Stochastic Control Problems with Probability and Risk-sensitive Criteria

A Dissertation

submitted in partial fulfilment

of the requirements for the award of the

degree of

Doctor of Philosophy

by

Arnab Bhabak

Department of Mathematics

Indian Institute of Technology Guwahati

Guwahai - 781039

March, 2023

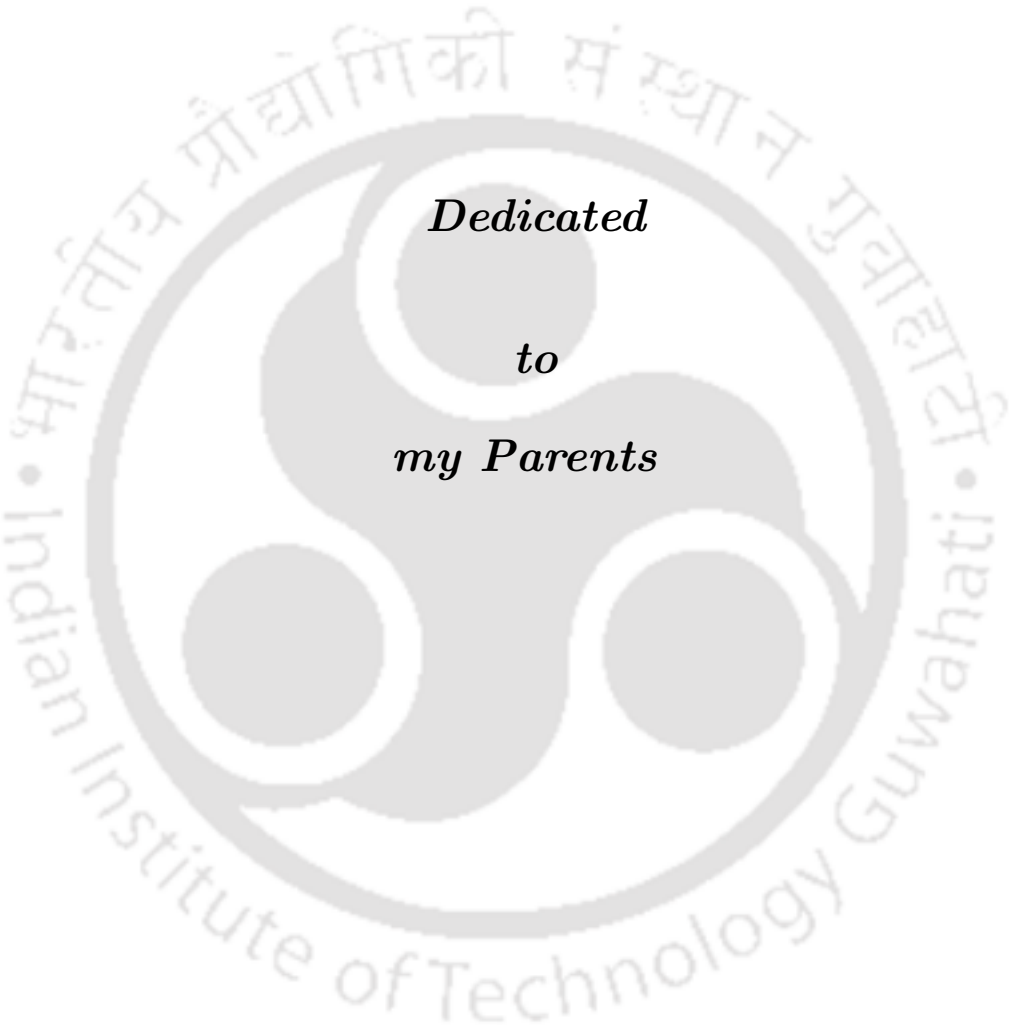
Declaration

I hereby declare that the work reported in this thesis is entirely original and has been carried out by me under the supervision of Dr. Subhamay Saha at the Department of Mathematics, Indian Institute of Technology Guwahati. I further declare that this work has not been the basis for the award of any degree, diploma, fellowship, associateship or similar title of any University or Institution.

Arnab Bhabak
Roll No-186123004

Certified

Dr. Subhamay Saha
(Research advisor)



Dedicated

to

my Parents

Acknowledgement

As with any piece of research that results in the production of a thesis, on the cover there should be not only the name of the researcher, but also the names of all those unsung heroes, those who to varying degrees provided assistance, encouragement and guidance, and without whom I would not have succeeded. I am very grateful to all those people, my heroes, who have given me so much of their time, love and energy. In producing this thesis, I faced my final academic challenge, to gain a PhD.

First and foremost I would like to thank my research supervisor Dr. Subhamay Saha for his careful guidance and constant encouragement for all these years. It was he who introduced me to the subject of stochastic control theory and then guided me in my exploration of the subject. I went to him time and again with all my doubts and he was more than happy to clarify them. But apart from academic matters, he is one of the most fantastic human beings I have ever met in my life. He was not only my research guide but my friend and philosopher and that's what makes him so special. And lastly I just wanna thank him for his extreme patience that he has demonstrated in tolerating a mischievous student like me.

I take this opportunity to express my gratefulness to the members of my doctoral committee Prof. N. Selvaraju, Dr. Sriparna Bandopadhyay, and Dr. Ayon Ganguly for investing their valuable time throughout the progress of my research. I'm also highly indebted to Dr. Chandan Pal because of the several useful discussions towards the initial part of the work of this thesis. His constant encouragement throughout these five years is highly appreciated. Also, I convey my heartfelt thankfulness to all the other faculties in our department for their immense help and support in every bit of my learning phase from the beginning.

Next, I express my deep sense of gratitude towards Dr. Sukanta Bhattacharjee, who is kind of an elder brother to me, he is always there for me through my ups and downs. We played some epic Badminton and Cricket matches as a part of the same team. And yes, of course I had some delicious dishes at his venue which I crave the most. He made my stay at IITG an experience worth remembering.

And how can I forget one of my best buddy Bivakar Bose! He is one of a kind of a guy anyone will ask for by his side in every possible situation. A master-chef, who can make

literally make any dishes with the best of his efforts. A genius of a guy academically, but also a greater human being.

Also, I would like to thank Madhusudan Mandal, Saptarshi Mandal, Gargi di for being there in my side in every possible situations. It will be disrespectful if I don't mention Subrata Jana Sir, who first introduced the fun facts of mathematics to me.

Research life is quite a topsy turvy one and at times it can be quite frustrating unless you have some nice company around you. In this regard I have been quite fortunate. Right from my first year I have enjoyed some fantastic camaraderie. Among the seniors I should definitely mention the names of Dibakar da, Avijit da. Among the juniors Sagar, Mizanur, Arindam, Bera were all very enjoyable companionship. Special thanks to our krantikari group members Sirshendu, Aryan, Ashish. Also, I would like to thank Chottu, my childhood friend cum neighbor for his unconditional support towards my family and me.

Since sports activities are integral part of my life, finally I would like to salute my badminton teammates Sukanta da, Partha, Prakash, Rishav, Bikram, Rik and also our little short cricket family.

My father always wanted me to be a good human being rather than just a good student. In the process of that I always draw motivations from a few whom I would like to thank. Dr. APJ Abdul kalam is one of them. Ever since my childhood I always looked upto captain of the ship M.S.Dhoni. Roger Federer, one of the few I am very much fond of on and off the court.

At the end, I don't want to even try to acknowledge my parents Swapan Bhabak and Runu Bhabak and my brother Parnab because I think English dictionary does not have enough adjectives to describe their impact on my life.

March, 2023

Arnab Bhabak

Abstract

In this thesis we consider stochastic control problems with probability and risk-sensitive criterion. We consider both single and multi controller problems. Under probability criterion we first consider a zero-sum game with semi-Markov state process. We consider a general state and finite action spaces. Under suitable assumptions, we establish the existence of value of the game and also characterize it through an optimality equation. In the process we also prescribe a saddle point equilibrium. Next we consider a zero-sum game with probability criterion for continuous time Markov chains. We consider denumerable state space and unbounded transition rates. Again under suitable assumptions, we show the existence of value of the game and also characterize it as the unique solution of a pair of Shapley equations. We also establish the existence of a randomized stationary saddle point equilibrium.

In the risk-sensitive setup we consider a single controller problem with semi-Markov state process. The state space is assumed to be discrete. In place of the classical risk-sensitive utility function, which is the exponential function, we consider general utility functions. The optimization criteria also contains a discount factor. We investigate random finite horizon and infinite horizon problems. Using a state augmentation technique we characterize the value functions and also prescribe optimal controls. We then consider risk-sensitive game problems. We study zero and non-zero sum risk-sensitive average criterion games for semi-Markov processes with a finite state space. For the zero-sum case, under suitable assumptions we show that the game has a value. We also establish the existence of a stationary saddle point equilibrium. For the non-zero sum case, under suitable assumptions we establish the existence of a stationary Nash equilibrium.

Finally, we also consider a partially observable model. More specifically, we investigate partially observable zero sum games where the state process is a discrete time Markov chain. We consider a general utility function in the optimization criterion. We show the existence of value for both finite and infinite horizon games and also establish the existence of optimal polices. The main step involves converting the partially observable game into a completely observable game which also keeps track of the total discounted accumulated reward/cost.

Contents

CHAPTER 1

Introduction

In this thesis we consider stochastic control problems, both with single and multiple controllers, under probability and risk-sensitive criteria. A stochastic control problem is an optimization problem, wherein one or more *controllers* try to control the evolution of a stochastic process by taking some *actions*. These actions either result in a cost or yield a reward. The aim of the controller or controllers is to maximize, in case of reward, or minimize, in case of cost, which is accumulated either over a finite or infinite time horizon. When there is more than one controller the stochastic control problem is referred to as stochastic game problem. A stochastic game problem is further classified into zero-sum game, wherein one *player* is trying to maximize his or her payoff and the other *player* is trying to minimize the same. While in a non-zero sum game each player is trying to optimize his or her individual payoff. In real life, stochastic control problems arises naturally in the fields of telecommunication, queueing systems, epidemiology, finance etc. Stochastic control problems are considered for both discrete and continuous time processes. In discrete time the state process considered is discrete time Markov chain. Such control problems are popularly referred to in literature as Discrete Time Markov Markov Decision Processes or DTMDP in short. In continuous-time the state processes that are considered are continuous time Markov chains also known in literature as Continuous Time Markov Decision Processes or CTMDP, semi-Markov processes also known as Semi-Markov Decision Processes or SMDP, and diffusions. Stochastic control problems are considered for both finite and infinite time horizons. Finite horizon control problems with both discounted as well as undiscounted cost/reward are studied. In case of infinite horizon two popular criteria are infinite horizon discounted cost/reward criterion

and infinite horizon average cost/reward criterion.

Now since the cost or reward functional which the controller(s) is(are) trying to optimize, depends on the random evolution of the controlled stochastic process, it itself is a random quantity. Thus the most obvious approach is to try to optimize the expectation of the random quantity. This approach is known as risk-neutral approach. But expectation optimization has the short coming that the risk of the optimal control can be quite large and hence this can be a major concern in real life applications. In order to address this concern, there are two existing remedies in the literature. One of the approaches that is available in the literature is to try to optimize the expectation of the exponential of the cost or reward functional. This approach is known as the *risk-sensitive approach*. Since the Taylor series expansion of the exponential function involves higher powers, the risk-sensitive approach takes care of the higher moments as well. Thus it takes care of the "risk". Risk-sensitive control problems have found wide applications in the field of mathematical finance. The other approach is what is called the probability criterion approach. Here the controller tries to maximize (or minimize in case of cost) the probability that the reward will exceed a predetermined threshold. Stochastic control problems with probability criterion has applications in the analysis of manufacturing systems.

Another classification of stochastic control problems arises from the point of view of observability. Sometimes the underlying stochastic process may not be completely observable to the controller(s). In such situations the controller needs to decide on his/her actions based on the available partial observation. Such control problems are referred to as partially observable stochastic control problems. Partially observable control problems arise naturally in telecommunications, where instead of the actual signal, a noisy observation of the same is available for decision making.

In this thesis we consider zero-sum stochastic game problems with probability criterion for continuous time Markov chains and semi-Markov processes. Before this thesis, in the existing literature only discrete time stochastic games with probability criterion were considered. This thesis thus extends the literature to the continuous time setup as well. In the risk-sensitive setup we consider a single controller problem with semi-Markov state process. In fact we consider discounted cost and general utility functions, which subsumes the classical risk-sensitive case, where the utility function is exponential. Prior literature on risk-sensitive SMDPs consisted of works on finite horizon criterion and average cost criterion. This work thus complements the existing literature. In the multi controller setup, we investigate risk-sensitive average cost stochastic games, both zero-sum and non-zero sum, again for semi-

Markov processes. To the best of our knowledge, this seems to be the first work on risk-sensitive games for semi-Markov processes. We also investigate a risk-sensitive zero-sum game problem in discrete time, under partial observation. Although there were works on partially observable risk-neutral games, there was no literature on risk-sensitive partially observable games.

Now we give a brief and verbal description of the basic aims in various stochastic control problems and the primary techniques used in analyzing them. The primary quantity associated with a single controller stochastic control problem is the value function, which is the infimum/supremum over all admissible control policies of the cost/reward functional. The two basic aims for such control problems is to characterize the value function and to establish the existence of an optimal control. For analyzing N stage finite horizon problems, value functions for all n stages, $n = 0, 1, ..., N$ are defined. The zero stage value function is equal to the terminal reward. Then the $n+1$ stage value function is expressed in terms of the n stage value function via an appropriately defined operator. The optimal control is given by the minimizing/maximizing selectors. In order to ensure the existence of minimizing/maximizing selectors, suitable continuity-compactness conditions are assumed. The infinite horizon discounted optimization problems are solved by characterizing the value function as the unique solution to a suitably defined optimality equation. The average reward problem is studied via limit of infinite horizon discounted optimality problem. In the discrete time, where discount factor is given by some $\beta \in (0, 1)$, the discount factor is taken to 1, while in the continuous time setup where the discount factor is given by $e^{-\alpha}, \alpha \in (0, \infty)$, α is taken to 0. This approach is popularly known as vanishing discount approach. Again, optimal controls are specified by minimizing/maximizing selectors of the optimality equation.

Risk-sensitive stochastic control problems with single controller are solved by either converting it to equivalent stochastic games or by converting them to equivalent risk-neutral problems using a state-augmentation technique. Risk-sensitive average optimality problems are also studied using an eigenvalue approach. This is because the optimality equation here looks like a nonlinear eigenvalue equation.

For zero-sum stochastic games, two primary quantities of interest are upper and lower value functions, given by infsup and supinf of the reward functional, respectively. In general the lower value function is dominated by the upper value function. The game is said to have a value if this two functions agree. One of the aims towards solving a zero-sum stochastic game is to show that the game has a value. The other requirement is to establish the existence of a saddle-point equilibrium. A saddle-point equilibrium is a pair of control policies such

that if one player unilaterally deviates from the pair, then his/her payoff will be worse as compared to the payoff under saddle-point equilibrium policy. It turns out that existence of saddle-point equilibrium ensures the existence of the value of the game. In the analysis of zero-sum stochastic games it is required to consider relaxed control policies, wherein a controller instead of choosing a particular action, chooses a probability distribution on the set of available actions. This is required in order to apply minimax theorems which generally gives the existence of saddle-point equilibrium. For non-zero sum games the main aim is to show the existence of Nash-equilibrium. A Nash-equilibrium is again a tuple of control policies such that if one controller unilaterally deviates from the the Nash-equilibrium tuple then his/her payoff will be worse as compared to the payoff under the Nash-equilibrium policy. Like in zero-sum games, the analysis of non-zero sum games also require the use of relaxed control framework. The main technique followed in literature in order to establish the existence of a Nash-equilibrium is to show that an appropriately defined set valued map has a fixed point. This is achieved by invoking Fan's fixed point theorem.

For partially observable control problems, the popular technique in literature is to convert it to an equivalent completely observable problem. This is done by replacing the unobservable component with its conditional distribution given the available history. Suitable assumptions are made on the original partially observable model, so that the equivalently defined completely observable problem can be solved. Then the solution of the completely observable problem is used to solve the partially observable problem.

1.1 Literature Review

We now provide a brief survey of the various stochastic control problems, both single and multiplayer, that has been considered in the literature. The study of discrete-time Markov decision processes(MDP) began quite a long time back. For one of the earlier literature we refer to Howard [1960]. After that there has been a lot of work on discrete-time MDP, for both finite and infinite horizons, for countable as well as uncountable state spaces. For detailed and upto date literature on discrete-time MDP we refer to the books Hernández-Lerma and Lasserre [1996]; Bäuerle and Rieder [2011] and references there in. Continuous-time Markov decision processes have also been widely investigated in literature. There are works on both bounded and unbounded transition rates. See Guo and Hernández-Lerma [2009] and references there in for countable state space literature, Piunovskiy and Zhang [2020]

and references there in for Borel state space literature. There are also ample amount of literature on semi-Markov decision process, see Lippman [1973]; Federgruen et al. [1979]; Bhattacharya and Majumdar [1989] and references there in.

The literature described above corresponds to risk-neutral control. To overcome the shortcoming of risk-neutral criterion, the two popular approaches that are prevalent in literature are probability criterion approach and risk-sensitive approach. In discrete-time one controller setup probability criterion has been considered by several authors, see White [1993]; Bouakiz and Kebir [1995]; Wu and Lin [1999]; Kira et al. [2012]; Sakaguchi and Ohtsubo [2013]. In the continuous-time setup, for semi-Markov processes probability criterion has been investigated in Sakaguchi and Ohtsubo [2010]; Huang et al. [2011, 2013]. Probability criterion in continuous-time Markov chains has been studied in Huo et al. [2017]; Huo and Guo [2020]; Huo and Wen [2021]. Since the pioneering work in Howard and Matheson [1972], risk-sensitive control problems has been widely studied in literature. For discrete time Markov chains risk-sensitive criterion has been considered in Chung and Sobel [1987]; Hernández-Hernández and Marcus [1996]; Borkar and Meyn [2002]; Jaśkiewicz [2007]; Bäuerle and Rieder [2014]. For diffusions, see Whittle [1990]; Fleming and McEneaney [1995]; Nagai [1996]; Menaldi and Robin [2005]; Biswas et al. [2010]; Basu and Ghosh [2012]; Arapostathis and Biswas [2018]. The works on risk-sensitive control of continuous-time Markov chains include Suresh Kumar and Pal [2013]; Ghosh and Saha [2014]; Guo and Liao [2019]; Guo and Zhang [2019]; Wei and Chen [2019a]; Biswas and Pradhan [2022]. The literate on risk-sensitive control problems for semi-Markov processes is comparatively few. In Chávez-Rodríguez et al. [2016] the authors consider risk-sensitive control of semi-Markov processes with average cost criterion. There the state space is assumed to be finite and the sojourn time distributions are assumed to have a compact support. In Huang et al. [2018], the authors consider risk-sensitive control of semi-Markov processes on a fixed finite horizon $[0, T]$.

