < \tau_z]\, U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau)I[\tau < \infty]\right)\right]\\ &\quad= E_x^\pi\left[I[\tau < \infty, \tau < \tau_z]\, U_\lambda\left(\sum_{t=0}^{\tau\wedge\tau_z-1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z})I[\tau \wedge \tau_z < \infty]\right)\right].\end{aligned}$$

Since $1 = I[\tau = \infty] + I[\tau_z \le \tau < \infty] + I[\tau < \infty, \tau < \tau_z]$, the three last displays together imply that

$$\begin{aligned} &E_x^\pi\left[U_\lambda\left(\sum_{t=0}^{\tau-1} R(X_t, A_t) + G(X_\tau) I[\tau < \infty]\right)\right] \\ &\qquad = E_x^\pi\left[U_\lambda\left(\sum_{t=0}^{\tau\wedge\tau_z - 1} R(X_t, A_t) + G(X_{\tau\wedge\tau_z}) I[\tau\wedge\tau_z < \infty]\right)\right]. \end{aligned}$$

Via (3) and (6), this relation leads to $U_\lambda(V_\lambda(x; \pi, \tau)) = U_\lambda(V_\lambda(x; \pi, \tau \wedge \tau_z))$, and (37) follows using that $U_\lambda(\cdot)$ is strictly increasing. □

6. PROOF OF THEOREM 3.3

In this section the existence of a Nash equilibrium will be verified. Since the proof is rather technical, to ease the presentation the essential steps have stated separately in Theorems 6.1 and Lemma 6.2 below. At this point it is convenient to have a glance at (21)–(24).

Theorem 6.1. For each $\tau \in \mathcal{T}$,

$$W_\lambda^*(\cdot) \le V_\lambda(\cdot; f^*, \tau). \tag{39}$$

The proof of this result relies on the following two lemmas.

Lemma 6.2. For every $n \in \mathbb{N}$, $x \in S$ and $\tau \in \mathcal{T}$,

$$\begin{aligned} &U_\lambda(W_\lambda^*(x)) \\ &\quad \le \sum_{k=0}^{n} E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k)\right) I[\tau = k]\right] \\ &\qquad + E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau \ge n+1]\right]. \end{aligned} \tag{40}$$

Proof. To begin with, observe that (20) and (21) together yield that, for every state x,

$$U_\lambda(W_\lambda^*(x)) \le \sum_{y\in S} p_{x,y}(f^*(x)) U_\lambda(R(x, f^*(x)) + W_\lambda^*(y)), \tag{41}$$

a relation that via the Markov property implies that, for every $x \in S$ and $n \in \mathbb{N}$,

$$U_\lambda(W_\lambda^*(X_n)) \le E_x^{f^*}\left[U_\lambda(R(X_n, A_n) + W_\lambda^*(X_{n+1})) | \mathcal{F}_n\right]. \tag{42}$$

Next, (40) will be verified by induction. Let $x \in S$ and $\tau \in \mathcal{T}$ be arbitrary. Combining the convention (1) with the relations $\tau \geq 0$ and $P_x^{f^*}[X_0 = x] = 1$, it follows that

$$\begin{aligned}
&U_\lambda(W_\lambda^*(x)) \\
&\quad = U_\lambda(W_\lambda^*(X_0))I[\tau = 0] + U_\lambda(W_\lambda^*(X_0))I[\tau \geq 1] \\
&\quad = U_\lambda\left(\sum_{t=0}^{0-1} R(X_t, A_t) + W_\lambda^*(X_0)\right) I[\tau = 0] + U_\lambda(W_\lambda^*(X_0))I[\tau \geq 1] \\
&\quad \leq U_\lambda\left(\sum_{t=0}^{0-1} R(X_t, A_t) + W_\lambda^*(X_0)\right) I[\tau = 0] \\
&\qquad\qquad + I[\tau \geq 1]E_x^{f^*}\left[U_\lambda(R(X_0, A_0) + W_\lambda^*(X_1))|\, \mathcal{F}_0\right] \\
&\quad = U_\lambda\left(\sum_{t=0}^{0-1} R(X_t, A_t) + W_\lambda^*(X_0)\right) I[\tau = 0] \\
&\qquad\qquad + E_x^{f^*}\left[U_\lambda(R(X_0, A_0) + W_\lambda^*(X_1))I[\tau \geq 1]|\, \mathcal{F}_0\right], \quad P_x^{f^*}\text{-a. s.}
\end{aligned}$$

where the inequality is due to (42) with $n = 0$, and the inclusion $[\tau \geq 1] \in \mathcal{F}_0$ was used to set the last equality. After taking the expectation with respect to $P_x^{f^*}$, the above display yields that the case $n = 0$ of (40) holds. Next, assume that $n \in \mathbb{N}$ is such that (40) is valid, and observe that

$$\begin{aligned}
&U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau \geq n+1] \\
&\quad = U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau = n+1] \\
&\qquad\qquad + U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau \geq n+2]
\end{aligned}$$

whereas, using (4),

$$\begin{aligned}
&U_\lambda\left(\sum_{t=0}^{n} R(X_t, A_t) + W_\lambda^*(X_{n+1})\right) I[\tau \geq n+2] \\
&\quad = e^{\lambda \sum_{t=0}^n R(X_t, A_t)} I[\tau \geq n+2] U_\lambda\left(W_\lambda^*(X_{n+1})\right) \\
&\quad \leq e^{\lambda \sum_{t=0}^n R(X_t, A_t)} I[\tau \geq n+2] E_x^{f^*}\left[U_\lambda(R(X_{n+1}, A_{n+1}) + W_\lambda^*(X_{n+2}))|\, \mathcal{F}_{n+1}\right] \\
&\quad = E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n+1} R(X_t, A_t) + W_\lambda^*(X_{n+2})\right) I[\tau \geq n+2] \middle|\, \mathcal{F}_{n+1}\right]
\end{aligned}$$

where (42) with $n+1$ instead of n was used to set the inequality, and the second equality was obtained combining (4) with the fact that the random variable $e^{\lambda \sum_{t=0}^n R(X_t, A_t)} I[\tau \geq$

$n+2]$ is $\mathcal{F}_{n+1}$-measurable. These two last displays together imply that

$$\begin{aligned} E_x^{f^*}&\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t,A_t)+W_\lambda^*(X_{n+1})\right)I[\tau\geq n+1]\right]\\ &\leq E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n} R(X_t,A_t)+W_\lambda^*(X_{n+1})\right)I[\tau= n+1]\right]\\ &\qquad + E_x^{f^*}\left[U_\lambda\left(\sum_{t=0}^{n+1} R(X_t,A_t)+W_\lambda^*(X_{n+2})\right)I[\tau\geq n+2]\right]. \end{aligned}$$

Combining this relation with the induction hypothesis, it follows that (40) holds with $n+1$ instead of n, completing the induction argument. □

Lemma 6.3. Given $x\in S$, let $f\in\mathbb{F}$ and $\tau\in\mathcal{T}$ be such that

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

In this case

$$\lim_{n\to\infty} E_x^f\left[\left|U_\lambda\left(\sum_{k=0}^{n} R(X_t,A_t)\right)\right| I[\tau>n+1]\right]=0. \tag{43}$$

Proof. Since G is bounded, from (6) and (7) it follows that the condition $V_\lambda(x;f,\tau)<\infty$ is equivalent to

$$E_x^f\left[e^{\lambda\sum_{k=0}^{\tau-1}R(X_t,A_t)}\right]\in(0,\infty), \tag{44}$$

so that $P_x^f[e^{\lambda\sum_{k=0}^{\tau-1}R(X_t,A_t)}<\infty]=1$. Combining this fact with the condition $P_x^f[\tau<\infty]=1$ it follows that

$$(1\vee e^{\lambda\sum_{k=0}^{\tau-1}R(X_t,A_t)})I[\tau>n+1]\to 0 \quad \text{as } n\to\infty \quad P_x^f\text{-a. s.},$$

and then (44) and the dominated convergence theorem together imply that

$$E_x^f\left[(1\vee e^{\lambda\sum_{k=0}^{\tau-1}R(X_t,A_t)})I[\tau>n+1]\right]\to 0 \quad \text{as } n\to\infty;$$