The work on stochastic games was pioneered by Shapley in Shapley [1953]. Since then there has been a lot of study in both zero and non-zero sum games. For a survey on discrete-time zero-sum games see Vrieze [1989]. For works on discrete-time non-zero sum games see Parthasarathy and Sinha [1989]; Ghosh and Bagchi [1998] and references there in. Stochastic games for continuous-time Markov chain has also been studied by a lot of authors for reference see Guo and Hernández-Lerma [2003, 2005, 2007] and references there in. Semi-Markov games have also been considered in literature, see Lal and Sinha [1992]; Jaśkiewicz [2002]; Luque-Vásquez [2002] and references there in. All these literature are for risk-neutral set up.

There also has been a lot of work on risk-sensitive stochastic games. Risk sensitive games for discrete time Markov chains has been studied by several authors, see for instance Basu and Ghosh [2014]; Bäuerle and Rieder [2017b]; Cavazos-Cadena and Hernández-Hernández [2019] for zero-sum games and Basu and Ghosh [2018]; Wei and Chen [2019b, 2021] for non-zero sum games. Risk-sensitive games for continuous-time diffusions has been studied in Biswas and Saha [2020]; Ghosh and Pradhan [2020]; Ghosh et al. [2021]. Similarly, risk-sensitive games for continuous-time Markov chains has been studied in Ghosh et al. [2016]; Wei [2018]; Golui et al. [2022]. Before the work of this thesis the literature on stochastic games with probability criterion was rather restricted. Discrete-time zero sum game with probability criterion has been studied in Huang et al. [2017] and non-zero sum game has been explored in Huang and Guo [2020].

All the literature described above pertains to completely observable case. Risk-sensitive control problems for discrete-time Markov chains with partial observation has been studied in James et al. [1994]; Di Masi and Stettner [1999]; Bäuerle and Rieder [2017a]. While for multiplayer setup, partially observable risk-neutral games has been studied in Ghosh et al. [2004]; Ghosh and Goswami [2006, 2008]; Saha [2014]. To the best of our knowledge, before this thesis there was no existing literature on partially observable risk-sensitive games.

1.2 Outline of the Thesis

Against the backdrop of the existing literature, we now outline the contributions of this thesis.

In Chapter 2, we consider a zero-sum semi-Markov game with a probability criterion. The state space is assumed to be Borel and action spaces are finite. The players start with an initial fixed reward level. Player 1 wishes to maximise the probability that the accumulated reward will exceed the pre fixed level before the failure of the system, while player 2 aims to minimise the same. Thus this is a two player stochastic optimization problem over a random time horizon. This work thus extends Huang et al. [2011] to the stochastic game setup, while it generalizes Huang et al. [2017] to the continuous-time framework. Here we establish that the zero-sum semi-Markov game with probability criterion has a value. We also characterise the value as the unique solution to a pair of Shapley equations. We also show the existence of optimal policies for both the players. The content of this Chapter is based on the published article Bhabak et al. [2022].

In Chapter 3, we study a zero-sum stochastic game for continuous-time Markov chain. The state space is assumed to be denumerable and the transition rates are possibly unbounded. The game is investigated under the probability criterion, wherein, player 1 tries to maximise the probability that his/her accumulated reward will exceed a pre-determined level before the system reaches a given target set, while player 2 wishes to minimise the same. Although, in general continuous time Markov chain can be considered as a special case of semi-Markov process, but the model considered in Chapter 3 is not a special case of the one considered in Chapter 2. This is because the non-explosivity condition assumed in Chapter 2 forces the transition rates to be bounded in the case of continuous-time Markov chains. But in Chapter 3 we deal with unbounded transition rates. Under suitable assumptions, we show that the value of the zero-sum game exists and is given by the unique solution of an appropriate pair of Shapley equations. Using the Shapley equations we also prescribe a saddle point equilibrium. This work generalizes the works in Huo et al. [2017]; Huo and Guo [2020] to the case of zero-sum game. The content of this Chapter is based on the published article Bhabak and Saha [2021].

In Chapter 4, we consider risk-sensitive optimization problems for semi-Markov processes on a countable state space. We consider non-negative running cost function, which is also assumed to be bounded above. The cost functional also includes the discount factor. The aim of the decision maker is to minimize the expected utility of the discounted cost accumulated over an infinite time horizon. In order to solve the infinite horizon problem we first consider an auxiliary random finite horizon problem, where the optimization is upto the Nth jump time of the process. The finite horizon problem can also be of independent interest. Then using the finite horizon problem we solve the infinite horizon problem via an appropriate limiting argument. We characterize the infinite horizon value function as the unique fixed point of an appropriate operator. This work generalizes the work on discrete-time Markov chains in Bäuerle and Rieder [2014] to the semi-Markov case. It also complements the risk-sensitive average optimality problem for semi-Markov processes considered in Chávez-Rodríguez et al. [2016] and finite horizon problem in Huang et al. [2018]. The content of this Chapter is based on the published article Bhabak and Saha [2022b].

In Chapter 5, we consider both zero and non-zero sum risk-sensitive average criterion games for semi-Markov processes. The state space is assumed to be finite and action spaces are Borel. We also assume that the sojourn times are supported on a fixed compact interval. Under general continuity-compactness assumptions and an additional assumption of irreducibility, we show that the zero-sum game admits a value. We also prescribe a saddle point

equilibrium which is given by minimizing and maximizing selectors of a pair of optimality equations. For the non-zero sum game problem, under certain additional assumptions we show the existence of a Nash equilibrium. In the non-zero sum case the main step involves showing the existence of solution of a coupled system of equations. In the analysis of both the zero-sum and non-zero sum games, risk sensitive games for discrete-time Markov chains serve as an important intermediate step. This work extends the work in Chávez-Rodríguez et al. [2016] to the case of both zero and non-zero sum games. The content of this Chapter is based on the published article Bhabak and Saha [2023].

In Chapter 6, we consider a discrete time zero-sum game where the state process evolves like a controlled Markov chain. We also assume that the state has two components one of which is observable while the other is not observable. In the optimization criterion we consider general utility function and discounted reward/cost. Thus the optimization criteria considered subsumes the classical risk-sensitive case. Player 1 is assumed to be the maximizer and player 2 the minimizer. Both finite and infinite horizon problems are investigated. In both cases we show that the game has value and also establish the existence of optimal policies for both players. Like in the risk-neutral case here also we convert the partially observable game into an equivalent completely observable game. But the difference with the risk-neutral case is that here we need to keep track of the accumulated cost as well. Since in the considered model the reward/cost function is assumed to depend on the unobservable component, so we need to consider the joint conditional distribution of the unobservable state component and the accumulated reward/cost. To the best of our knowledge this is the first work on partially observable risk-sensitive games. This work thus generalizes the work in Bäuerle and Rieder [2017a] to the case of multiple controllers. The content of this Chapter is based on the communicated article Bhabak and Saha [2022a].

Finally in Chapter 7, we discuss some of the open problems that can be pursued as a followup of this thesis.

1.3 Some Preliminaries

In this thesis, in three chapters, the underlying process is semi-Markov process. A semi-Markov process is a continuous-time pure jump process, which generalizes continuous-time Markov chains. The main generalization pertains to the fact that for semi-Markov processes the sojourn time distributions in each state are assumed to be some general non-negative dis-

tribution and not necessarily exponential as is the case with continuous-time Markov chains. This flexibility is very useful in modelling real scenarios, because memoryless property of sojourn times is seldom true in practice. Since the sojourn times are general, given the current time, the future evolution of the process depends both on the current state as well as the amount of time the process has been in that state. If E is the Borel state space of the process, then the evolution of semi-Markov process is described via a stochastic kernel $Q(\cdot,\cdot|\cdot)$ on $\mathbb{R}_+ \times E$ given E. Thus for a given $t \in \mathbb{R}_+$ and a Borel subset S of E, $Q(t,S|\cdot)$ is a measurable function of E equipped with the Borel $\sigma-$algebra. Given $x \in E$, $Q(\cdot,\cdot|x)$ is a probability measure on $\mathbb{R}_+ \times E$. For a fixed $x \in E$, $Q(\cdot,E|x)$ is the distribution function of the of the sojourn time in state x. For a given $t \in \mathbb{R}_+$, $Q(t,\cdot|x))$ is a sub-probability measure on E. The distribution of the next step is given by $Q(\infty,\cdot|x)$. Thus for a given $t \in \mathbb{R}_+$ and $S \subset E$, $Q(t,S|x))$ is the joint probability that the sojourn time in state x will be less than or equal to t and the next transition will be in S. If the state space is countable, there is another equivalent way of describing the evolution of a semi-Markov process. If E is the countable state space then the evolution is described by two quantities. First the transition probability matrix $(p_{ij})_{i,j\in E}$ of the embedded discrete-time Markov chain which keeps track of the states at successive jump times. The other quantity is $(F_{ij}(\cdot))_{i,j\in E}$, which is the conditional distribution function of the sojourn time of the process given that the process has just entered i and the next transition will be into state j. Thus in this case the joint distribution as in the above description is given by $Q(t,j|i) = F_{ij}(t)p_{ij}$. In the thesis we work with controlled semi-Markov processes for which the above quantities also depend on additional control parameter(s). The quantities involving control parameters are defined in the respective chapters.

Now we state few important theorems which we use in the upcoming chapters.

Theorem 1.3.1 (Ionescu-Tulcea's Theorem, Bäuerle and Rieder [2011]). *Let E be a borel space and ν be a probability measure on E and Q_n a sequence of stochastic kernels. Then there exists a unique probability measure $\mathbb{P}_\nu$ on E^∞ such that*

$$\mathbb{P}(B_0 \times ... \times B_N \times E \times ...) = \int_{B_0} ... \int_{B_N} Q_{N-1}(dx_N|x_{N-1})...Q_0(dx_1|x_0)\nu(dx_0)$$

for every measurable rectangle set $B_0 \times ... \times B_N \in E_{N+1}$.

Theorem 1.3.2 (Measurable Selection Theorem, Arapostathis et al. [1993]). *Let V and W*

be two Borel spaces. Let Φ *be a compact-valued measurable set function from* V *to* W*. Let* $f : \text{Graph}(\Phi) \to \mathbb{R}$ *be a measurable function, such that for each* $v \in V$*,* $f(v, \cdot)$ *is lower semi-continuous on* $\Phi(v)$*, then there exists a measurable function* $\phi^* : V \to W$ *such that* $\phi^*(v) \in \Phi(v)$ *and*

$$f(v, \phi^*(v)) = \min_{w \in \Phi(v)} \{f(v, w)\} \quad \forall v \in V\,.$$

Theorem 1.3.3 (Fan's Minimax Theorem, Fan [1952])**.** *Let* L_1 *and* L_2 *be two locally convex topological spaces and let* K_1*,* K_2 *be two compact convex sets in* L_1, L_2*, respectively. Let* f *be a real-valued continuous function on* $K_1 \times K_2$*. If, for any* $x_0 \in K_1$ *and* $y_0 \in K_2$*, the sets* $\{x \in K_1 | f(x, y_0) = \max_{\xi \in K_1} f(\xi, y_0)\}$ *and* $\{y \in K_2 | f(x_0, y) = \min_{\eta \in K_2} f(x_0, \eta)\}$ *are convex, then*

$$\max_{x \in K_1} \min_{y \in K_2} f(x, y) = \min_{y \in K_2} \max_{x \in K_1} f(x, y)\,.$$

Theorem 1.3.4 (Fan's Fixed-point Theorem, Fan [1952])**.** *Let* L *be a locally convex topological space and* K *a compact convex set in* L*. Let* $\mathcal{A}(K)$ *be the collection of all closed convex subsets of* K*. Further, let* $f : K \to \mathcal{A}(K)$ *be a upper semi-continuous map, i.e., for any point* $x_0 \in K$ *and any open set* U *in* K *such that* $f(x_0) \in U$*, there exists a neighbourhood* W *of* x_0 *such that* $f(x) \in U$ *for all* $x \in W$*. Then, there exists* $x^* \in K$ *such that* $x^* \in f(x^*)$*.*

In the setting of metric spaces, a function $f : K \to \mathcal{A}(K)$ is said to be upper semi-continuous at $x \in K$ if $x_n \to x$ and $y_n \to y$ with $y_n \in f(x_n)$, then $y \in f(x)$. We will be using this definition of upper semi-continuity.

CHAPTER 2

Zero-Sum Semi-Markov Games with Probability Criterion

In this Chapter we consider a semi-Markov zero-sum stochastic game with general state and finite action spaces. The performance is analyzed via a probability criterion. Under suitable assumptions, we establish the existence of value of the game and also characterize it through an optimality equation. In the process we also prescribe a saddle point equilibrium. The Chapter is organized as follows: In section 1 we describe the model and the optimization problem. In section 2 we prove our main result. Finally, in section 3 we provide an illustrative example. This Chapter is based on Bhabak et al. [2022].

2.1 Model Description and Problem Formulation

The semi-Markov stochastic game model that we are interested in is given by a five tuple:

$$\{E, D, (A(x) \subset A, B(x) \subset B, x \in E), Q(.,.|x,a,b), r(x,a,b)\}, \tag{2.1.1}$$

where the individual entities have the following interpretation:

1. E is the state space, assumed to be a Borel space endowed with a Borel σ-algebra $\mathcal{E}$;
2. $D \in \mathcal{E}$ is a given target set, can be thought of as the set of all bad states of the system,

with $D^c := E \setminus D$ denoting the complement of D with respect to E;

3. A, B are the action spaces, which are assumed to be countable. $A(x) \subset A$ and $B(x) \subset B$ are finite sets of admissible actions in state $x \in E$ for player 1 and 2 respectively;

4. Let $K := \{(x, a, b) | x \in E, a \in A(x), b \in B(x)\}$ be the set of all admissible state-action pairs. The function $Q(., .|x, a, b)$ is a stochastic kernel on $\mathbb{R}_+ \times E$ given K, which describes the transition mechanism of the controlled semi-Markov process (SMP). It is assumed that
i) $Q(., S|x, a, b)$, for any fixed $S \in \mathcal{E}$ and $(x, a, b) \in K$, is a non-decreasing, right continuous real valued function on $\mathbb{R}_+$ such that $Q(0, S|x, a, b) = 0$;
ii) $Q(t, .|x, a, b)$, for each fixed $t \in \mathbb{R}_+$, is a sub-stochastic kernel on E given K; and
(iii) $\mathbb{P}(.|x, a, b) := Q(\infty, .|x, a, b)$ is a stochastic kernel on E given K. If actions $a \in A(x)$ and $b \in B(x)$ are selected in state x by player 1 and 2 respectively, then $Q(t, S|x, a, b)$ is the joint probability that the sojourn time in state x will be less than or equal to $t \in \mathbb{R}_+$, and the next state will belong to $S \in \mathcal{E}$;

5. Lastly, the measurable function $r : K \to \mathbb{R}_+$ denotes the reward rate (representing profit for player 1 and cost for player 2), which satisfies $r(x, a, b) > 0$ for every $x \in D^c$ and $a \in A(x)$, $b \in B(x)$.

We now describe the evolution of the controlled SMP. Suppose that the system is in state x_0 at time $t = 0$, and the decision makers has a common profit goal λ_0 in mind (represents reward level for player 1, and cost level for player 2, respectively), then player 1 chooses an action a_0 from $A(x_0)$ and player 2 selects an action b_0 from $B(x_0)$ based on the system state x_0 and profit goal λ_0. As a consequences of these action choices, the system remains in x_0 until time t_1, at which point the system state changes to x_1 according to the transition law $Q(dt_1, dx_1|x_0, a_0, b_0)$. At time t_1, a reward $r(x_0, a_0, b_0)t_1$ is generated which represents profit for player 1 and cost for player 2 respectively and thus the decision makers are left with a remaining profit goal $\lambda_1 := \lambda_0 - r(x_0, a_0, b_0)t_1$. Based on the current state x_1, the jump time t_1, the current profit goal λ_1 as well as the previous state x_0, previous profit goal λ_0 and previous actions a_0 and b_0, a second action a_1 from $A(x_1)$ and b_1 from $B(x_1)$ are chosen by player 1 and 2 respectively and the same sequence of events repeats. The game evolves in this way, and hence we obtain an admissible history h_n of the controlled SMP up to the nth

decision epoch, i.e.,

$$h_n = (0, x_0, \lambda_0, a_0, b_0, ..., t_{n-1}, x_{n-1}, \lambda_{n-1}, a_{n-1}, b_{n-1}, t_n, x_n, \lambda_n),$$

where, $0 = t_0 < t_1 < \ldots < t_n$, $(x_m, a_m, b_m) \in K$, $\lambda_0 \in \mathbb{R}_+$, $\lambda_{m+1} := \lambda_m - r(x_m, a_m, b_m)(t_{m+1} - t_m)$, for $m = 0, 1, \ldots, n-1$ and $x_n \in E$. Let H_n denote the set of all admissible histories h_n of the system up to the nth decision epoch, where H_n is endowed with a Borel σ-algebra.

Remark 2.1.1. *Here the parameters λ_n denote the remaining reward levels at the nth decision epochs for the decision makers, which would affect the action selection of the decision makers. In particular, the case of "$\lambda_n < 0$" means that the decision maker's reward level is achieved.*

Next, we define policies, which specifies a decision rule for the decision makers to select actions.

Definition 2.1.2. *A randomized history dependent policy or simply a policy for player 1 is a sequence $\pi^1 = \{\pi_n^1 \ : n \geq 0\}$ of stochastic kernels π_n^1 on A given H_n such that*

$$\pi_n^1(A(x_n)|h_n) = 1 \quad \forall\ h_n \in H_n, n = 0, 1, ...$$

A randomized history dependent policy for player 2 can be defined analogously.

Let us denotes the set of all policies for player i by Π_i for $i = 1, 2$.

Notation: Let Φ_1 denote the set of all stochastic kernels ψ on A given $E \times \mathbb{R}$ satisfying $\psi(A(x)|x, \lambda) = 1$, for any $(x, \lambda) \in E \times \mathbb{R}$.

Definition 2.1.3. *(a) A policy $\pi = \{\pi_n\}$ is said to be randomized Markov for player 1 if there is a sequence $\{\psi_n\}$ of stochastic kernels $\psi_n \in \Phi_1$, such that $\pi_n(.|h_n) = \psi_n(.|x_n, \lambda_n)$ for every $h_n \in H_n$ and $n \geq 0$. In this case we write it as $\pi = \{\psi_n\}$.*
(b) A randomized Markov policy $\pi = \{\psi_n\}$ is said to be randomized stationary for player 1 if ψ_n is independent of n. By an abuse of notation we will sometimes denote a randomized stationary policy by ψ.

Similar policies can be defined for player 2. We denote by Π_i, Π_i^{RM}, Π_i^{RS} the families of all randomized history dependent, randomized Markov and stationary policies, respectively for player i, where $i = 1, 2$. Obviously, $\Phi_i = \Pi_i^{RS} \subset \Pi_i^{RM} \subset \Pi_i$.

For each $(s, x, \lambda) \in \mathbb{R}_+ \times E \times \mathbb{R}$ and $\pi^1 \in \Pi_1$ and $\pi^2 \in \Pi_2$ by the well-known Tulcea's Theorem, there exist a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}^{\pi^1,\pi^2}_{(s,x,\lambda)})$ and stochastic process $\{S_n, J_n, \lambda_n, A_n, B_n\}$ such that, for each $t \in \mathbb{R}_+$, $S \in \mathcal{E}$, $a \in A$, $b \in B$ and $n \geq 0$,

$$\mathbb{P}^{\pi^1,\pi^2}_{(s,x,\lambda)}(S_0 = s, J_0 = x, \lambda_0 = \lambda) = 1, \tag{2.1.2}$$

$$\mathbb{P}^{\pi^1,\pi^2}_{(s,x,\lambda)}(A_n = a, B_n = b|h_n) = \pi^1_n(a|h_n)\pi^2_n(b|h_n), \tag{2.1.3}$$

$$\mathbb{P}^{\pi^1,\pi^2}_{(s,x,\lambda)}(S_{n+1} - S_n \leq t, J_{n+1} \in S|h_n, a_n, b_n) = Q(t, S|x_n, a_n, b_n), \tag{2.1.4}$$

where S_n, J_n, $\lambda_n := \lambda_{n-1} - r(J_{n-1}, A_{n-1}, B_{n-1})(S_n - S_{n-1})$, A_n, B_n denote the nth decision epoch, the state, the reward level and the actions chosen by player 1 and 2 at the nth decision epoch. Equation (2.1.3) highlights the fact that the players choose their respective actions independent of each other. The expectation operator with respect to $\mathbb{P}^{\pi^1,\pi^2}_{(s,x,\lambda)}$ is denoted by $\mathbb{E}^{\pi^1,\pi^2}_{(s,x,\lambda)}$. For simplicity, $\mathbb{P}^{\pi^1,\pi^2}_{(0,x,\lambda)}$ and $\mathbb{E}^{\pi^1,\pi^2}_{(0,x,\lambda)}$ is denoted by $\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}$ and $\mathbb{E}^{\pi^1,\pi^2}_{(x,\lambda)}$. Without loss of generality we suppose that $S_0 = 0$.

In applications, it is natural to avoid the possibility of an infinite number of jumps within a finite time. For that purpose, we impose the following assumption.

Assumption 2.1.4. $\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(\lim_{n\to\infty} S_n = \infty) = 1$, *for all* $(x, \lambda) \in E \times \mathbb{R}$, $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$.

The following provides a sufficient condition for Assumption 2.1.4.

Proposition 2.1.5. *If there exist* $\delta > 0$ *and* $\epsilon > 0$ *such that*

$$Q(\delta, E|x, a, b) \leq 1 - \epsilon \quad \forall\ (x, a, b) \in K, \tag{2.1.5}$$

then, Assumption 2.1.4 holds.

Proof. Using the properties (2.1.2)-(2.1.4) and condition (2.1.5), it can be easily shown using induction that

$$\mathbb{E}^{\pi^1,\pi^2}_{(x,\lambda)}(e^{-S_n}) \leq (1 - \epsilon + \epsilon e^{-\delta})^n \quad \forall\ (x, \lambda) \in E \times \mathbb{R}, (\pi^1, \pi^2) \in \Pi_1 \times \Pi_2 \quad \text{and } n \geq 1.$$

Now fix any $t > 0$. By Markov inequality, we get,

$$\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(S_n \leq t) = \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(e^{-S_n} \geq e^{-t}) \leq e^t\, \mathbb{E}^{\pi^1,\pi^2}_{(x,\lambda)}(e^{-S_n}),$$

which gives $\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(S_n \leq t) \leq e^t\,(1-\epsilon+\epsilon e^{-\delta})^n$. Note that $(1-\epsilon+\epsilon e^{-\delta}) < 1$, and hence

$$\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(\lim_{n\to\infty} S_n \leq t) = \lim_{n\to\infty} \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(S_n \leq t) = 0,$$

which implies the desired conclusion. □

We now define a continuous time state action process $\{Z(t), U_1(t), U_2(t), t \in \mathbb{R}_+\}$ by

$$Z(t) = J_n, \quad U_1(t) = A_n, \quad U_2(t) = B_n \quad \text{for} \quad S_n \leq t < S_{n+1},\ t \in \mathbb{R}_+ \text{ and } n \geq 0\,.$$

Obviously, these processes depend on the policies π^1, π^2 of the players, but for economy of notation we don't include them in the notation.

Definition 2.1.6. *The stochastic process $\{Z(t), U_1(t), U_2(t), t \in \mathbb{R}_+\}$ is called a controlled semi-Markov process.*

Let $\mathcal{T}_D$ be the first hitting time of the target set D of the process $\{Z(t), t \geq 0\}$, i.e.,

$$\mathcal{T}_D := \inf\{t \geq 0 | Z(t) \in D\} \quad (\text{with } \inf\emptyset := +\infty)\,.$$

The performance function U^{π^1,π^2} of the semi-Markov stochastic game under consideration is defined by

$$U^{\pi^1,\pi^2}(x,\lambda) = \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\int_0^{\mathcal{T}_D} r(Z(t), U_1(t), U_2(t))dt > \lambda\right) \quad \forall\ (x,\lambda) \in E \times \mathbb{R}, (\pi^1,\pi^2) \in \Pi_1 \times \Pi_2. \tag{2.1.6}$$

Thus from player 1's perspective (2.1.6) quantifies the chance that the accumulated reward/profit will exceed the level λ before the failure of the system, while from player 2's perspective (2.1.6) measures the risk that the cost level will exceed λ until the failure of the system, when the pair of policies (π^1, π^2) are being used by the players.

To introduce our optimality problem, we also need the following functions:

$$I(x,\lambda) = \sup_{\pi^1\in\Pi_1}\inf_{\pi^2\in\Pi_2} U^{\pi^1,\pi^2}(x,\lambda)$$

$$L(x,\lambda) = \inf_{\pi^2\in\Pi_2}\sup_{\pi^1\in\Pi_1} U^{\pi^1,\pi^2}(x,\lambda)$$

defined on $E\times\mathbb{R}$. The function $I(\cdot)$ is called the lower value and $L(\cdot)$ is the upper value of the game, respectively. Clearly, $I(x,\lambda)\leq L(x,\lambda)$ for every $(x,\lambda)\in E\times\mathbb{R}$.

Definition 2.1.7. *If $I(x,\lambda) = L(x,\lambda)$ for every $(x,\lambda)\in E\times\mathbb{R}$, then we call the common function the value of the game, which is denoted by V.*

Here player 1 is interested in maximizing $U^{\pi^1,\pi^2}(\cdot,\cdot)$ over $\pi^1\in\Pi_1$ for each $\pi^2\in\Pi_2$, and player 2 wants to minimize $U^{\pi^1,\pi^2}(\cdot,\cdot)$ over $\pi^2\in\Pi_2$ for each $\pi^1\in\Pi_1$. Thus, we are interested in finding a pair of optimal policies $(\pi^{*^1},\pi^{*^2})\in\Pi_1\times\Pi_2$ as below.

Definition 2.1.8. *Suppose that the value of the game V exists. A policy $\pi^{*^1}\in\Pi_1$ is said to be optimal for player 1 if,*

$$\inf_{\pi^2\in\Pi_2} U^{\pi^{*^1},\pi^2}(x,\lambda) = V(x,\lambda),\quad \forall(x,\lambda)\in E\times\mathbb{R}.$$

similarly, $\pi^{*^2}\in\Pi_2$ is said to be optimal for player 2 if,

$$\sup_{\pi^1\in\Pi_1} U^{\pi^1,\pi^{*^2}}(x,\lambda) = V(x,\lambda),\quad \forall(x,\lambda)\in E\times\mathbb{R}.$$

If $\pi^{*^k}\in\Pi_k$ is optimal for player k, then $(\pi^{*^1},\pi^{*^2})\in\Pi_1\times\Pi_2$ is called a pair of optimal policies, also known as a saddle point equilibrium.

Remark 2.1.9.

i) Using standard arguments as in Lemma 3.1 of Huang et al. [2017] it can be shown that, for any fixed $(\pi^1,\pi^2)\in\Pi_1\times\Pi_2$, we have that

$$\sup_{\pi'\in\Pi_1} U^{\pi',\pi^2}(x,\lambda) = \sup_{\pi'\in\Pi_1^{RM}} U^{\pi',\pi^2}(x,\lambda),$$

$$\inf_{\pi' \in \Pi_2} U^{\pi^1,\pi'}(x,\lambda) = \inf_{\pi' \in \Pi_2^{RM}} U^{\pi^1,\pi'}(x,\lambda),$$

which implies that it is sufficient to limit ourselves to $\Pi_1^{RM} \times \Pi_2^{RM}$ *in the upcoming arguments.*

ii) For all $(\pi^1,\pi^2) \in \Pi_1 \times \Pi_2$*, it is obvious that* $U^{\pi^1,\pi^2}(x,\lambda) = 1_{(-\infty,0)}(\lambda)$ *for every* $(x,\lambda) \in D \times \mathbb{R}$*, where* 1_C *is the indicator function on any set* C*. Therefore, in order to avoid triviality, we restrict our attention to the case of* $(x,\lambda) \in D^c \times \mathbb{R}$*.*

2.2 Main results

In this section we prove our main result, that the game has a value and also show the existence of a saddle point equilibrium. To that end, let $\mathcal{P}(X)$ be the family of probability measures on a metric space X, endowed with weak topology. Let $\mathcal{F}_m := \{H : D^c \times \mathbb{R} \to [0,1]$, such that $H(\cdot,\cdot)$ is Borel measurable on $D^c \times \mathbb{R}$ and $H(x,\lambda) = 1$ if $\lambda < 0$ for each $x \in D^c\}$. In addition, we define the operators $T^{\psi,\phi}$, T and $T^{a,b}$ on $\mathcal{F}_m$ as follows: for any $H \in \mathcal{F}_m$, $x \in D^c$, $\psi \in \mathcal{P}(A(x))$, $\phi \in \mathcal{P}(B(x))$ and $(\pi^1,\pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$, if $\lambda \geq 0$,

$$\begin{aligned}
T^{a,b}H(x,\lambda) :=& 1 - Q(\lambda/r(x,a,b), E|x,a,b) \\
&+ \int_{D^c} \int_0^{\lambda/r(x,a,b)} H(y, \lambda - r(x,a,b)t) Q(dt,dy|x,a,b), \\
T^{\psi,\phi}H(x,\lambda) :=& \sum_{a \in A(x)} \psi(a) \sum_{b \in B(x)} \phi(b) T^{a,b}H(x,\lambda), \qquad (2.2.1) \\
T^{\pi^1,\pi^2}H(x,\lambda) :=& T^{\pi^1(\cdot|x,\lambda),\pi^2(\cdot|x,\lambda)}H(x,\lambda), \\
TH(x,\lambda) :=& \sup_{\psi \in \mathcal{P}(A(x))} \inf_{\phi \in \mathcal{P}(B(x))} T^{\psi,\phi}H(x,\lambda), \qquad (2.2.2)
\end{aligned}$$

and $T^{a,b}H(x,\lambda) = T^{\psi,\phi}H(x,\lambda) = TH(x,\lambda) = 1$ for $\lambda < 0$.

Remark 2.2.1. *The operators defined above will be crucial in characterizing the value of the game and the pair of optimal policies. Also it is easy to see that the above operators are all monotone.*

Now for U^{π^1,π^2} as defined in (2.1.6), note that for each $(x,\lambda) \in D^c \times \mathbb{R}_+$ and $(\pi^1,\pi^2) \in$

$\Pi_1 \times \Pi_2$,

$$\begin{aligned} U^{\pi^1,\pi^2}(x,\lambda) &= \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\int_0^{\mathcal{T}_D} r(Z(t),U_1(t),U_2(t))dt > \lambda\right) \\ &= \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\sum_{m=0}^{\infty}\int_{S_m}^{S_{m+1}} 1_{\{\mathcal{T}_D>t\}} r(Z(t),U_1(t),U_2(t))dt > \lambda\right) \\ &= \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\sum_{m=0}^{\infty} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda\right) \\ &= \lim_{n\to\infty} \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\sum_{m=0}^{n} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda\right), \end{aligned} \tag{2.2.3}$$

where $T_{m+1} := S_{m+1} - S_m$ (with $T_0 = 0$) denote the sojourn times between two successive decision epochs. The second equality follows from Assumption 2.1.4, and the last equality follows from the non-negativity of the reward function and the continuity of probability measures. Motivated by (2.2.3), we define $U^{\pi^1,\pi^2}_{-1}(x,\lambda) := 1_{(-\infty,0)}(\lambda)$, and

$$U^{\pi^1,\pi^2}_{n}(x,\lambda) = \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\sum_{m=0}^{n} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda\right),$$

if $\lambda \geq 0$ and $U^{\pi^1,\pi^2}_{n}(x,\lambda) := 1$ otherwise for every $x \in D^c$ and $n \geq 0$. Clearly, $U^{\pi^1,\pi^2}_{n} \leq U^{\pi^1,\pi^2}_{n+1}$ for every $n \geq -1$, and $\lim_{n\to\infty} U^{\pi^1,\pi^2}_{n} = U^{\pi^1,\pi^2}$.

Lemma 2.2.2. *Suppose that Assumption 2.1.4 hold. For any $\pi^1 = \{\psi_0,\psi_1,...\} \in \Pi_1^{RM}$, $\pi^2 = \{\phi_0,\phi_1,...\} \in \Pi_2^{RM}$ define the shifted policies ${}^{(1)}\pi^1 = \{\psi_1,\psi_2,...\} \in \Pi_1^{RM}$ and ${}^{(1)}\pi^2 = \{\phi_1,\phi_2,...\} \in \Pi_2^{RM}$. Then, for all $n \geq -1$, we have*
(a) $U^{\pi^1,\pi^2}_{n} \in \mathcal{F}_m$ and $U^{\pi^1,\pi^2} \in \mathcal{F}_m$.
(b) $U^{\pi^1,\pi^2}_{n+1} = T^{\psi_0,\phi_0}U^{({}^{(1)}\pi^1,{}^{(1)}\pi^2)}_{n}$ and $U^{\pi^1,\pi^2} = T^{\psi_0,\phi_0}U^{({}^{(1)}\pi^1,{}^{(1)}\pi^2)}$. In particular $U^{\psi,\phi} = T^{\psi,\phi}U^{\psi,\phi}$ for every $(\psi,\phi) \in \Pi_1^{RS} \times \Pi_2^{RS}$.

Proof. (a) To show that $U^{\pi^1,\pi^2}_{n} \in \mathcal{F}_m$; it is enough to prove that $U^{\pi^1,\pi^2}_{n}(.,.)$ is Borel measurable on $D^c \times \mathbb{R}$. We establish this by induction. When $n = -1$, it is obviously true. Now assume $U^{\pi^1,\pi^2}_{n}(.,.)$ is Borel-measurable for some $n \geq -1$ and every $(\pi^1,\pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$.

It then follows that for any $\pi^1 = \{\psi_0, \psi_1, ...\} \in \Pi_1^{RM}$, and $\pi^2 = \{\phi_0, \phi_1, ...\} \in \Pi_2^{RM}$,

$$T^{\psi_0,\phi_0}U_n^{(1)\pi^1,(1)\pi^2}(x,\lambda) := \sum_{a\in A(x)} \psi_0(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda)\Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) + \int_{D^c}\int_0^{\lambda/r(x,a,b)} U_n^{(1)\pi^1,(1)\pi^2}(y, \lambda - r(x,a,b)t)Q(dt,dy|x,a,b)\Big]$$

is well defined and measurable in (x,λ). On the other hand, for $\lambda < 0$,

$$U_{n+1}^{\pi^1,\pi^2}(x,\lambda) = T^{\psi_0,\phi_0}U_n^{(1)\pi^1,(1)\pi^2}(x,\lambda) = 1,$$

while for $\lambda \geq 0$, we have

$$\begin{aligned}
U_{n+1}^{\pi^1,\pi^2}(x,\lambda) =& \mathbb{P}_{(x,\lambda)}^{\pi^1,\pi^2}\left(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda - r(J_0,A_0,B_0)T_1\right)\\
=& \mathbb{E}_{(x,\lambda)}^{\pi^1,\pi^2}\Bigg[\mathbb{P}_{(x,\lambda)}^{\pi^1,\pi^2}\Bigg(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda - r(J_0,A_0,B_0)T_1 |\\
&\qquad T_0,J_0,\lambda_0,A_0,B_0,T_1,J_1,\lambda_1 = \lambda_0 - r(J_0,A_0,B_0)T_1\Bigg)\Bigg]\\
=& \sum_{a\in A(x)} \psi_0(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda) \int_E\int_0^\infty Q(du,dy|x,a,b)\\
&\Bigg[\mathbb{P}_{(x,\lambda)}^{\pi^1,\pi^2}\Bigg(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda - r(x,a,b)u |\\
&\qquad T_0 = 0, J_0 = x, \lambda_0 = \lambda, A_0 = a, B_0 = b,\\
&\qquad T_1 = u, J_1 = y, \lambda_1 = \lambda - r(x,a,b)u\Bigg)\Bigg]\\
=& \sum_{a\in A(x)} \psi_0(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda)\Bigg[\int_E\int_{\lambda/r(x,a,b)}^\infty Q(du,dy|x,a,b)\times 1\\
&+ \int_E\int_0^{\lambda/r(x,a,b)} Q(du,dy|x,a,b)\\
&\times \mathbb{P}_{(y,\lambda-r(x,a,b)u)}^{(1)\pi^1,(1)\pi^2}\left(\sum_{m=0}^{n} 1_{\cap_{k=0}^m (J_k\in D^c)} r(J_m,A_m,B_m)T_{m+1} > \lambda - r(x,a,b)u\right)\Bigg]
\end{aligned}$$

$$= \sum_{a\in A(x)} \psi_0(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda)\Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) + \int_{D^c}\int_0^{\lambda/r(x,a,b)} Q(du,dy|x,a,b) \times U_n^{(1)\pi^1,(1)\pi^2}(y, \lambda - r(x,a,b)u)\Big)\Big],$$

where the third equality follows from properties (2.1.2)-(2.1.4), and the fourth equality is due to the Markov property of the policy pair (π^1, π^2) and properties (2.1.2)-(2.1.4) again. Hence,

$$U_{n+1}^{\pi^1,\pi^2} = T^{\psi_0,\phi_0} U_n^{(1)\pi^1,(1)\pi^2} \quad \forall (x,\lambda) \in D^c \times \mathbb{R},$$

and thus $U_{n+1}^{\pi^1,\pi^2}(.,.)$ is measurable. Therefore, by induction, $U_n^{\pi^1,\pi^2}(.,.)$ is measurable for every $n \geq -1$. Furthermore, since limit of measurable functions is again measurable, we have $\lim_{n\to\infty} U_n^{\pi^1,\pi^2} = U^{\pi^1,\pi^2} \in \mathcal{F}_m$.

(b) From the proof of (a), we have $U_{n+1}^{\pi^1,\pi^2} = T^{\psi_0,\phi_0} U_n^{(1)\pi^1,(1)\pi^2}$. Letting $n \to \infty$, by the dominated convergence theorem we obtain $U^{\pi^1,\pi^2} = T^{\psi_0,\phi_0} U^{(1)\pi^1,(1)\pi^2}$.
The last statement is obvious. □

Remark 2.2.3. *Following the proof of Lemma 2.2.2, it is easy to see that*

$$U^{(k)\pi^1,(k)\pi^2} = T^{\psi_k,\phi_k} U^{(k+1)\pi^1,(k+1)\pi^2} \tag{2.2.4}$$

holds for any ${}^{(k)}\pi^1 = \{\psi_{k+m}, m \geq 0\} \in \Pi_1^{RM}$ *and* ${}^{(k)}\pi^2 = \{\phi_{k+m}, m \geq 0\} \in \Pi_2^{RM}$ *with* $k = 0, 1, ...,$ ${}^{(0)}\pi^1 = \pi^1$ *and* ${}^{(0)}\pi^2 = \pi^2$.