This convergence and the inequality $1\vee e^{\lambda\sum_{k=0}^{\tau-1}R(X_t,A_t)}\geq e^{\lambda\sum_{k=0}^{n}R(X_t,A_t)}$ lead to $\lim_{n\to\infty} E_x^f\left[e^{\lambda\sum_{k=0}^{n}R(X_t,A_t)}I[\tau>n+1]\right]=0$, and the deisred conclusion (43) follows via (3). □

Now, Lemmas 6.2 and 6.3 will be used to establish Theorem 6.1.

Proof. (Proof of Theorem 6.1) By Lemma 5.2, without loss of generality τ can be replaced by $\tau\wedge\tau_z$, and then Assumption 2.3 yields that it is sufficient to establish the conclusion under the condition that τ is a finite stopping time:

$$P_x^{f^*}[\tau<\infty]=1, \quad x\in S. \tag{45}$$

Since (39) certainly holds if $V_\lambda(\cdot; f^*, \tau) = \infty$, in the following argument it will be supposed that

$$V_\lambda(\cdot; f^*, \tau) < \infty. \tag{46}$$

Observe that (4) and the inclusion $W_\lambda^* \in [\![0, G]\!]$ together yield that

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

Notice that, via Lemma 6.3, (45) and (46) 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

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

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

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

where the inequality is due to the inclusion $W_\lambda^* \in [\![0, G]\!]$ and the monotonicity of $U_\lambda(\cdot)$ and, using (45), the last equality is due to (3) and (6). Taking the limit as n goes to ∞ in the right-hand side of (40), the two previous displays yield that $U_\lambda(W_\lambda^*(x)) \leq U_\lambda(V_\lambda(x, f^*, \tau))$ and then (39) follows using that $U_\lambda(\cdot)$ is strictly increasing. □

The last major step before the proof of Theorem 3.3 is the following.

Theorem 6.4. For every $x \in S$

$$V_\lambda(x; \pi, \tau^*) \leq W_\lambda^*(x), \quad \pi \in \mathcal{P}. \tag{47}$$

The proof of this theorem depends on the following lemma.

Lemma 6.5. (i) For each $x \in S$ and $\pi \in \mathcal{P}$, and $n = 1, 2, \ldots$

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

(ii) For every $n \in \mathbb{N}$, $x \in S \setminus S^*$ and $\pi \in \mathcal{P}$,

$$\begin{aligned}
U_\lambda(W_\lambda^*(x)) \geq \sum_{k=1}^{n} E_x^\pi &\left[U_\lambda \left(\sum_{t=0}^{k-1} R(X_t, A_t) + W_\lambda^*(X_k) \right) I[\tau^* = k] \right] \\
&+ E_x^\pi \left[U_\lambda \left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n) \right) I[\tau^* > n] \right]. \qquad (48)
\end{aligned}$$

Proof. First, observe that $U_\lambda(W_\lambda^*(x)) < U_\lambda(G(x))$ when $x \notin S^*$, by (19) and (23), and then

$$\begin{aligned}
&U_\lambda(W_\lambda^*(x)) \\
&\quad = \sup_{a \in A(x)} \sum_{y \in S} p_{x,y}(a) U_\lambda \left(R(x,a) + W_\lambda^*(y) \right) \\
&\quad \geq \sum_{y \in S} p_{x,y}(a) U_\lambda \left(R(x,a) + W_\lambda^*(y) \right), \quad x \in S \setminus S^*, \quad a \in A(x). \qquad (49)
\end{aligned}$$