Now in order to establish the existence of value of the game and saddle point equilibrium we need to impose the following assumption.

Assumption 2.2.4. $\mathbb{P}_{(x,\lambda)}^{\pi^1,\pi^2}(\mathcal{T}_D < \infty) = 1$ *for every* $(x,\lambda) \in D^c \times \mathbb{R}_+$ *and* $(\pi^1, \pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$.

Under Assumption 2.2.4, we have

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cap_{k=1}^{\infty}\{J_k \in D^c\}\right) = 1 - \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cup_{k=1}^{\infty}\{J_k \in D\}\right)\\
&= 1 - \sum_{n=1}^{\infty}\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(J_k \in D^c, 1 \le k \le n-1, J_n \in D\right)\\
&= 1 - \mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(\mathcal{T}_D < \infty) = 0.
\end{aligned}$$

Remark 2.2.5. *Thus Assumption 2.2.4 indicates that, no matter what the initial state is, what the level is and what the pair of randomized Markov policies are being followed, the system will eventually fail within finite time almost surely.*

Next we give a sufficient condition imposed on the primitive data of the model (2.1.1) which ensures Assumption 2.2.4. Intuitively, the system will eventually reach D if for each state in D^c there is a fixed positive probability of landing in D in the next transition.

Condition 1. *Suppose there exists some constant $\beta > 0$ satisfying $\mathbb{P}(D|x,a,b) \ge \beta$ for all $(a,b) \in A(x) \times B(x)$ and $x \in D^c$. Then Assumption 2.2.4 holds.*

Proof. We will use induction to show that

$$\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cap_{k=1}^{n}\{J_k \in D^c\}\right) \le (1-\beta)^n \tag{2.2.5}$$

for all $(x,\lambda) \in D^c \times \mathbb{R}$, $\pi^1 := \{\psi_k, k \ge 0\} \in \Pi_1^{RM}$, $\pi^2 := \{\phi_k, k \ge 0\} \in \Pi_2^{RM}$ and $n \ge 1$.

Indeed, for $n = 1$ and every $(x,\lambda) \in D^c \times \mathbb{R}$, we have the fact. Now assume that

$$\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cap_{k=1}^{n}\{J_k \in D^c\}\right) \le (1-\beta)^n$$

for every $(x,\lambda) \in D^c \times \mathbb{R}$ and some $n \geq 1$. Then, we see that

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cap_{k=1}^{n+1} \{J_k \in D^c\}\right) \\
&= \mathbb{E}^{\pi^1,\pi^2}_{(x,\lambda)}\left[1_{\{J_1\in D^c\}}...1_{\{J_n\in D^c\}}\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}(J_{n+1} \in D^c | T_0, J_0, \lambda_0, A_0, B_0, ..., T_n, J_n, \lambda_n)\right] \\
&= \mathbb{E}^{\pi^1,\pi^2}_{(x,\lambda)}\left[1_{\{J_1\in D^c\}}...1_{\{J_n\in D^c\}}\mathbb{P}^{(n)\pi^1,(n)\pi^2}_{(J_n,\lambda_n)}(J_1 \in D^c)\right] \\
&\leq (1-\beta)^{n+1},
\end{aligned}$$

where the second equality follows from the properties (2.1.2)-(2.1.4) and Markov property of policy (π^1,π^2), and the inequality is due to the induction hypothesis. Hence, by induction we have (2.2.5), which gives (upon taking limit over $n \to \infty$), $\mathbb{P}^{\pi^1,\pi^2}_{(x,\lambda)}\left(\cap_{k=1}^{\infty} \{J_k \in D^c\}\right) \leq \lim_{n\to\infty}(1-\beta)^n \to 0$, and thus implies the required Assumption. □

Before stating the next lemma, we need to introduce one more notation as below. Let $\bar{\mathcal{F}}_m := \{H : D^c \times \mathbb{R} \to [-1,1]$, satisfying $H(\cdot,\cdot)$ is Borel measurable on $D^c \times \mathbb{R}$ and for each $x \in D^c$, $H(x,\lambda) := 0$ for $\lambda < 0\}$. Define an operator on $\bar{\mathcal{F}}_m$ by

$$\bar{T}^{\psi,\phi}H(x,\lambda) := \sum_{a\in A(x)} \psi(a) \sum_{b\in B(x)} \phi(b) \int_{D^c}\int_0^\infty H(y, \lambda - r(x,a,b)t)Q(dt,dy|x,a,b),$$

for $\lambda \geq 0$ and $\bar{T}^{\psi,\phi}H(x,\lambda) := 0$ otherwise for each $(\psi,\phi) \in \Phi_1 \times \Phi_2$ and $x \in D^c$.

Lemma 2.2.6. *Suppose that Assumptions 2.1.4 and 2.2.4 are satisfied. Then for any function u in $\mathcal{F}_m$ and $(x,\lambda) \in D^c \times \mathbb{R}$, the following statements hold.*
(a) If $u(x,\lambda) \leq T^{\pi^1,\phi_k}u(x,\lambda)$ for all $k \geq 0$, for policies $\pi^1 \in \Pi_1^{RS}$ and $\bar{\pi}^2 = \{\phi_k, k \geq 0\} \in \Pi_2^{RM}$, then $u(x,\lambda) \leq U^{\pi^1,\bar{\pi}_2}(x,\lambda)$.
(b) If $u(x,\lambda) \geq T^{\psi_k,\pi^2}u(x,\lambda)$ for all $k \geq 0$, for policies $\pi^2 \in \Pi_2^{RS}$ and $\bar{\pi}^1 = \{\psi_k, k \geq 0\} \in \Pi_1^{RM}$, then $u(x,\lambda) \geq U^{\bar{\pi}^1,\pi_2}(x,\lambda)$.
(c) For every $(\pi^1,\pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$, $U^{\pi^1,\pi^2}(.,.)$ is the unique solution in $\mathcal{F}_m$ to the equation $H = T^{\pi^1,\pi^2}H$.

Proof. (a) Obviously, the assertion holds trivially for $\lambda < 0$. So, we only need to prove for $\lambda \geq 0$. On the other hand, for any $x \in D^c$, policies $\pi^1 \in \Pi_1^{RS}$, $\bar{\pi}^2 = \{\phi_k, k \geq 0\} \in \Pi_2^{RM}$ and

$n \geq 1$, we have

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(x,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}\right)\\
&= \mathbb{E}^{\pi^1,\bar{\pi}^2}_{(x,\lambda)}\left[\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(x,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}|T_0, J_0, \lambda_0, A_0, B_0, T_1, J_1, \lambda_1 = \lambda_0 - r(J_0, A_0, B_0)T_1\right)\right]\\
&= \sum_{a\in A(x)} \pi^1(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda) \int_E \int_0^\infty Q(du, dy|x, a, b) \times \left[\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(x,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}\right|\right.\\
&\left.T_0 = 0, J_0 = x, \lambda_0 = \lambda, A_0 = a, B_0 = b, T_1 = u, J_1 = y, \lambda_1 = \lambda - r(x, a, b)u\right)\Big]\\
&= \sum_{a\in A(x)} \pi^1(a|x,\lambda) \sum_{b\in B(x)} \phi_0(b|x,\lambda) \int_{D^c} \int_0^\infty \mathbb{P}^{\pi^1,(1)\bar{\pi}^2}_{(y,\lambda-r(x,a,b)u)}(\cap_{k=0}^{n}\{J_k \in D^c\}) Q(du, dy|x, a, b)\,.
\end{aligned} \tag{2.2.6}$$

On the other hand, it follows from (2.2.4) that $U^{\pi^1,(k)\bar{\pi}^2}(x,\lambda) = T^{\pi^1,\phi_k}U^{\pi^1,(k+1)\bar{\pi}^2}(x,\lambda)$ for all $k \geq 0$, which, together with the condition $u(x,\lambda) \leq T^{\pi^1,\phi_k}u(x,\lambda)$ and (2.2.6), for $\lambda \geq 0$, gives that

$$\begin{aligned}
&u(x,\lambda) - U^{\pi^1,\bar{\pi}^2}(x,\lambda) \leq \bar{T}^{\pi^1,\phi_0}[u(x,\lambda) - U^{\pi^1,(1)\bar{\pi}^2}(x,\lambda)]\\
&\leq \bar{T}^{\pi^1,\phi_0}\bar{T}^{\pi^1,\phi_1}...\bar{T}^{\pi^1,\phi_n}[u(x,\lambda) - U^{\pi^1,(n+1)\bar{\pi}^2}(x,\lambda)]\\
&\leq \sum_{a_0\in A(x)} \pi^1(a_0|x,\lambda) \sum_{b_0\in B(x)} \phi_0(b_0|x,\lambda) \int_{D^c}\int_0^\infty \sum_{a_1\in A(x_1)} \pi^1(a_1|x_1,\lambda_1) \sum_{b_1\in B(x_1)} \phi_1(b_1|x_1,\lambda_1)\\
&\int_{D^c}\int_0^\infty ... \sum_{a_{n-1}\in A(x_{n-1})} \pi^1(a_{n-1}|x_{n-1},\lambda_{n-1}) \sum_{b_{n-1}\in B(x_{n-1})} \phi_{n-1}(b_{n-1}|x_{n-1},\lambda_{n-1})\\
&\int_{D^c}\int_0^\infty \mathbb{P}^{\pi^1,(n)\bar{\pi}^2}_{(x_n,\lambda_{n-1}-r(x_{n-1},a_{n-1},b_{n-1})t_n)}(J_1 \in D^c) Q(dt_n, dx_n|x_{n-1}, a_{n-1}, b_{n-1})...\\
&Q(dt_2, dx_2|x_1, a_1, b_1) Q(dt_1, dx_1|x, a_0, b_0)
\end{aligned}$$

$$= \sum_{a_0 \in A(x)} \pi^1(a_0|x,\lambda) \sum_{b_0 \in B(x)} \phi_0(b_0|x,\lambda) \int_{D^c} \int_0^\infty \sum_{a_1 \in A(x_1)} \pi^1(a_1|x_1,\lambda_1) \sum_{b_1 \in B(x_1)} \phi_1(b_1|x_1,\lambda_1)$$

$$\int_{D^c} \int_0^\infty \ldots \sum_{a_{n-2} \in A(x_{n-2})} \pi^1(a_{n-2}|x_{n-2},\lambda_{n-2}) \sum_{b_{n-2} \in B(x_{n-2})} \phi_{n-2}(b_{n-2}|x_{n-2},\lambda_{n-2}) \int_{D^c} \int_0^\infty$$

$$\mathbb{P}^{\pi^1,(n-1)\bar{\pi}^2}_{(x_{n-1},\lambda_{n-2}-r(x_{n-2},a_{n-2},b_{n-2})t_{n-1})}(\cap_{k=1}^2\{J_k \in D^c\}) Q(dt_{n-1}, dx_{n-1}|x_{n-2}, a_{n-2}, b_{n-2}) \ldots$$

$$Q(dt_2, dx_2|x_1, a_1, b_1) Q(dt_1, dx_1|x, a_0, b_0)$$

$$= \ldots = \mathbb{P}^{\pi^1,\bar{\pi}^2}_{(x,\lambda)} \left(\cap_{k=1}^{n+1} \{J_k \in D^c\} \right)$$

for all $n \geq 0$. Letting $n \to \infty$ in the above inequality, using Assumption 2.2.4, we obtain that $u(x,\lambda) \leq U^{\pi^1,\bar{\pi}_2}(x,\lambda)$.

(b) The fact that $U^{(k)\bar{\pi}^1,\pi^2}(x,\lambda) = T^{\psi_k,\pi^2} U^{(k+1)\bar{\pi}^1,\pi^2}(x,\lambda)$ in (2.2.4) and the condition in this part yield that $U^{\bar{\pi}^1,\pi^2}(x,\lambda) - u(x,\lambda) \leq \bar{T}^{\psi_0,\pi^2}[U^{(1)\bar{\pi}^1,\pi^2}(x,\lambda) - u(x,\lambda)]$. Then, by similar arguments as in part (a) gives the desired result.

(c) Lemma 2.2.2(b), together with (a) and (b) of this lemma, gives the uniqueness of the solution $U^{\pi^1,\pi^2}(.,.)$ in $\mathcal{F}_m$ to the equation $H = T^{\pi^1,\pi^2} H$. □

The following lemma will be useful in prescribing a saddle point equilibrium for the game in hand.

Lemma 2.2.7. *For each $(x,\lambda) \in D^c \times \mathbb{R}_+$ and $u \in \mathcal{F}_m$, $T^{\psi,\phi}u(x,\lambda)$ is continuous in $(\psi,\phi) \in \mathcal{P}(A(x)) \times \mathcal{P}(B(x))$ with the operator $T^{\psi,\phi}u(x,\lambda)$ as in (2.2.1).*

Proof. By the finiteness of $A(x)$ and $B(x)$

$$1 - Q(\lambda/r(x,a,b), E|x,a,b) + \int_{D^c} \int_0^{\lambda/r(x,a,b)} H(y, \lambda - r(x,a,b)t) Q(dt, dy|x,a,b)$$

is a bounded continuous function on $A(x) \times B(x)$. Thus, by definition of weak convergence, for any sequence $\{(\psi_l, \phi_l)\} \subset \mathcal{P}(A(x)) \times \mathcal{P}(B(x))$ converging to (ψ,ϕ) with respect to the

weak topology, by letting $l \to \infty$ we get,

$$\begin{aligned}\lim_{l\to\infty} T^{\psi_l,\phi_l}u(x,\lambda) &= \sum_{a\in A(x)}\psi(a)\sum_{b\in B(x)}\phi(b)\Big[1-Q(\lambda/r(x,a,b),E|x,a,b)\\ &\quad+\int_{D^c}\int_0^{\lambda/r(x,a,b)} u(y,\lambda-r(x,a,b)t)Q(dt,dy|x,a,b)\Big]\\ &= T^{\psi,\phi}u(x,\lambda),\end{aligned}$$

which completes the proof. □

Remark 2.2.8. *Using Assumptions 2.1.4 and 2.2.4 and Lemma 2.2.7, we have by Fan's minimax theorem Fan [1953], for any $H \in \mathcal{F}_m$, $\lambda \geq 0$ and T as in (2.2.2),*

$$TH(x,\lambda) = \sup_{\psi\in\mathcal{P}(A(x))}\inf_{\phi\in\mathcal{P}(B(x))} T^{\psi,\phi}H(x,\lambda) = \inf_{\phi\in\mathcal{P}(B(x))}\sup_{\psi\in\mathcal{P}(A(x))} T^{\psi,\phi}H(x,\lambda)\,.$$

Let $u_{-1}(x,\lambda) := 1_{(-\infty,0)}(\lambda)$ and $u_n(x,\lambda) := Tu_{n-1}(x,\lambda)$ for each $(x,\lambda) \in D^c \times \mathbb{R}$ and $n \geq 0$. Now we state the main result of the Chapter.

Theorem 2.2.9. *Suppose Assumptions 2.1.4 and 2.2.4 hold. Then we have the following statements.*
(a) The limit $\lim_{n\to\infty} u_n(x,\lambda) := u^*(x,\lambda)$ exists and belongs to $\mathcal{F}_m$, moreover, u^* satisfies the Shapley equation $u^*(x,\lambda) = Tu^*(x,\lambda)$, i.e.,

$$\begin{aligned}u^*(x,\lambda) &= \sup_{\psi\in\mathcal{P}(A(x))}\inf_{\phi\in\mathcal{P}(B(x))}\sum_{a\in A(x)}\psi(a)\sum_{b\in B(x)}\phi(b)\Big[1-Q(\lambda/r(x,a,b),E|x,a,b)\\ &\quad+\int_{D^c}\int_0^{\lambda/r(x,a,b)} u^*(y,\lambda-r(x,a,b)t)Q(dt,dy|x,a,b)\Big] \qquad (2.2.7)\\ &= \inf_{\phi\in\mathcal{P}(B(x))}\sup_{\psi\in\mathcal{P}(A(x))}\sum_{a\in A(x)}\psi(a)\sum_{b\in B(x)}\phi(b)\Big[1-Q(\lambda/r(x,a,b),E|x,a,b)\\ &\quad+\int_{D^c}\int_0^{\lambda/r(x,a,b)} u^*(y,\lambda-r(x,a,b)t)Q(dt,dy|x,a,b)\Big] \qquad (2.2.8)\end{aligned}$$

for any $(x,\lambda) \in D^c \times \mathbb{R}_+$.
(b) There exists a pair of stationary policies $(\pi^{*^1},\pi^{*^2}) \in \Pi_1^{RS} \times \Pi_2^{RS}$ such that, for all

$(x, \lambda) \in D^c \times \mathbb{R}$,

$$u^*(x,\lambda) = T^{\pi^{*^1},\pi^{*^2}} u^*(x,\lambda) = \max_{\psi \in \mathcal{P}(A(x))} T^{\psi,\pi^{*^2}} u^*(x,\lambda) = \min_{\phi \in \mathcal{P}(B(x))} T^{\pi^{*^1},\phi} u^*(x,\lambda). \tag{2.2.9}$$

(c) $u^*(x,\lambda)$ *is the value of the game, and* $u^*(x,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(x,\lambda)$.
(d) $(\pi^{*^1}, \pi^{*^2}) \in \Pi_1^{RS} \times \Pi_2^{RS}$ *in (b) above is a saddle point equilibrium.*

Proof. (a) By Prohorov's Theorem, the finiteness of $A(x)$ and $B(x)$ implies the compactness of $\mathcal{P}(A(x))$ and $\mathcal{P}(B(x))$. Hence, lemma 2.2.7, and measurable selection theorem in Nowak [1985], gives that there exists $(\pi^1, \pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$ (which may depend on n) such that $u_n(x,\lambda) = Tu_{n-1}(x,\lambda) = T^{\pi^1,\pi^2} u_{n-1}(x,\lambda)$ for all $n \geq 0$, which shows the measurability of $u_n(\cdot,\cdot)$. Moreover, $u_n(x,\lambda) = 1$ for $\lambda < 0$. Therefore, $u_n \in \mathcal{F}_m$ for any $n \geq -1$. For $\lambda \geq 0$, we have

$$\begin{aligned} u_0(x,\lambda) &= \sup_{\psi \in \mathcal{P}(A(x))} \inf_{\phi \in \mathcal{P}(B(x))} \sum_{a \in A(x)} \psi(a) \sum_{b \in B(x)} \phi(b) \Big[1 - Q(\lambda / r(x,a,b), E | x,a,b) \Big] \\ &\geq u_{-1}(x,\lambda). \end{aligned}$$

Therefore, by the definition of $\{u_n, n \geq -1\}$ and monotonicity of the operator T, we have $u_{-1} \leq u_0 \leq ... \leq u_n ...$, i.e., $\{u_n, n \geq -1\}$ is a non-decreasing sequence, and converges to some function $u^* \in \mathcal{F}_m$.

Now, for $\lambda < 0$, $u^*(x,\lambda) = Tu^*(x,\lambda) = 1$. To show the Shapley equation for $\lambda \geq 0$, using the monotonicity again, we have $Tu^* \geq Tu_n = u_{n+1}$ for all $n \geq -1$, which shows that

$$Tu^* \geq u^*. \tag{2.2.10}$$

For the reverse inequality, it follows from the definition of the operator T that

$$\begin{aligned} u_{n+1}(x,\lambda) \geq & \inf_{\phi\in\mathcal{P}(B(x))} \sum_{a\in A(x)} \psi(a) \sum_{b\in B(x)} \phi(b) \Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) \\ & + \int_{D^c} \int_0^{\lambda/r(x,a,b)} u_n(y, \lambda - r(x,a,b)t) Q(dt, dy|x,a,b) \Big] \\ = & \sum_{a\in A(x)} \psi(a) \sum_{b\in B(x)} \phi_n^*(b) \Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) \\ & + \int_{D^c} \int_0^{\lambda/r(x,a,b)} u_n(y, \lambda - r(x,a,b)t) Q(dt, dy|x,a,b) \Big] \end{aligned} \tag{2.2.11}$$

for any $\psi \in \mathcal{P}(A(x))$, where the existence of $\phi_n^* \in \mathcal{P}(B(x))$ (may be dependent on ψ) is guaranteed by Lemma 2.2.7. By the compactness of $\mathcal{P}(B(x))$, without loss of generality, we suppose that $\phi_n^* \to \phi^* \in \mathcal{P}(B(x))$. Taking $n \to \infty$ in above equation, it follows from the extended Fatou's lemma (Lemma 8.3.7 in Hernández-Lerma and Lasserre [1999]) that,

$$\begin{aligned} u^*(i,\lambda) \geq & \sum_{a\in A(x)} \psi(a) \sum_{b\in B(x)} \phi^*(b) \Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) \\ & + \int_{D^c} \int_0^{\lambda/r(x,a,b)} u^*(y, \lambda - r(x,a,b)t) Q(dt, dy|x,a,b) \Big] \\ \geq & \inf_{\phi\in\mathcal{P}(B(x))} \Big\{ \sum_{a\in A(x)} \psi(a) \sum_{b\in B(x)} \phi(b) \Big[1 - Q(\lambda/r(x,a,b), E|x,a,b) \\ & + \int_{D^c} \int_0^{\lambda/r(x,a,b)} u^*(y, \lambda - r(x,a,b)t) Q(dt, dy|x,a,b) \Big] \Big\}. \end{aligned}$$

Since $\psi \in \mathcal{P}(A(x))$ is arbitrary, we get,

$$Tu^* \leq u^*,$$

which together with (2.2.10) gives $Tu^* = u^*$.

(b) Obviously, (2.2.9) holds for $\lambda < 0$. For each $\lambda \geq 0$, if we choose $\pi^{*^1} \in \Pi_1^{RS}$ as the outer maximizing selector in (2.2.7) and $\pi^{*^2} \in \Pi_2^{RS}$ as the outer minimizing selector in (2.2.8) then (2.2.9) follows.

(c) From (b) of this theorem, we have $u^*(x,\lambda) = T^{\pi^{*^1},\pi^{*^2}} u^*(x,\lambda)$. Thus by Lemma

2.2.6(c), we have $u^*(x,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(x,\lambda)$. Let π^{*^2} be fixed. For any policy $\pi^1 := \{\psi_n, n \geq 0\} \in \Pi_1^{RM}$, we get $\psi_n(.|x,\lambda) \in \mathcal{P}(A(x))$ for all $n \geq 0$ and $(x,\lambda) \in D^c \times \mathbb{R}$. Then, from (2.2.9), for any $n \geq 0$, we have

$$u^*(x,\lambda) \geq T^{\psi_n,\pi^{*^2}} u^*(x,\lambda),$$

which combined with Lemma 2.2.6(b) gives $u^*(x,\lambda) \geq U^{\pi^1,\pi^{*^2}}(x,\lambda)$ for all $\pi^1 \in \Pi_1^{RM}$. Therefore, $u^*(x,\lambda) \geq \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^{*^2}}(x,\lambda)$, while the converse inequality follows from $u^*(x,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(x,\lambda)$, which implies that, for any $(x,\lambda) \in D^c \times \mathbb{R}$,

$$u^*(x,\lambda) = \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^{*^2}}(x,\lambda) \geq \inf_{\pi^2 \in \Pi_2^{RM}} \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^2}(x,\lambda) = L(x,\lambda). \tag{2.2.12}$$

Using (2.2.9) again, a similar argument together with Lemma 2.2.6(a) shows that

$$u^*(x,\lambda) = \inf_{\pi^2 \in \Pi_2^{RM}} U^{\pi^{*^1},\pi^2}(x,\lambda) \leq \sup_{\pi^1 \in \Pi_1^{RM}} \inf_{\pi^2 \in \Pi_2^{RM}} U^{\pi^1,\pi^2}(x,\lambda) = I(x,\lambda) \tag{2.2.13}$$

Hence, by (2.2.12) and (2.2.13), $L(x,\lambda) = I(x,\lambda) = V(x,\lambda) = u^*(x,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(x,\lambda)$.

(d) This follows directly from both (2.2.12) and (2.2.13). □

2.3 Example

In this section we give an example to indicate potential situations where our model can be used. Consider an inventory which can have stock levels $\{0, 1, \ldots, M\}$. There are two parties. The manufacturer manufactures goods and the retailer orders goods. The state of the process under consideration is the stock level. At time 0, there is some initial stock level i_0. The manufacturer based on the initial stock level and his/her profit goal decides to manufacture $a_0 \in \{1, 2, \ldots, M-1\}$ units of stock. While the retailer pre-orders $b_0 \in \{1, 2, \ldots, M-1\}$ units of stock. The manufacturing takes a random amount of time with distribution function given by $F(\cdot|a_0)$. Suppose the manufacturing is over at time t_1 and the number of demands

that has arrived upto t_1 is k, then the new inventory level becomes i_1 given by

$$i_1 = (i_0 + a_0 - k)^+ \wedge M,$$

where we use the notation that for two real numbers a, b, $a^+ = \max\{a, 0\}$ and $a \wedge b = \min\{a, b\}$. Further, at time t_1 a reward equal to $(\bar{r}(i_0, b_0) + \alpha(i_0)(a_0 - b_0)^+)\, t_1$ is received by the manufacturer which is also the cost paid by the retailer, for certain positive real valued functions $\bar{r}(\cdot, \cdot)$ and $\alpha(\cdot)$. This continues until the stock level either hits 0 or M. Suppose demands arrive according to an independent Poisson process with rate 1. Thus, if Z_n is the number of demands which arrive at the nth decision epoch and A_{n-1} is the action chosen by the manufacturer at $(n-1)th$ decision epoch, then

$$p(k|a) = P(Z_n = k | A_{n-1} = a) = \int_0^\infty e^{-t} \frac{t^k}{k!} F(dt|a) \quad \forall k \geq 0.$$

Suppose $F(0|a) = 0$ for all $a \in \{1, 2, \ldots, M-1\}$. Thus in the notation of (2.1.1), $E = \{0, 1, \ldots, M\}, D = \{0, M\}, A(i) = B(i) = A = \{1, 2, \ldots, M-1\}$ for all $i \in E$. Here $r(i, a, b) = \bar{r}(i, b) + \alpha(i)(a - b)^+$ for all $(i, a, b) \in E \times A \times A$. The transition kernel Q has the following description: suppose that the process has just entered the state i and the action chosen by the manufacturer is a, then the distribution of the time of the next transition is given by $F(\cdot|a)$ and given that the transition happens at time t, the next state is $i + a - k$ with probability $e^{-t}\frac{t^k}{k!}$, when $i + a - k \in \{1, \ldots, M-1\}$. If $i + a - k \leq 0$, then the next state is 0, whereas if $i + a - k \geq M$, then the next state is M. For any state $i \in \{1, 2, \ldots, M-1\}$ and any action a of the manufacturer, the probability that the next state will be 0 is greater than $p(2M - 1|a)$, which is strictly positive. Thus, this example satisfies the assumptions of our model and thereby can analyzed using our results.

CHAPTER 3

Continuous-time Zero-Sum Games with Probability Criterion

In this Chapter, we investigate a zero-sum stochastic game for continuous-time Markov chain with denumerable state space and unbounded transition rates, under the probability criterion. Under suitable assumptions, we show the existence of value of the game and also characterize it as the unique solution of a pair of Shapley equations. We also establish the existence of a randomized stationary saddle point equilibrium. This Chapter is organized as follows: In Section 1, we describe the problem and the assumptions. In Section 2, we state and prove our main theorem. This Chapter is based on Bhabak and Saha [2021].

3.1 The model and probability criterion

The continuous-time zero-sum game model that we are interested in is a seven tuple

$$\{E, D, A, B, (A(i), i \in E, B(i), i \in E), q(j|i, a, b), r(i, a, b)\} \tag{3.1.1}$$

consisting of the following elements:

- a denumerable state space E;
- $D \subset E$ is a given target set;

- A, B are the action spaces, which are Borel spaces endowed with Borel σ-algebras $\mathcal{A}$ and $\mathcal{B}$ respectively; $A(i) \in \mathcal{A}$ and $B(i) \in \mathcal{B}$ are the sets of admissible actions in state $i \in E$ for player 1 and 2 respectively. Let $K := \{(i, a, b) | i \in E, a \in A(i), b \in B(i)\}$ be the set of admissible state-action pairs;

- the transition rates $q(j|i, a, b)$, which satisfy $q(j|i, a, b) \geq 0$ for all $(i, a, b) \in K$ and $i \neq j$. Moreover, the transition rates $q(j|i, a, b)$ are assume to be conservative, that is

$$\sum_{j \in E} q(j|i, a, b) = 0, \ \ \forall (i, a, b) \in K, \tag{3.1.2}$$

  and stable, that is

$$q^*(i) = \sup_{a \in A(i), b \in B(i)} q_i(a, b) < \infty, \ \ \forall (i, a, b) \in K, \tag{3.1.3}$$

  where $q_i(a, b) = -q(i|i, a, b) \geq 0$ for all $(i, a, b) \in K$;

- the reward rate $r(i, a, b)$, which is a non-negative real-valued measurable function on K and it is assumed that $r(i, ., .) > 0$ for all $i \in D^c$.

Now we describe the evolution of the controlled continuous-time Markov decision process (CTMDP). Suppose the system state is i_0 at the initial decision epoch $S_0 = 0$, and the decision makers have a common reward level (representing the profit goal for player 1 and cost level for player 2) λ_0 in mind. Depending on i_0, λ_0, player 1 selects an action $a_0 \in A(i_0)$ and player 2 selects an action $b_0 \in B(i_0)$. As a consequence of these choices, the system remains at i_0 until time t_1, at which point the system jumps to a new state i_1 with transition law $e^{-q_{i_0}(a_0,b_0)t_1} q(i_1|i_0, a_0, b_0)dt_1$. At time t_1, a reward $r(i_0, a_0, b_0)t_1$ is earned by player 1, which also denotes the cost for player 2. So, at state i_1 the remaining level is $\lambda_1 = \lambda_0 - r(i_0, a_0, b_0)t_1$ for both the players. On the basis of the current state i_1, the current reward level λ_1 as well as the previous state i_0, actions a_0 and b_0 and the previous reward level λ_0, player 1 chooses an action $a_1 \in A(i_1)$ and player 2 chooses an action $b_1 \in B(i_1)$. And the same sequence of events continues. Based on the above evolution, we obtain an admissible history h_n of the CTMDP up to the nth decision epoch, i.e.,

$$h_n = (0, i_0, \lambda_0, a_0, b_0, ..., t_{n-1}, i_{n-1}, \lambda_{n-1}, a_{n-1}, b_{n-1}, t_n, i_n, \lambda_n),$$

where, $0 < t_1 < \ldots < t_n$, $(i_m, a_m, b_m) \in K$, $\lambda_0 \in \mathbb{R}_+ = [0, \infty)$, $\lambda_{m+1} := \lambda_m - r(i_m, a_m, b_m)(t_{m+1} - t_m)$, and $i_n \in E$. Let H_n denote the set of all admissible histories h_n of the system up to the nth decision epoch, where H_n is endowed with a Borel σ-algebra.

Next, we define policies, which specifies a decision rule for the decision makers to select actions.

Definition 3.1.1. *A randomized history dependent policy for player 1 is a sequence* $\pi^1 = \{\pi_n^1 : n \geq 0\}$ *of stochastic kernels* π_n^1 *on* A *given* H_n *such that*

$$\pi_n^1(A(i_n)|h_n) = 1 \quad \forall\ h_n \in H_n, n = 0, 1, ...$$

A randomized history dependent policy for player 2 can be defined analogously.

We denote the set of all history dependent policies for player i by Π_i for $i = 1, 2$.

Notation: Let Φ_1 denote the set of all stochastic kernels ψ on A given $E \times \mathbb{R}$ satisfying $\psi(A(i)|i, \lambda) = 1, \quad \forall\ (i, \lambda) \in E \times \mathbb{R}$.

Definition 3.1.2. *(a) A policy* $\pi = \{\pi_n\}$ *is said to be randomized Markov for player 1 if there is a sequence* $\{\psi_n\}$ *of stochastic kernels* $\psi_n \in \Phi_1$*, such that* $\pi_n(.|h_n) = \psi_n(.|i_n, \lambda_n)$ *for every* $h_n \in H_n$ *and* $n \geq 0$*. In this case we write it as* $\pi = \{\psi_n\}$*.*
(b) A randomized Markov policy $\pi = \{\psi_n\}$ *is said to be randomized stationary for player 1 if* ψ_n *is independent of* n*. By an abuse of notation we will sometimes denote a randomized stationary policy by* ψ*.*

Similar policies can be defined for player 2. We denote by Π_i, Π_i^{RM}, Π_i^{RS} the families of all randomized history dependent, randomized Markov and stationary policies, respectively for player i, where $i = 1, 2$. Obviously, $\Phi_i = \Pi_i^{RS} \subset \Pi_i^{RM} \subset \Pi_i$.

For each $(i, \lambda) \in E \times \mathbb{R}$ and $\pi^1 \in \Pi_1$ and $\pi^2 \in \Pi_2$, by standard construction analogous to Guo and Piunovskiy [2011], there exist a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2})$ and stochastic process $\{S_n, J_n, \lambda_n, A_n, B_n\}$ such that, for each $t \in \mathbb{R}_+$, $j \in E$, $C \in \mathcal{A}$, $G \in \mathcal{B}$ and $n \geq 0$,

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}(S_0 = 0, J_0 = i, \lambda_0 = \lambda) = 1, \tag{3.1.4}$$

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}(A_n \in C, B_n \in G|h_n) = \pi_n^1(C|h_n)\pi_n^2(G|h_n), \tag{3.1.5}$$

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}(S_{n+1} - S_n \leq t, J_{n+1} = j|h_n, a_n, b_n) = \int_0^t e^{-q_{i_n}(a_n,b_n)s} q(j|i_n, a_n, b_n)ds\,, \tag{3.1.6}$$

where S_n, J_n, $\lambda_n := \lambda_{n-1} - r(J_{n-1}, A_{n-1}, B_{n-1})(S_n - S_{n-1})$, A_n, B_n denote the nth decision epoch, the state, the reward level and the actions chosen by player 1 and 2 at the nth decision epoch. The expectation operator with respect to $\mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}$ is denoted by $\mathbb{E}^{\pi^1,\pi^2}_{(i,\lambda)}$. Let $S_\infty = \lim_{n\to\infty} S_n$. In applications, it is natural to avoid the possibility of an infinite number of jumps within a finite time. In order to avoid explosion of the CTMDP, we impose the following assumption.

Assumption 3.1.3. *There exists a function $V \geq 1$ on E and constants $c_0 > 0$, $b_0 \geq 0$, and $L_0 \geq 0$ such that*
(a) $\sum_{j\in E} V(j)q(j|i,a,b) \leq c_0 V(i) + b_0$, for all $(i,a,b) \in K$; and
(b) $q^*(i) \leq L_0 V(i)$ for all $i \in E$, with $q^*(i)$ defined as in (3.1.3).

Then arguing analogous to Theorem 3.1 in Guo and Piunovskiy [2011], the following theorem can be proved.

Theorem 3.1.4. *Under Assumption 3.1.3, for any $(\pi^1,\pi^2) \in \Pi_1 \times \Pi_2$, $i \in E$, $\lambda \in \mathbb{R}$ we have $\mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}(S_\infty = \infty) = 1$.*

We also define the continuous time state action processes $\{X(t), U_1(t), U_2(t), t \in \mathbb{R}_+\}$ by

$$X(t) = J_n, \quad U_1(t) = A_n, \quad U_2(t) = B_n \quad \text{for} \quad S_n \leq t < S_{n+1}, n \geq 0\,.$$

For the given target set D, we define the the random variable,

$$\mathcal{T}_D := \inf\{t \geq 0 | X(t) \in D\} \quad (\text{with } \inf\emptyset := +\infty)\,,$$

which is the first hitting time into the set D of the state process $\{X(t)\}$. Now we define the probability criterion $U^{\pi^1,\pi^2}(i,\lambda)$ under a pair of policies $(\pi^1,\pi^2) \in \Pi_1 \times \Pi_2$ by

$$U^{\pi^1,\pi^2}(i,\lambda) := \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\int_0^{\mathcal{T}_D} r(X(t), U_1(t), U_2(t))dt > \lambda\right),$$

which gives capacity for player 1 to reach the profit level λ and also measures the risk of player 2 to control the cost level λ. To introduce our optimality problem, we also need the following functions:

$$I(i,\lambda) = \sup_{\pi^1\in\Pi_1} \inf_{\pi^2\in\Pi_2} U^{\pi^1,\pi^2}(i,\lambda)$$

$$L(i,\lambda) = \inf_{\pi^2 \in \Pi_2} \sup_{\pi^1 \in \Pi_1} U^{\pi^1,\pi^2}(i,\lambda)\,.$$

The function $I(\cdot,\cdot)$ is called the lower value of the game, while $L(\cdot,\cdot)$ is called the upper value of the game. Clearly, $I(i,\lambda) \leq L(i,\lambda)$ for every $(i,\lambda) \in E \times \mathbb{R}$.

Definition 3.1.5. *If $I(i,\lambda) = L(i,\lambda)$ for every $(i,\lambda) \in E \times \mathbb{R}$, then we call the common function the value of the game, which is denoted by V.*

Here player 1 is interested in maximizing $U^{\pi^1,\pi^2}(.,.)$ over $\pi^1 \in \Pi_1$ for each $\pi^2 \in \Pi_2$, and player 2 wants to minimize $U^{\pi^1,\pi^2}(.,.)$ over $\pi^2 \in \Pi_2$ for each $\pi^1 \in \Pi_1$. That is, we aim at finding a pair of optimal policies $(\pi^{*^1}, \pi^{*^2}) \in \Pi_1 \times \Pi_2$ as below.