(i) Let $\pi \in \mathcal{P}$ be arbitrary and, using that $X_t \notin S^*$ for $0 \leq t < \tau^*$, by (24), the above display and the Markov property together yield that for each $n \in \mathbb{N}$ the following relation holds almost surely with respect to P_x^π:

$$\begin{aligned}
U_\lambda(W_\lambda^*(X_n)) &\geq \sum_{y \in S} p_{X_n,y}(A_n) U_\lambda \left(R(X_n, A_n) + W_\lambda^*(y) \right) \\
&= E_x^\pi \left[U_\lambda \left(R(X_n, A_n) + W_\lambda^*(X_{n+1}) \right) \middle| \mathcal{F}_n, A_n \right] \text{ on } [\tau^* > n].
\end{aligned}$$

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 (4) leads to

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

From this point, the conclusion follows taking the expectation with respect to P_x^π and using the equality $I[\tau^* > n] = I[\tau^* = n+1] + I[\tau^* > n+1]$.

(ii) The argument is by induction. Let $x \in S \setminus S^*$ and $\pi \in \mathcal{P}$ be arbitrary, and note that (49) leads to $U_\lambda(W_\lambda^*(x)) \geq E_x^\pi \left[U_\lambda\left(R(X_0, A_0) + W_\lambda^*(X_1)\right)\right]$; since $P_x^\pi[\tau^* > 0] = 1$, by (24) it follows that

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

an expression that is equivalent to (48) with $n = 1$. Suppose now that $n \in \mathbb{N}$ is such that (48) is valid. In this case, direct calculations combining part (i) with the induction hypothesis show that (48) also holds with $n+1$ instead of n, completing the induction argument. □

Proof. (Proof of Theorem 6.4) First, note that (25) and Lemma 5.1(ii) together imply that

$$P_x^\pi[\tau^* < \infty] = 1, \quad x \in S. \tag{50}$$

Now, let $\pi \in \mathcal{P}$ be arbitrary and suppose that $x \in S^*$, so that (23) and (24) yield that

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

whereas (1) and (6) together lead to $V_\lambda(x; \pi, \tau^*) = G(x)$, and then (47) holds with equality. Next, the desired conclusion will be verified when the initial state x does not belong to S^*. Consider the following claim:

$$\begin{aligned} &\text{For every } x \in S \setminus S^*, \text{ and } \pi \in \mathcal{P}, \\ &\liminf_{n\to\infty} E_x^\pi\left[U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n]\right] \geq 0. \end{aligned} \tag{51}$$

Observing that $U_\lambda(\cdot) > 0$ when λ is positive, it is clear that the above assertion holds if $\lambda > 0$. To complete the proof of (51), suppose that $\lambda < 0$ and note that (3) and the non-negativity of R and W_λ^* together yield that $\left|U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n]\right| \leq I[\tau^* > n]$, by (3), and via (50) it follows that, as $n \to \infty$,

$$E_x^\pi\left[\left|U_\lambda\left(\sum_{t=0}^{n-1} R(X_t, A_t) + W_\lambda^*(X_n)\right) I[\tau^* > n]\right|\right] \leq P_x^\pi[\tau^* > n] \to 0,$$

a convergence that immediately yields that (51) holds with equality when λ is negative.

Next, using that $U_\lambda(\cdot)$ has constant sign, the monotone convergence theorem yields that

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

where the last equality follows combining (6) and (50). To conclude, take the inferior limit as n goes to ∞ in the right-hand side of (48) to obtain, via the the above display and (51), that $U_\lambda(W_\lambda^*(x)) \geq U_\lambda(V_\lambda(x;\pi,\tau^*))$, an inequality that using that U_λ is strictly increasing leads to $W_\lambda^*(x) \geq V_\lambda(x;\pi,\tau^*)$, showing that (47) is also valid for $x \in S \setminus S^*$.

□

Finally, the two previous theorems will be used to establish the existence of a Nash equilibrium.

Proof. (Proof of Theorem 3.3) By Theorems 6.1 and 6.4

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

Setting $(\pi,\tau) = (f^*,\tau^*)$ it follows that $W_\lambda^*(\cdot) = V_\lambda(\cdot; f^*, \tau^*)$, establishing part (i), and combining this fact with the above display it follows from Definition 2.2 that (f^*,τ^*) is a Nash equilibrium, completing the proof. □

7. CONCLUSION

In this note Markov stopping games with bounded rewards and risk-sensitive total reward criterion were studied. Besides mild continuity-compactness requirements, the framework of the paper was determined by the existence of an absorbing state postulated in Assumption 2.3, and in that context the existence of a Nash equilibrium was studied. The main conclusion of the paper, stated in Theorem 3.3, establishes that given a fixed point W_λ^* of the operator T_λ in Definition 3.1, it is possible to define a Nash equilibrium $(f^*,\tau^*) \in \mathbb{F} \times \mathcal{T}$, and that the value function of the game $V_\lambda(\cdot; f^*, \tau^*)$ equals $W_\lambda^*(\cdot)$, a fact that immediately yields that T_λ has a *unique* fixed point. On the other hand, studying the existence of Nash equilibria for Markov stopping games under more general conditions than those assumed in this work, for instance, for models that do not satisfy Assumption 2.3, seems to be an interesting problem.

(Received November 25, 2021)

REFERENCES

[1] A. Alanís-Durán and R. Cavazos-Cadena: An optimality system for finite average Markov decision chains under risk-aversion. Kybernetika *48* (2012), 83–104.

[2] E. Altman and A. Shwartz: Constrained Markov games: Nash equilibria. In: Annals of Dynamic Games (V. Gaitsgory, J. Filar, and K. Mizukami, eds.), Birkhauser, Boston 2000, pp. 213–221.

[3] R. Atar and A. Budhiraja: A stochastic differential game for the inhomogeneous Laplace equation. Ann. Probab. *38* (2010), 2, 498–531. DOI:10.1214/09-aop494

[4] S. Balaji and S. P. Meyn: Multiplicative ergodicity and large deviations for an irreducible Markov chain. Stoch. Proc. Appl. *90* (2000), 1, 123–144. DOI:10.1016/s0304-4149(00)00032-6

[5] N. Bäuerle and U. Rieder: Markov Decision Processes with Applications to Finance. Springer, New York 2011.

[6] N. Bäuerle and U. Rieder: More risk-sensitive Markov decision processes. Math. Oper. Res. *39* (2014), 1, 105–120. DOI:10.1287/moor.2013.0601

[7] N. Bäuerle and U. Rieder: Zero-sum risk-sensitive stochastic games. Stoch. Proc. Appl. *127* (2017), 2, 622–642. DOI:10.1016/j.spa.2016.06.020

[8] T. R. Bielecki, D. Hernández-Hernández, and S. R. Pliska: Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. Mathematical Methods of OR *50* (1999), 167–188. DOI:10.1007/s001860050094

[9] V. S. Borkar and S. F. Meyn: Risk-sensitive optimal control for Markov decision process with monotone cost. Math. Oper. Res. *27* (2002), 1, 192–209. DOI:10.1287/moor.27.1.192.334

[10] R. Cavazos-Cadena and D. Hernández-Hernández: A system of Poisson equations for a non-constant Varadhan functional on a finite state space. Appl. Math. Optim. *53* (2006), 101–119. DOI:10.1007/s00245-005-0840-3

[11] R. Cavazos-Cadena and D. Hernández-Hernández: Nash equilibria in a class of Markov stopping games. Kybernetika *48* (2012), 5, 1027–1044.