Definition 3.1.6. *Suppose that the value of the game V exists. A policy $\pi^{*^1} \in \Pi_1$ is said to be optimal for player 1 if,*

$$\inf_{\pi^2 \in \Pi_2} U^{\pi^{*^1},\pi^2}(i,\lambda) = V(i,\lambda), \quad \forall (i,\lambda) \in E \times \mathbb{R}.$$

similarly, $\pi^{*^2} \in \Pi_2$ is said to be optimal for player 2 if,

$$\sup_{\pi^1 \in \Pi_1} U^{\pi^1,\pi^{*^2}}(i,\lambda) = V(i,\lambda), \quad \forall (i,\lambda) \in E \times \mathbb{R}.$$

If $\pi^{*^k} \in \Pi_k$ is optimal for player k, then $(\pi^{*^1}, \pi^{*^2}) \in \Pi_1 \times \Pi_2$ is called a pair of optimal policies, also known as saddle point equilibrium.

Remark 3.1.7.
i) By mimicking arguments as in Huang et al. [2017] it can be shown that, for any fixed $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, we have that

$$\sup_{\pi' \in \Pi_1} U^{\pi',\pi^2}(i,\lambda) = \sup_{\pi' \in \Pi_1^{RM}} U^{\pi',\pi^2}(i,\lambda),$$

$$\inf_{\pi' \in \Pi_2} U^{\pi^1,\pi'}(i,\lambda) = \inf_{\pi' \in \Pi_2^{RM}} U^{\pi^1,\pi'}(i,\lambda),$$

which implies that it is sufficient to limit ourselves to $\Pi_1^{RM} \times \Pi_2^{RM}$ in the upcoming arguments.

ii) For all $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$*, it is obvious that* $U^{\pi^1,\pi^2}(i,\lambda) = 1_{(-\infty,0)}(\lambda)$ *for every* $(i,\lambda) \in D \times \mathbb{R}$*, where* 1_C *is the indicator function on any set* C*. Thus, in order to avoid triviality, we restrict our attention to the case of* $(i,\lambda) \in D^c \times \mathbb{R}$*.*

3.2 Main results

In this section, we state and prove our main results. Let $\mathcal{P}(\mathcal{S})$ be the family of probability measures on the set $\mathcal{S}$, endowed with weak topology. We first introduce the following notation: Let $\mathcal{F}_m := \{U : D^c \times \mathbb{R} \to [0,1]$, such that $U(\cdot,\cdot)$ is Borel measurable on $D^c \times \mathbb{R}$, and $U(i,\lambda) = 1$ for $(i,\lambda) \in D^c \times (-\infty, 0)\}$. We also define operators $M^{\psi,\phi}$, M on $\mathcal{F}_m$ as follows: for $U \in \mathcal{F}_m$, $i \in D^c$, $a \in A(i)$, $b \in B(i)$, $\psi \in \mathcal{P}(A(i))$, $\phi \in \mathcal{P}(B(i))$, if $\lambda \geq 0$,

$$M^{\psi,\phi}U(i,\lambda) := \int_{A(i)} \psi(da) \int_{B(i)} \phi(db) \Bigg[e^{-q_i(a,b)\times \frac{\lambda}{r(i,a,b)}} + \sum_{j \neq i, j \in D^c} \int_0^{\lambda/r(i,a,b)} U(j, \lambda - r(i,a,b)t) e^{-q_i(a,b)t} q(j|i,a,b) dt \Bigg], \tag{3.2.1}$$

$$MU(i,\lambda) := \sup_{\psi \in \mathcal{P}(A(i))} \inf_{\phi \in \mathcal{P}(B(i))} M^{\psi,\phi}U(i,\lambda). \tag{3.2.2}$$

Moreover, if $\lambda < 0$, we define

$$M^{\psi,\phi}U(i,\lambda) = MU(i,\lambda) = 1. \tag{3.2.3}$$

For $(\pi^1, \pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$, define

$$M^{\pi^1,\pi^2}U(i,\lambda) := M^{\pi^1(\cdot|i,\lambda),\pi^2(\cdot|i,\lambda)}U(i,\lambda).$$

Note that, for each $(i,\lambda) \in D^c \times \mathbb{R}$ and $(\pi^1, \pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$, we can rewrite $U^{\pi^1,\pi^2}(i,\lambda)$ as follows:

$$
\begin{aligned}
U^{\pi^1,\pi^2}(i,\lambda) =& \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\int_0^{\mathcal{T}_D} r(X(t),U_1(t),U_2(t))dt > \lambda\right)\\
=& \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\sum_{m=0}^{\infty}\int_{S_m}^{S_{m+1}} 1_{\{\mathcal{T}_D>t\}} r(X(t),U_1(t),U_2(t))dt > \lambda\right)\\
=& \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\sum_{m=0}^{\infty} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda\right)\\
=& \lim_{n\to\infty} \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\sum_{m=0}^{n} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda\right), \qquad (3.2.4)
\end{aligned}
$$

where $\theta_{m+1} := S_{m+1} - S_m$ denote the sojourn times between two successive decision epochs. The last equality follows by the non-negativity of the reward rate $r(i,a,b)$ and the continuity of probability measures. Thus, we define a sequence $\{U_n^{\pi^1,\pi^2}(i,\lambda), n = -1,0,1,...\}$ by

$$
U_{-1}^{\pi^1,\pi^2}(i,\lambda) = 1_{(-\infty,0)}(\lambda),\ U_n^{\pi^1,\pi^2}(i,\lambda) := \mathbb{P}^{\pi^1,\pi^2}_{(i,\lambda)}\left(\sum_{m=0}^{n} 1_{\cap_{k=0}^{m}(J_k\in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda\right),
$$

for $n \geq 0$ and $(i,\lambda) \in D^c \times \mathbb{R}$.

Obviously, we have $U_n^{\pi^1,\pi^2}(i,\lambda) \leq U_{n+1}^{\pi^1,\pi^2}(i,\lambda)$, $\forall n \geq -1$, and $\lim_{n\to\infty} U_n^{\pi^1,\pi^2}(i,\lambda) = U^{\pi^1,\pi^2}(i,\lambda)$. We further impose the following conditions.

Assumption 3.2.1. *For every $(i,\lambda) \in D^c \times \mathbb{R}_+$,*
(a) A(i) and B(i) are compact;
(b) For all $i,j \in E$, the function $r(i,a,b)$ and $q(j|i,a,b)$ are continuous in $(a,b) \in A(i)\times B(i)$.
(c) For each fixed $U \in \mathcal{F}_m$, $\sum_{j\in D^c, j\neq i}\int_0^{\lambda/r(i,a,b)} U(j,\lambda - r(i,a,b)t)e^{-q_i(a,b)t}q(j|i,a,b)dt$ is continuous in $(a,b) \in A(i)\times B(i)$.

Lemma 3.2.2. *Suppose that Assumptions 3.1.3 and 3.2.1 hold. For any fixed $(i,\lambda) \in D^c\times\mathbb{R}$ and $\pi^1 = \{\psi_0,\psi_1,...\} \in \Pi_1^{RM}$, $\pi^2 = \{\phi_0,\phi_1,...\} \in \Pi_2^{RM}$ define the shifted policies ${}^{(1)}\pi^1 = \{\psi_1,\psi_2,...\} \in \Pi_1^{RM}$ and ${}^{(1)}\pi^2 = \{\phi_1,\phi_2,...\} \in \Pi_2^{RM}$. Then, for all $n \geq -1$, we have*
(a) $U_n^{\pi^1,\pi^2} \in \mathcal{F}_m$ and $U^{\pi^1,\pi^2} \in \mathcal{F}_m$.
(b) $U_{n+1}^{\pi^1,\pi^2} = M^{\psi_0,\phi_0}U_n^{{}^{(1)}\pi^1,{}^{(1)}\pi^2}$ and $U^{\pi^1,\pi^2} = M^{\psi_0,\phi_0}U^{{}^{(1)}\pi^1,{}^{(1)}\pi^2}$.
In particular $U^{\psi,\phi} = M^{\psi,\phi}U^{\psi,\phi}$ for every $(\psi,\phi) \in \Pi_1^{RS}\times\Pi_2^{RS}$.

Proof. (a) For any $(i,\lambda) \in D^c \times \mathbb{R}$, $(\pi^1,\pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$, we will show $U_n^{\pi^1,\pi^2} \in \mathcal{F}_m$ by induction. It is obviously true when $n = -1$. Now assume that $U_n^{\pi^1,\pi^2} \in \mathcal{F}_m$ for some $n \geq -1$. For $\lambda < 0$, it is easy to see that $U_{n+1}^{\pi^1,\pi^2} = 1$. For $\lambda \geq 0$,

$$
\begin{aligned}
U_{n+1}^{\pi^1,\pi^2}(i,\lambda) =& \mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\left(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k \in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda - r(J_0,A_0,B_0)\theta_1\right) \\
=& \mathbb{E}_{(i,\lambda)}^{\pi^1,\pi^2}\Bigg[\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Bigg(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k \in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda - r(J_0,A_0,B_0)\theta_1 \Bigg| \\
& \qquad S_0, J_0, \lambda_0, A_0, B_0, \theta_1, J_1, \lambda_1 = \lambda_0 - r(J_0,A_0,B_0)\theta_1\Bigg)\Bigg] \\
=& \int_{A(i)} \psi_0(da|i,\lambda) \int_{B(i)} \phi_0(db|i,\lambda) \sum_{j\neq i, j\in E} \int_0^\infty q(j|i,a,b) e^{-q_i(a,b)t} dt \\
& \times \Bigg[\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Bigg(\sum_{m=1}^{n+1} 1_{\cap_{k=0}^m (J_k \in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda - r(i,a,b)t \Bigg| \\
& S_0 = 0, J_0 = i, \lambda_0 = \lambda, A_0 = a, B_0 = b, \theta_1 = t, J_1 = j, \lambda_1 = \lambda - r(i,a,b)t\Bigg)\Bigg] \\
=& \int_{A(i)} \psi_0(da|i,\lambda) \int_{B(i)} \phi_0(db|i,\lambda) \Bigg[\sum_{j\neq i, j\in E} \int_{\lambda/r(i,a,b)}^\infty q(j|i,a,b) e^{-q_i(a,b)t} dt \times 1 \\
& + \sum_{j\neq i, j\in E} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \\
& \times \mathbb{P}_{(j,\lambda - r(i,a,b)t)}^{(1)\pi^1,(1)\pi^2}\Bigg(\sum_{m=0}^{n} 1_{\cap_{k=0}^m (J_k \in D^c)} r(J_m,A_m,B_m)\theta_{m+1} > \lambda - r(i,a,b)t\Bigg)\Bigg] \\
=& \int_{A(i)} \psi_0(da|i,\lambda) \int_{B(i)} \phi_0(db|i,\lambda) \Bigg[e^{-q_i(a,b)\frac{\lambda}{r(i,a,b)}} + \\
& \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \times U_n^{(1)\pi^1,(1)\pi^2}(j, \lambda - r(i,a,b)t)\Bigg)\Bigg],
\end{aligned}
$$

where the third inequality follows from properties (3.1.4)-(3.1.6), and the fourth equality is due to the Markov property of the policy pair (π^1,π^2) and the properties (3.1.4)-(3.1.6) again. Hence,

$$U_{n+1}^{\pi^1,\pi^2} = M^{\psi_0,\phi_0} U_n^{(1)\pi^1,(1)\pi^2} \quad \forall (i,\lambda) \in D^c \times \mathbb{R},$$

and thus $U_{n+1}^{\pi^1,\pi^2}(.,.) \in \mathcal{F}_m$. Therefore, by induction, $U_n^{\pi^1,\pi^2}(.,.) \in \mathcal{F}_m$ for every $n \geq -1$. Fur-

thermore, since limit of measurable functions is again measurable, we have $\lim_{n\to\infty} U_n^{\pi^1,\pi^2} = U^{\pi^1,\pi^2} \in \mathcal{F}_m$.

(b) From the proof of (a), we have $U_{n+1}^{\pi^1,\pi^2} = M^{\psi_0,\phi_0} U_n^{(1)\pi^1,(1)\pi^2}$. Letting $n \to \infty$, by the dominated convergence theorem we obtain $U^{\pi^1,\pi^2} = M^{\psi_0,\phi_0} U^{(1)\pi^1,(1)\pi^2}$.

The last statement is obvious. □

Remark 3.2.3. *(a) Following the proof of Lemma 3.2.2, we see that*

$$U^{(k)\pi^1,(k)\pi^2} = M^{\psi_k,\phi_k} U^{(k+1)\pi^1,(k+1)\pi^2} \tag{3.2.5}$$

holds for any ${}^{(k)}\pi^1 := \{\psi_{k+m}, m \geq 0\} \in \Pi_1^{RM}$ *and* ${}^{(k)}\pi^2 := \{\phi_{k+m}, m \geq 0\} \in \Pi_2^{RM}$ *with* $k = 0, 1, ...,$ ${}^{(0)}\pi^1 := \pi^1$ *and* ${}^{(0)}\pi^2 := \pi^2$.
(b) Lemma 3.2.2 gives a way of computing $U^{\pi^1,\pi^2}(.,.)$ *for each pair of policies* $(\pi^1, \pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$, *i.e.,* $U^{\pi^1,\pi^2}(.,.) = \lim_{n\to\infty} U_n^{\pi^1,\pi^2}(.,.)$ *with* $U_n^{\pi^1,\pi^2}(.,.)$ *recursively defined by* $U_{-1}^{\pi^1,\pi^2}(.,.) = 1_{(-\infty,0)}(\lambda)$, *and* $U_n^{\pi^1,\pi^2}(.,.) = M^{\pi^1,\pi^2} U_{n-1}^{\pi^1,\pi^2}(.,.)$ *for each* $n \geq 0$.

We need the following assumption to ensure the existence of optimal policies.

Assumption 3.2.4. $\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}(\mathcal{T}_D < \infty) = 1$ *for every* $(i, \lambda) \in D^c \times \mathbb{R}_+$ *and* $(\pi^1, \pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$.

Assumption 3.2.4 can be written equivalently as:

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\left(\cup_{n=1}^{\infty}\{J_n \in D\}\right) = 1 \text{ or } \mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\left(\cap_{n=1}^{\infty}\{J_n \in D^c\}\right) = 0 \text{ for every } (i,\lambda) \in D^c \times \mathbb{R}_+.$$

A sufficient condition for Assumption 3.2.4 is given by the following proposition.

Proposition 3.2.5. *If there exists some positive constant* $\beta < 1$, *such that* $\sum_{j\in D, j\neq i} q(j|i,a,b) \geq \beta q_i(a,b)$ *for all* $i \in D^c$ *and* $(a,b) \in A(i) \times B(i)$, *then Assumption 3.2.4 holds.*

Proof. We will prove by induction that

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\left(\cap_{k=1}^{n}\{J_k \in D^c\}\right) \leq (1-\beta)^n, \quad n \geq 1, \tag{3.2.6}$$

for all $(i,\lambda) \in D^c \times \mathbb{R}$, $(\pi^1,\pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$. When $n = 1$, for all $(i,\lambda) \in D^c \times \mathbb{R}$, $(\pi^1,\pi^2) \in \Pi_1^{RM} \times \Pi_2^{RM}$, we have

$$\begin{aligned}\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Big(J_1 \in D^c\Big) &= \mathbb{E}_{(i,\lambda)}^{\pi^1,\pi^2}\Big[1_{\{J_1\in D^c\}}\Big] \\ &= \int_{a\in A(i)} \psi_0(da|i,\lambda) \int_{b\in B(i)} \phi_0(db|i,\lambda) \sum_{j\neq i, j\in D^c} \frac{q(j|i,a,b)}{q_i(a,b)} \\ &\leq 1-\beta.\end{aligned}$$

Thus, the assertion is true for $n = 1$. Now assume that the assertion (3.2.6) holds for $n = k$. When $n = k+1$,

$$\begin{aligned}&\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Big(\cap_{l=1}^{k+1}\{J_l \in D^c\}\Big) = \mathbb{E}_{(i,\lambda)}^{\pi^1,\pi^2}\Big[1_{(\cap_{l=1}^{k+1}\{J_l\in D^c\})}\Big] \\ &= \mathbb{E}_{(i,\lambda)}^{\pi^1,\pi^2}\Big[\mathbb{E}_{(i,\lambda)}^{\pi^1,\pi^2}\Big[1_{(\cap_{l=1}^{k+1}\{J_l\in D^c\})}\Big|A_0,B_0,J_1,\theta_1,\lambda_1\Big]\Big] \\ &= \int_{a\in A(i)} \psi_0(da|i,\lambda) \int_{b\in B(i)} \phi_0(db|i,\lambda) \sum_{j\neq i, j\in E} \int_0^{+\infty} \mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Big(\cap_{l=1}^{k+1}\{J_l\in D^c\}\Big| \\ &A_0 = a, B_0 = b, J_1 = j, \theta_1 = t, \lambda_1 = \lambda - r(i,a,b)t\Big) \times e^{-q_i(a,b)t} q(j|i,a,b)dt \\ &= \int_{a\in A(i)} \psi(da|i,\lambda) \int_{b\in B(i)} \phi(db|i,\lambda) \\ &\times \sum_{j\neq i, j\in D^c} \int_0^{+\infty} \mathbb{P}_{(j,\lambda-r(i,a,b)t)}^{(1)\pi^1,(1)\pi^2}\Big(\cap_{l=1}^{k}\{J_l\in D^c\}\Big) \\ &\times e^{-q_i(a,b)t} q(j|i,a,b)dt \\ &\leq (1-\beta)^{k+1},\end{aligned}$$

where the second equality is obtained by using the conditional expectation property, and the last inequality follows from the induction hypothesis. Hence, we proved (3.2.6) by induction. Thus, by letting $n \to \infty$ in (3.2.6), we get

$$\mathbb{P}_{(i,\lambda)}^{\pi^1,\pi^2}\Big(\cap_{k=1}^{\infty}\{J_k \in D^c\}\Big) \leq \lim_{n\to\infty}(1-\beta)^n = 0\,.$$

□

Before stating the next lemma, we need to introduce one more notation as below. Let $\bar{\mathcal{F}}_m := \{H : D^c \times \mathbb{R} \to [-1, 1]$, satisfying $H(\cdot, \cdot)$ is Borel measurable and $H(i, \lambda) := 0$ for $\lambda < 0\}$. Define an operator on $\bar{\mathcal{F}}_m$ by

$$\bar{M}^{\psi,\phi}H(i,\lambda) := \int_{a\in A(i)} \psi(da|i,\lambda) \int_{b\in B(i)} \phi(db|i,\lambda) \sum_{j\neq i, j\in D^c} \int_0^\infty q(j|i,a,b)e^{-q_i(a,b)t} \times H(j, \lambda - r(i,a,b)t)dt \tag{3.2.7}$$

for $\lambda \geq 0$ and $\bar{M}^{\psi,\phi}H(i,\lambda) := 0$ otherwise for each $(\psi, \phi) \in \Phi_1 \times \Phi_2$ and $i \in D^c$.

Lemma 3.2.6. *Suppose that Assumptions 3.1.3, 3.2.1 and 3.2.4 are satisfied. Then for any function u in $\mathcal{F}_m$ and $(i, \lambda) \in D^c \times \mathbb{R}$, the following statements hold.*
(a) If $u(i,\lambda) \leq M^{\pi^1,\phi_k}u(i,\lambda)$ for all $k \geq 0$, and any policies $\pi^1 \in \Pi_1^{RS}$ and $\bar{\pi}^2 = \{\phi_k, k \geq 0\} \in \Pi_2^{RM}$, then $u(i,\lambda) \leq U^{\pi^1,\bar{\pi}^2}(i,\lambda)$.
(b) If $u(i,\lambda) \geq M^{\psi_k,\pi^2}u(i,\lambda)$ for all $k \geq 0$, and any policies $\pi^2 \in \Pi_2^{RS}$ and $\bar{\pi}^1 = \{\psi_k, k \geq 0\} \in \Pi_1^{RM}$, then $u(i,\lambda) \geq U^{\bar{\pi}^1,\pi^2}(i,\lambda)$.
(c) For every $(\pi^1, \pi^2) \in \Pi_1^{RS} \times \Pi_2^{RS}$, $U^{\pi^1,\pi^2}(.,.)$ is the unique solution in $\mathcal{F}_m$ to the equation $H = M^{\pi^1,\pi^2}H$.

Proof. (a) Obviously, the assertion holds for $\lambda < 0$. Now, we show the case for $\lambda \geq 0$. On the other hand, for any $i \in D^c$, policies $\pi^1 \in \Pi_1^{RS}$, $\bar{\pi}^2 = \{\phi_k, k \geq 0\} \in \Pi_2^{RM}$ and $n \geq 1$, we have

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(i,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}\right)\\
&= \mathbb{E}^{\pi^1,\bar{\pi}^2}_{(i,\lambda)}\left[\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(i,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}\middle| J_0, \lambda_0, A_0, B_0, \theta_1, J_1, \lambda_1 = \lambda_0 - r(J_0, A_0, B_0)\theta_1\right)\right]\\
&= \int_{A(i)} \pi^1(da|i,\lambda) \int_{B(i)} \phi_0(db|i,\lambda) \sum_{j\neq i, j\in E} \int_0^\infty q(j|i,a,b)e^{-q_i(a,b)t}dt\\
&\times \left[\mathbb{P}^{\pi^1,\bar{\pi}^2}_{(i,\lambda)}\left(\cap_{k=1}^{n+1}\{J_k \in D^c\}|J_0 = i, \lambda_0 = \lambda, A_0 = a, B_0 = b, \theta_1 = t, J_1 = j, \lambda_1 = \lambda - r(i,a,b)t\right)\right]\\
&= \int_{A(i)} \pi^1(da|i,\lambda) \int_{B(i)} \phi_0(db|i,\lambda) \sum_{j\neq i, j\in D^c} \int_0^\infty \mathbb{P}^{\pi^1,(1)\bar{\pi}^2}_{(j,\lambda-r(i,a,b)t)}(\cap_{k=0}^{n}\{J_k \in D^c\})\\
&q(j|i,a,b)e^{-q_i(a,b)t}dt.
\end{aligned} \tag{3.2.8}$$

On the other hand, it follows from (3.2.5) that $U^{\pi^1,(k)\bar{\pi}^2}(i,\lambda) = M^{\pi^1,\phi_k}U^{\pi^1,(k+1)\bar{\pi}^2}(i,\lambda)$ for all $k \geq 0$, which, together with the condition $u(i,\lambda) \leq M^{\pi^1,\phi_k}u(i,\lambda)$ and (3.2.8), gives for $\lambda \geq 0$,

$$
\begin{aligned}
&u(i,\lambda) - U^{\pi^1,\bar{\pi}^2}(i,\lambda) \leq \bar{M}^{\pi^1,\phi_0}[u(i,\lambda) - U^{\pi^1,(1)\bar{\pi}^2}(i,\lambda)] \\
&\leq \bar{M}^{\pi^1,\phi_0}\bar{M}^{\pi^1,\phi_1}...\bar{M}^{\pi^1,\phi_n}[u(i,\lambda) - U^{\pi^1,(n+1)\bar{\pi}^2}(i,\lambda)] \\
&= \int_{A(i)} \pi^1(da_0|i,\lambda) \int_{B(i)} \phi_0(db_0|i,\lambda) \sum_{i_1\neq i, i_1\in D^c} \int_0^\infty \int_{A(i_1)} \pi^1(da_1|i_1,\lambda_1) \int_{B(i_1)} \phi_1(db_1|i_1,\lambda_1) \\
&\sum_{i_2\neq i_1, i_2\in D^c} \int_0^\infty ... \int_{A(i_{n-1})} \pi^1(da_{n-1}|i_{n-1},\lambda_{n-1}) \int_{B(i_{n-1})} \phi_{n-1}(db_{n-1}|i_{n-1},\lambda_{n-1}) \sum_{i_n\neq i_{n-1}, i_n\in D^c} \int_0^\infty \\
&\mathbb{P}^{\pi^1,(n)\bar{\pi}^2}_{(i_n,\lambda_{n-1}-r(i_{n-1},a_{n-1},b_{n-1})t_n)}(J_1 \in D^c) q(i_n|i_{n-1},a_{n-1},b_{n-1}) e^{-q_{i_{n-1}}(a_{n-1},b_{n-1})t_n} ... \\
&q(i_1|i_0,a_0,b_0) e^{-q_{i_0}(a_0,b_0)t_1} dt_n \dots dt_1 \\
&= \int_{A(i)} \pi^1(da_0|i,\lambda) \int_{B(i)} \phi_0(db_0|i,\lambda) \sum_{i_1\neq i, i_1\in D^c} \int_0^\infty \int_{A(i_1)} \pi^1(da_1|i_1,\lambda_1) \int_{B(i_1)} \phi_1(db_1|i_1,\lambda_1) \\
&\sum_{i_2\neq i_1, i_2\in D^c} \int_0^\infty ... \int_{A(i_{n-2})} \pi^1(da_{n-2}|i_{n-2},\lambda_{n-2}) \int_{B(i_{n-2})} \phi_{n-2}(db_{n-2}|i_{n-2},\lambda_{n-2}) \sum_{i_{n-1}\neq i_{n-2}, i_{n-1}\in D^c} \int_0^\infty \\
&\mathbb{P}^{\pi^1,(n-1)\bar{\pi}^2}_{(i_{n-1},\lambda_{n-2}-r(i_{n-2},a_{n-2},b_{n-2})t_{n-1})}(\cap_{k=1}^2\{J_k \in D^c\}) q(i_{n-1}|i_{n-2},a_{n-2},b_{n-2}) e^{-q_{i_{n-2}}(a_{n-2},b_{n-2})t_{n-1}} ... \\
&q(i_2|i_1,a_1,b_1) e^{-q_{i_1}(a_1,b_1)t_2} q(i_1|i_0,a_0,b_0) e^{-q_{i_0}(a_0,b_0)t_1} dt_{n-1} \dots dt_1 \\
&= ... = \mathbb{P}^{\pi^1,\bar{\pi}^2}_{(i,\lambda)}\left(\cap_{k=1}^{n+1} \{J_k \in D^c\} \right),
\end{aligned}
$$

for all $n \geq 0$. Letting $n \to \infty$ in the above inequality, we obtain that $u(i,\lambda) \leq U^{\pi^1,\bar{\pi}^2}(i,\lambda)$.

(b) The fact that $U^{(k)\bar{\pi}^1,\pi^2}(i,\lambda) = M^{\psi_k,\pi^2}U^{(k+1)\bar{\pi}^1,\pi^2}(i,\lambda)$ in (3.2.5) and the given condition in this part yield that $U^{\bar{\pi}^1,\pi^2}(i,\lambda) - u(i,\lambda) \leq \bar{M}^{\psi_0,\pi^2}[U^{(1)\bar{\pi}^1,\pi^2}(i,\lambda) - u(i,\lambda)]$. Then, preceding similarly as in part (a) gives the desired result.

(c) Lemma 3.2.2, together with (a) and (b) of this lemma, gives the uniqueness of the solution $U^{\pi^1,\pi^2}(.,.)$ in $\mathcal{F}_m$ to the equation $H = M^{\pi^1,\pi^2}H$. □

Lemma 3.2.7. *For each $(i,\lambda) \in D^c \times \mathbb{R}_+$ and $u \in \mathcal{F}_m$, $M^{\psi,\phi}u(i,\lambda)$ is continuous in $(\psi,\phi) \in \mathcal{P}(A(i)) \times \mathcal{P}(B(i))$ with the operator $M^{\psi,\phi}u(i,\lambda)$ as in (3.2.1).*

Proof. Since $A(i)$ and $B(i)$ are compact, by Prohorov's theorem, so are $\mathcal{P}(A(i))$ and $\mathcal{P}(B(i))$.

Further,

$$e^{-q_i(a,b)\times\frac{\lambda}{r(i,a,b)}} + \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b)e^{-q_i(a,b)t}dt \times u(j,\lambda - r(i,a,b)t),$$

is a bounded continuous function on $A(i) \times B(i)$. By the definition of weak convergence of probability measures, for any sequence $\{(\psi_l, \phi_l)\} \subset \mathcal{P}(A(i)) \times \mathcal{P}(B(i))$ converging to (ψ, ϕ) with respect to the weak topology, letting $l \to \infty$ we obtain that

$$\begin{aligned}\lim_{l\to\infty} M^{\psi_l,\phi_l}u(i,\lambda) =& \int_{A(i)} \psi(da) \int_{B(i)} \phi(db) \Big[e^{-q_i(a,b)\times\frac{\lambda}{r(i,a,b)}} + \\ & \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b)e^{-q_i(a,b)t}dt \times u(j,\lambda - r(i,a,b)t)\Big] \\ =& M^{\psi,\phi}u(i,\lambda),\end{aligned}$$

which completes the proof. □

Remark 3.2.8. *Since $M^{\psi,\phi}$ as defined in (3.2.1) is bilinear in (ψ, ϕ), we have by our assumptions and Lemma 3.2.7, using Fan's minimax theorem Fan [1953] that*

$$MU(i,\lambda) = \sup_{\psi\in\mathcal{P}(A(i))} \inf_{\phi\in\mathcal{P}(B(i))} M^{\psi,\phi}U(i,\lambda) = \inf_{\phi\in\mathcal{P}(B(i))} \sup_{\psi\in\mathcal{P}(A(i))} M^{\psi,\phi}U(i,\lambda),$$

for $\lambda \geq 0$.

Let $u_{-1}(i,\lambda) := 1_{(-\infty,0)}(\lambda)$ and $u_n(i,\lambda) := Mu_{n-1}(i,\lambda)$ for each $(i,\lambda) \in E \times \mathbb{R}$ and $n \geq 0$, with operator M as in (3.2.2). Now we state the main result on the existence of a pair of optimal policies and value of the game.

Theorem 3.2.9. *Under Assumptions 3.1.3,3.2.1 and 3.2.4 using operators in (3.2.1)-(3.2.2), we have the following statements.*
(a) The limit $\lim_{n\to\infty} u_n(i,\lambda) = u^*(i,\lambda)$ exists and belongs to $\mathcal{F}_m$. Moreover, u^* satisfies the

pair of Shapley equations $u^*(i,\lambda) = Mu^*(i,\lambda)$, *i.e.,*

$$u^*(i,\lambda) = \sup_{\psi\in\mathcal{P}(A(i))}\inf_{\phi\in\mathcal{P}(B(i))}\left\{\int_{A(i)}\psi(da)\int_{B(i)}\phi(db)\left[e^{-q_i(a,b)\times\frac{\lambda}{r(i,a,b)}}+\sum_{j\neq i,j\in D^c}\int_0^{\lambda/r(i,a,b)}q(j|i,a,b)e^{-q_i(a,b)t}dt\times u^*(j,\lambda-r(i,a,b)t)\right]\right\} \tag{3.2.9}$$

$$= \inf_{\phi\in\mathcal{P}(B(i))}\sup_{\psi\in\mathcal{P}(A(i))}\left\{\int_{A(i)}\psi(da)\int_{B(i)}\phi(db)\left[e^{-q_i(a,b)\times\frac{\lambda}{r(i,a,b)}}+\sum_{j\neq i,j\in D^c}\int_0^{\lambda/r(i,a,b)}q(j|i,a,b)e^{-q_i(a,b)t}dt\times u^*(j,\lambda-r(i,a,b)t)\right]\right\} \tag{3.2.10}$$

for any $(i,\lambda)\in D^c\times\mathbb{R}_+$.

(b) There exists a pair of stationary policies $(\pi^{*^1},\pi^{*^2})\in\Pi_1^{RS}\times\Pi_2^{RS}$ *such that, for all* $(i,\lambda)\in D^c\times\mathbb{R}$,

$$u^*(i,\lambda) = M^{\pi^{*^1},\pi^{*^2}}u^*(i,\lambda) = \max_{\psi\in\mathcal{P}(A(i))}M^{\psi,\pi^{*^2}}u^*(i,\lambda) = \min_{\phi\in\mathcal{P}(B(i))}M^{\pi^{*^1},\phi}u^*(i,\lambda). \tag{3.2.11}$$

(c) $u^*(i,\lambda)$ *is the value of the game, and* $u^*(i,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(i,\lambda)$.

(d) (π^{*^1},π^{*^2}) *in (b) above is a saddle point equilibrium.*

Proof. (a) The compactness of $P(A(i))$ and $P(B(i))$, Lemma 3.2.7, and measurable selection theorem in Nowak [1985] gives that there exists $(\pi^1,\pi^2)\in\Pi_1^{RS}\times\Pi_2^{RS}$ (which may be dependent on n) such that $u_n(i,\lambda) = Mu_{n-1}(i,\lambda) = M^{\pi^1,\pi^2}u_{n-1}(i,\lambda)$ for all $n\geq 0$, which shows the measurability of $u_n(i,\lambda)$ in $\lambda\in\mathbb{R}$ for each $i\in D^c$. Moreover, $u_n(i,\lambda) = 1$ for $\lambda<0$. Therefore, $u_n\in\mathcal{F}_m$ for any $n\geq -1$. Also, it is easy to see that,

$$u_0(i,\lambda) = \sup_{\psi\in\mathcal{P}(A(i))}\inf_{\phi\in\mathcal{P}(B(i))}\int_{A(i)}\psi(da)\int_{B(i)}\phi(db)\left[e^{-q_i(a,b)\times\frac{\lambda}{r(i,a,b)}}\right] \geq u_{-1}(i,\lambda).$$

Therefore, by the definition of $\{u_n, n\geq -1\}$ and monotonicity of the operator M, we have $u_{-1}\leq u_0\leq ...\leq u_n...$, i.e., $\{u_n, n\geq -1\}$ is a non-decreasing sequence, and thus converges to some function $u^*\in\mathcal{F}_m$.

Now, for $\lambda<0$, $u^*(i,\lambda) = Mu^*(i,\lambda) = 1$. To establish the Shapley equations for $\lambda\geq 0$,

using the monotonicity again, we have $Mu^* \geq Mu_n = u_{n+1}$ for all $n \geq -1$, which shows that

$$Mu^* \geq u^*. \tag{3.2.12}$$

To show the reverse inequality, it follows from the definition of the operator M that

$$\begin{aligned} u_{n+1}(i,\lambda) \geq & \inf_{\phi \in \mathcal{P}(B(i))} \int_{A(i)} \psi(da) \int_{B(i)} \phi(db) \Bigg[e^{-q_i(a,b)\times \frac{\lambda}{r(i,a,b)}} + \\ & \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \times u_n(j, \lambda - r(i,a,b)t) \Bigg] \\ = & \int_{A(i)} \psi(da) \int_{B(i)} \phi_n^*(db) \Bigg[e^{-q_i(a,b)\times \frac{\lambda}{r(i,a,b)}} + \\ & \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \times u_n(j, \lambda - r(i,a,b)t) \Bigg] \end{aligned} \tag{3.2.13}$$

for any $\psi \in \mathcal{P}(A(i))$, where the existence of $\phi_n^* \in \mathcal{P}(B(i))$ (may be dependent on ψ) is guaranteed by Lemma 3.2.7. By the compactness of $\mathcal{P}(B(i))$, without loss of generality, we suppose that $\phi_n^* \to \phi^* \in \mathcal{P}(B(i))$. Taking $n \to \infty$ in equation (3.2.13), it follows from the extended Fatou's lemma (Lemma 8.3.7 in Hernández-Lerma and Lasserre [1999]) that

$$\begin{aligned} u^*(i,\lambda) \geq & \int_{A(i)} \psi(da) \int_{B(i)} \phi^*(db) \Bigg[e^{-q_i(a,b)\times \frac{\lambda}{r(i,a,b)}} + \\ & \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \times u^*(j, \lambda - r(i,a,b)t) \Bigg] \\ \geq & \inf_{\phi \in \mathcal{P}(B(i))} \Bigg\{ \int_{A(i)} \psi(da) \int_{B(i)} \phi(db) \Bigg[e^{-q_i(a,b)\times \frac{\lambda}{r(i,a,b)}} + \\ & \sum_{j\neq i, j\in D^c} \int_0^{\lambda/r(i,a,b)} q(j|i,a,b) e^{-q_i(a,b)t} dt \times u^*(j, \lambda - r(i,a,b)t) \Bigg] \Bigg\}. \end{aligned}$$

Now, since $\psi \in \mathcal{P}(A(i))$ is arbitrary, we get that

$$Mu^* \leq u^*,$$

which together with (3.2.12) gives $Mu^* = u^*$.

(b) Obviously, (3.2.11) holds for $\lambda < 0$. For $\lambda \geq 0$, if we choose $\pi^{*^1} \in \Pi_1^{RS}$ as the outer maximizing selector in (3.2.9) and $\pi^{*^2} \in \Pi_2^{RS}$ as the outer minimizing selector in (3.2.10) then (3.2.11) follows.

(c) From (b) of this theorem, we have $u^*(i,\lambda) = M^{\pi^{*^1},\pi^{*^2}} u^*(i,\lambda)$. Thus by lemma 3.2.6(c), we have $u^*(i,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(i,\lambda)$. Let π^{*^2} be fixed. For any policy $\pi^1 := \{\psi_n, n \geq 0\} \in \Pi_1^{RM}$, we get $\psi_n(.|i,\lambda) \in \mathcal{P}(A(i))$ for all $n \geq 0$ and $(i,\lambda) \in D^c \times \mathbb{R}$. Then, from (3.2.11), for any $n \geq 0$, we have

$$u^*(i,\lambda) \geq M^{\psi_n,\pi^{*^2}} u^*(i,\lambda),$$

which combined with lemma 3.2.6(b) gives $u^*(i,\lambda) \geq U^{\pi^1,\pi^{*^2}}(i,\lambda)$ for all $\pi^1 \in \Pi_1^{RM}$. Therefore, $u^*(i,\lambda) \geq \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^{*^2}}(i,\lambda)$, while the reverse inequality follows from $u^*(i,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(i,\lambda)$, which implies that, for any $(i,\lambda) \in D^c \times \mathbb{R}$,

$$u^*(i,\lambda) = \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^{*^2}}(i,\lambda) \geq \inf_{\pi^2 \in \Pi_2^{RM}} \sup_{\pi^1 \in \Pi_1^{RM}} U^{\pi^1,\pi^2}(i,\lambda) = L(i,\lambda). \tag{3.2.14}$$

Using (3.2.11) again, a similar argument together with Lemma 3.2.6(a) shows that

$$u^*(i,\lambda) = \inf_{\pi^2 \in \Pi_2^{RM}} U^{\pi^{*^1},\pi^2}(i,\lambda) \leq \sup_{\pi^1 \in \Pi_1^{RM}} \inf_{\pi^2 \in \Pi_2^{RM}} U^{\pi^1,\pi^2}(i,\lambda) = I(i,\lambda). \tag{3.2.15}$$

Hence, by (3.2.14) and (3.2.15), $L(i,\lambda) = I(i,\lambda) = V(i,\lambda) = u^*(i,\lambda) = U^{\pi^{*^1},\pi^{*^2}}(i,\lambda)$.