[12] R. Cavazos-Cadena, L. Rodríguez-Gutiérrez, and D. M. Sánchez-Guillermo: Markov stopping games with an absorbing state and total reward criterion. Kybernetika *57* (2021), 474–492. DOI:10.14736/kyb-2021-3-0474

[13] E. V. Denardo and U. G. Rothblum: A turnpike theorem for a risk-sensitive Markov decision process with stopping. SIAM J. Control Optim. *45* (2006), 2, 414–431. DOI:10.1137/S0363012904442616

[14] G. B. Di Masi and L. Stettner: Risk-sensitive control of discrete time Markov processes with infinite horizon. SIAM J. Control Optim. *38* (1999), 1, 61–78. DOI:10.1137/S0363012997320614

[15] G. B. Di Masi and L. Stettner: Infinite horizon risk sensitive control of discrete time Markov processes with small risk. Syst. Control Lett. *40* (2000), 15–20. DOI:10.1016/S0167-6911(99)00118-8

[16] G. B. Di Masi and L. Stettner: Infinite horizon risk sensitive control of discrete time Markov processes under minorization property. SIAM J. Control Optim. *46* (2007), 1, 231–252. DOI:10.1137/040618631

[17] J. A.Filar and O. J. Vrieze: Competitive Markov Decision Processes. Springer, New York 1996.

[18] O. Hernández-Lerma: Adaptive Markov Control Processes. Springer, New York 1989.

[19] R. A. Howard and J. E. Matheson: Risk-sensitive Markov decision processes. Manage. Sci. *18* (1972), 7, 349–463. DOI:10.1287/mnsc.18.7.356

[20] A. Jaśkiewicz: Average optimality for risk sensitive control with general state space. Ann. Appl. Probab. *17* (2007), 2, 654–675. DOI:10.1214/105051606000000790

[21] V. N. Kolokoltsov and O. A. Malafeyev: Understanding Game Theory. World Scientific, Singapore 2010.

[22] I. Kontoyiannis and S. P. Meyn: Spectral theory and limit theorems for geometrically ergodic Markov processes. Ann. Appl. Probab. *13* (2003), 1, 304–362. DOI:10.1214/aoap/1042765670

[23] V. M. Martínez-Cortés: Bi-personal stochastic transient Markov games with stopping times and total reward criterion. Kybernetika *57* (2021), 1, 1–14. DOI:10.14736/kyb-2021-1-0001

[24] G. Peskir: On the American option problem. Math. Finance *15* (2007), 169–181. DOI:10.1111/j.0960-1627.2005.00214.x

[25] G. Peskir and A. Shiryaev: Optimal Stopping and Free-Boundary Problems. Birkhauser, Boston 2006.

[26] M. Pitera and L. Stettner: Long run risk sensitive portfolio with general factors. Math. Meth. Oper. Res. *82* (2016), 2, 265–293. DOI:10.1007/s00186-015-0514-0

[27] M. L. Puterman: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York 1994.

[28] L. S. Shapley: Stochastic games. Proc. National Academy Sci. *39* (1953), 10, 1095–1100.

[29] A. Shiryaev: Optimal Stopping Rules. Springer, New York 2008.

[30] K. Sladký: Growth rates and average optimality in risk-sensitive Markov decision chains. Kybernetika *44* (2008), 2, 205–226.

[31] K. Sladký: Risk-sensitive average optimality in Markov decision processes. Kybernetika *54* (2018), 6, 1218–1230. DOI:10.14736/kyb-2018-6-1218

[32] L. Stettner: Risk sensitive portfolio optimization. Math. Meth. Oper. Res. *50* (1999), 3, 463–474. DOI:10.1007/s001860050081

[33] L. E. Zachrisson: Markov Games. Princeton University Press *12*, Princeton 1964.

Jaicer López-Rivero, Facultad de Ciencias Físico-Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla, PUE. México.
e-mail: jaicer.lopez@alumno.buap.mx

Rolando Cavazos-Cadena, Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Saltillo, COAHV. México.
e-mail: rolando.cavazos@uaaan.edu.mx

Hugo Cruz-Suárez, Facultad de Ciencias Físico-Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla, PUE. México.
e-mail: hcs@fcfm.buap.mx