(d) This directly follows from both (3.2.14) and (3.2.15). □

Remark 3.2.10. *The proof techniques in chapters 2 and 3 are very similar. But there are some key differences in the model. In chapter 2, the state space is general while in chapter 3 it is denumerable. In chapter 3, the transition rates are allowed to be unbounded. Condition in equation (2.1.5) is satisfied by continuous-time Markov chain only when transition rates are bounded. In chapter 3, Assumption 3.1.3 is needed to ensure non-explosion of the process. Also, in chapter 3, we assume that action spaces are compact metric spaces and not necessarily finite. Thus, we need Assumption 3.2.1 in this chapter.*

CHAPTER 4

Risk-sensitive Semi-Markov Decision Problems with Discounted Cost and General Utilities

In this Chapter we consider risk-sensitive control of semi-Markov processes with a discrete state space. We consider general utility functions and discounted cost in the optimization criteria. We consider random finite horizon and infinite horizon problems. Using a state augmentation technique we characterize the value functions and also prescribe optimal controls. This Chapter is organized as follows: In Section 1, we describe our model and the control problem. In Section 2 we investigate the random finite horizon problem. Finally in Section 3, we analyze the infinite horizon problem. This Chapter is based on Bhabak and Saha [2022b].

4.1 The Control Model

The semi-Markov decision problem(SMDP) model that we are interested in is

$$\{E, A, (A(i), i \in E), Q(.,.|i,a), C(i,a), U(.)\},$$

where the individual components has the following interpretation:

- E is a countable state space. Without loss of generality, we take $E = \{1, 2, \ldots\}$.

- A is the action space, which is assumed to be Borel space endowed with the Borel σ-algebra $\mathcal{A}$.
- $A(i) \in \mathcal{A}$ denotes the set of all admissible actions in state i. Let $K := \{(i,a)|i \in E, a \in A(i)\}$ be the set of all admissible state action pairs.
- $Q(\cdot,\cdot|i,a)$ is a semi-Markov kernel on $[0,\infty)\times E$ given K. We assume that $Q(0,j|i,a) = 0$ for any $j \in E$ and $(i,a) \in K$. It describes the transition mechanism of the controlled process. Thus if $a \in A(i)$ is the action chosen in state i, then for any $t > 0$ and $j \in E$, $Q(t,j|i,a)$ is the joint probability that the sojourn time in state i will be less than or equal to t and the next transition will be into state j.
- $C : K \to [0,\bar{c}]$ is a measurable running cost function with $0 < \bar{c} < \infty$.
- $U : [0,\infty) \to \mathbb{R}$ denotes a utility function, which is assumed to be continuous and strictly increasing.

Now we describe the evolution of the controlled semi-Markov process. At time 0, which is the initial decision epoch, the system is in state i_0. Depending upon the state of the system the controller chooses an action $a_0 \in A(i_0)$. As a consequence of this choice of action the system remains at i_0 until time t_1. At time t_1 the system jumps to the next state i_1 according to the transition law $Q(dt_1, i_1|i_0, a_0)$. A discounted cost equal to $\int_0^{t_1} e^{-\alpha u} C(i_0,a_0)du$ is generated. Now in state i_1, depending on the current state, previous state, sojourn time, and previously selected action the controller chooses an action $a_1 \in A(i_1)$ and the same sequence of events repeat. Based on this evolution and we obtain a history $h_n = (i_0, a_0, t_1, i_1, a_1, t_2, ..., i_{n-1}, a_{n-1}, t_n, i_n)$ up to the nth jump of the described process. Here t_k denotes the *kth* jump time with the assumption $t_0 = 0$, i_k is the state after the *kth* jump and a_k is the action chosen at the *kth* jump time. Let H_n denote the set of all possible histories upto the *nth* jump time. H_n is endowed with a Borel $\sigma-$field.
Next, we describe the policies which govern the choice of action by the decision-maker.

Definition 4.1.1. *A history dependent policy $\pi := \{\pi_n, n \geq 0\}$ is sequence of measurable functions $\pi_n : H_n \to A$ such that $\pi_n(h_n) \in A(i_n)$. A history dependent policy π is said to be Markov if there exists a sequence $\{f_n\}$ of measurable functions $f_n : [0,\infty) \times E \to A$, such that $f_n(t,i) \in A(i)$ and $\pi_n(h_n) = f_n(t_n, i_n)$. In this case we write $\pi = \{f_n\}_{n\geq 0}$. If $f_n = f$ for some common function f for all n, then the Markov policy is said to be stationary. We will sometimes denote a stationary policy by the common function f. We denote by Π, Π^M, Π^S the set of all history dependent, Markov and stationary policies respectively.*

For each $i \in E$ and $\pi \in \Pi$ by the well-known Tulcea's Theorem (Hernández-Lerma and Lasserre [1996], Proposition C.10), there exist a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_i^\pi)$ and stochastic processes $\{S_n, X_n, A_n\}_{n\geq 0}$ such that, for each $t \in [0, \infty)$, $j \in E$, $C \in \mathcal{A}$ and $n \geq 0$,

$$\mathbb{P}_i^\pi(S_0 = 0, X_0 = i) = 1,$$

$$\mathbb{P}_i^\pi(A_n \in C|h_n) = \delta_{\pi_n(h_n)}(C),$$

$$\mathbb{P}_i^\pi(S_{n+1} - S_n \leq t, X_{n+1} = j|h_n, a_n) = Q(t, j|i_n, a_n),$$

where S_n, X_n and A_n denote the *nth* jump time, the state and the action chosen by the decision maker at the nth jump time and δ denotes the Dirac measure. The expectation operator with respect to $\mathbb{P}_i^\pi$ is denoted by $\mathbb{E}_i^\pi$. In order to avoid the possibility of an infinite number of jumps within a finite time interval we make the following standard assumption.

Assumption 4.1.2. *There exist constants $\delta > 0$ and $\epsilon > 0$ such that*

$$\sup_{(i,a)\in K} Q(\delta, E|i, a) \leq 1 - \epsilon. \tag{4.1.1}$$

Remark 4.1.3. *Assumption 4.1.2 means that the sojourn time at any state and under any action exceeds δ with a probability at least ϵ. If $S_\infty = \lim_{n\to\infty} S_n$, then it is well known that (see Proposition 2.1 of Huang et al. [2011]), if (4.1.1) holds then $\mathbb{P}_i^\pi(S_\infty = \infty) = 1$ for any $i \in E$ and $\pi \in \Pi$.*

We also define the continuous time processes $\{X(t), A(t), t \in [0, \infty)\}$ by

$$X(t) = X_n, \quad A(t) = A_n, \quad \text{for} \quad S_n \leq t < S_{n+1},\ t \in [0, \infty) \text{ and } n \geq 0\,.$$

Now we describe the cost criteria. Let $\alpha > 0$ be a discount factor. We consider SMDPs over both finite(random) and infinite horizons. For $N \geq 1$, the total discounted cost accumulated upto the Nth jump time is given by

$$C_N = \int_0^{S_N} e^{-\alpha u} C(X_u, A_u) du = \sum_{n=1}^{N} e^{-\alpha T_{n-1}} \int_0^{S_n - S_{n-1}} e^{-\alpha t} C(X_{n-1}, A_{n-1}) dt,$$

and the total discounted cost accumulated over the infinite time horizon is given by

$$C_\infty = \int_0^\infty e^{-\alpha u} C(X_u, A_u) du = \sum_{n=1}^{\infty} e^{-\alpha S_{n-1}} \int_0^{S_n - S_{n-1}} e^{-\alpha t} C(X_{n-1}, A_{n-1}) dt.$$

Although, not explicit in the notation, note that both C_N and C_∞ depends on the control policy π. Instead of the standard expectation minimization, here we consider the following minimization problems:

$$\inf_{\pi \in \Pi} \mathbb{E}_i^\pi \big[U(C_N)\big], \ \ i \in E \ \text{ and}$$

$$\inf_{\pi \in \Pi} \mathbb{E}_i^\pi \big[U(C_\infty)\big], \ \ i \in E.$$

For $i \in E$ and $\pi \in \Pi$, let

$$J_N^\pi(i) = \mathbb{E}_i^\pi \big[U(C_N)\big] \quad \text{and} \quad J_\infty^\pi(i) = \mathbb{E}_i^\pi \big[U(C_\infty)\big].$$

Also let

$$J_N(i) = \inf_{\pi \in \Pi} J_N^\pi(i) \quad \text{and} \quad J_\infty(i) = \inf_{\pi \in \Pi} J_\infty^\pi(i).$$

A policy $\pi^* \in \Pi$ is said to be optimal for the finite horizon problem if $J_N^{\pi^*}(i) = J_N(i)$ for all $i \in E$. Similarly, a policy π^* is said to be optimal for the infinite horizon problem if $J_\infty^{\pi^*}(i) = J_\infty(i)$ for all i in E. We wish to characterise $J_N(\cdot)$ and $J_\infty(\cdot)$ and find optimal policies for both finite and infinite horizon problems.

4.2 Finite Horizon Problem

We first consider the optimization problem upto the *Nth* jump time. We will use a state augmentation technique to convert the original problem to a standard risk-neutral problem. Similar state augmentation technique has been used in the context of discrete time MDP in Bäuerle and Rieder [2014] and in the context of SMDP in Huang et al. [2018]. We augment the state process to include the accumulated discounted cost. More precisely, we consider the augmented controlled state process $\{S_n, X_n, C_n, n \geq 0\}$ where S_n and X_n are

as before and C_n is the accumulated discounted cost upto the nth jump time. For $i \in E$, $(t, \lambda) \in [0, \infty) \times [0, \infty)$ and $a \in A(i)$, the controlled transition law of the augmented state process is given by

$$\hat{Q}(B \times \{j\} \times C|t, i, \lambda, a) = \int_0^\infty 1_B(t+s)\delta_{\lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})}(C)Q(ds, j|i, a),$$

where 1 is the indicator function, B and C are Borel subsets of $[0, \infty)$, $j \in E$. In this augmented set-up we redefine the various policy sets. But for economy of notation, we use the same notations for the augmented policy sets. Thus, in particular, in the augmented set-up a Markov policy is given by $\pi = \{f_n\}$ where $f_n : [0, \infty) \times E \times [0, \infty) \to A$ are measurable functions such that $f_n(t, i, \lambda) \in A(i)$ for all (t, i, λ).

Suppose $\mathbb{E}_{(t,i,\lambda)}$ is the expectation operator corresponding to the initial condition $S_0 = t, X_0 = i, C_0 = \lambda$. Then for $n = 0, 1, \ldots, N$ and $\pi \in \Pi$ define the value functions,

$$V_{n\pi}(t, i, \lambda) = \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u}C(X_u, A_u)du + \lambda\big)\big] = \mathbb{E}^\pi_{(t,i,\lambda)}[U(C_n)],$$

$$(t, i, \lambda) \in [0, \infty) \times E \times [0, \infty),$$

$$V_n(t, i, \lambda) = \inf_{\pi \in \Pi} V_{n\pi}(t, i, \lambda), \quad (t, i, \lambda) \in [0, \infty) \times E \times [0, \infty). \tag{4.2.1}$$

Thus $J_N(i) = V_N(0, i, 0)$. The state augmentation now allows us to think of the optimization problem as a finite horizon discrete time Markov decision process with state process $\{S_n, X_n, C_n, n \geq 0\}$, zero one stage cost and terminal cost function $g(t, i, \lambda) = U(\lambda)$. Now define the set

$$B([0, \infty) \times E \times [0, \infty)) = \{v : [0, \infty) \times E \times [0, \infty) \to [0, \infty) \text{ is measurable and}$$
$$U(\lambda) \leq v(t, i, \lambda) \leq U(e^{-\alpha t}\frac{\bar{c}}{\alpha} + \lambda)\ \forall (t, i, \lambda)\}. \tag{4.2.2}$$

Let F denote the set of all measurable functions $f : [0, \infty) \times E \times [0, \infty) \to A$ such that $f(t, i, \lambda) \in A(i)$ for all (t, i, λ). Then for $v \in B([0, \infty) \times E \times [0, \infty))$ and $f \in F$, we define

the operators

$$T_f v(t,i,\lambda) := \sum_j \int v\big(t+s, j, \lambda + \frac{C(i,f(t,i,\lambda))}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i, f(t,i,\lambda)) \tag{4.2.3}$$

and

$$(Tv)(t,i,\lambda) := \inf_{a\in A(i)} \sum_j \int v\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i, a). \tag{4.2.4}$$

We say that $f \in F$ is a minimizer of v if $Tv = T_f v$. Observe that the operators $T_f v$ and T are both monotone. We need to impose the following assumption.

Assumption 4.2.1. *For each $(t,i,\lambda) \in [0,\infty) \times E \times [0,\infty)$,*
(i) $A(i)$ is compact.
(ii) $C(i,a)$ is continuous on $A(i)$.
(iii) $\sum_j \int v\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i, a)$ is continuous on $A(i)$, for each $v \in B([0,\infty) \times E \times [0,\infty))$.

We have the following:

Theorem 4.2.2. *Suppose Assumptions 4.1.2 and 4.2.1 hold. For $n = 1, ..., N$, we have the following.*

(a) For any policy $\pi = (f_0, f_1, ...) \in \Pi^M$, we have the iteration: $V_{n\pi} = T_{f_0} T_{f_1} ... T_{f_{n-1}} U$, where T_{f_i} is as in (4.2.3).

(b) $V_0(t,i,\lambda) = U(\lambda)$ and $V_n = TV_{n-1}$ i.e.,

$$V_n(t,i,\lambda) = \inf_{a\in A(i)} \sum_j \int V_{n-1}\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i, a). \tag{4.2.5}$$

Also, $V_n \in B([0,\infty) \times E \times [0,\infty))$ for all $n = 0, 1, \ldots, N$.

(c) For every $n = 1, 2, ..., N$ there exists a minimizer $f_n^* \in F$ of V_{n-1} and $(f_N^*, f_{N-1}^*, \ldots, f_1^*)$ is an optimal Markov policy for the finite horizon optimization problem.

Proof. We prove (a) by induction on n.
Firstly, $V_{0\pi}(t, i, \lambda) = U(\lambda)$. Now for $n = 1$ we have,

$$\begin{aligned} V_{1\pi}(t,i,\lambda) =& \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\lambda + \int_t^{S_1} e^{-\alpha u} C(i, f_0(t,i,\lambda))du\big)\big] \\ =& (T_{f_0}U)(t,i,\lambda). \end{aligned}$$

Now suppose that the statement holds for $V_{n-1\pi}$. So we consider $V_{n\pi}$:

$$\begin{aligned} &(T_{f_0}...T_{f_{n-1}}U)(t,i,\lambda) = T_{f_0}(T_{f_1}...T_{f_{n-1}}U)(t,i,\lambda) \\ &= \sum_j \int V_{n-1\bar{\pi}}\big(t+s, j, \lambda + \frac{C(i, f_0(t,i,\lambda))}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i, f_0(t,i,\lambda)) \\ &= \sum_j \int \mathbb{E}^{\bar{\pi}}_{(t+s,j,\lambda')}\big[U\big(\lambda + \frac{C(i, f_0(t,i,\lambda))}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)}) \\ &+ \int_{t+s}^{S_{n-1}} e^{-\alpha u} C(X_u, a_u)du\big)\big] Q(ds, j|i, f_0(t,i,\lambda)) \\ &= \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u} C(X_u, A_u)du + \lambda\big)\big] \\ &= V_{n\pi}(i, \lambda, t), \end{aligned}$$

where $\bar{\pi}$ is the one shifted policy and $\lambda' = \lambda + \frac{C(i,f_0(t,i,\lambda))}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})$. Hence we get our desired result for $V_{n\pi}$. So by induction argument we have showed (a).

(b) and (c): The proof of (*b*) and (*c*) follows by Assumption 4.2.1 and standard theory of discrete time MDP, see Chapter 3 of Hernández-Lerma and Lasserre [1996]. □

Corollary 4.2.3. *In the case of $U(\lambda) = (\frac{1}{\gamma})e^{\gamma\lambda}$ with $\gamma \neq 0$, we have $V_n(i, \lambda, t) = e^{\gamma\lambda}h_n(t, i)$*

and $J_N(i) = h_N(0,i)$. *And* h_n *satisfies the iteration given by* $h_0(i,\lambda) = \frac{1}{\gamma}$ *and*

$$h_n(t,i) = \inf_{a\in A(i)} \sum_j \int e^{\gamma \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})} h_{n-1}(t+s,j) Q(ds, j|i,a).$$

Proof. We will prove by induction on n. For $n = 0$ we have $V_0(t,i,\lambda) = (\frac{1}{\gamma})e^{\gamma\lambda} = e^{\gamma\lambda}(\frac{1}{\gamma}) = e^{\gamma\lambda}h_0$. Hence the statement is true for $n = 0$. Now suppose it is true for $n-1$. From (4.2.5) we get,

$$\begin{aligned} V_n(t,i,\lambda) &= \inf_{a\in A(i)} \sum_j \int V_{n-1}\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) Q(ds, j|i,a) \\ &= \inf_{a\in A(i)} \sum_j \int e^{\gamma(\lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)}))} h_{n-1}(t+s,j) Q(ds, j|i,a) \\ &= e^{\gamma\lambda} \inf_{a\in A(i)} \sum_j \int e^{\gamma \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})} h_{n-1}(t+s,j) Q(ds, j|i,a) \end{aligned}$$

Hence, the statement follows by setting

$$h_n(t,i) = \inf_{a\in A(i)} \sum_j \int e^{\gamma \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})} h_{n-1}(t+s,j) Q(ds, j|i,a). \tag{4.2.6}$$

□

4.3 Infinite Horizon Problem

Now to consider the infinite horizon problem. Like in the finite horizon case we again consider the augmented set-up and define the following value functions.

$$V_{\infty\pi}(t,i,\lambda) := \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\int_t^{\infty} e^{-\alpha u} C(X_u, A_u) du + \lambda\big)\big],\ \pi \in \Pi,\ (t,i,\lambda) \in [0,\infty) \times E \times [0,\infty).$$

$$V_\infty(t,i,\lambda) = \inf_{\pi\in\Pi} V_{\infty\pi}(t,i,\lambda),\ (t,i,\lambda) \in [0,\infty)\times E\times[0,\infty).$$

For a stationary policy $\pi = \{f\}$, we will will write $V_{\infty\pi}$ as V_f. It is easy to see that, $J_\infty(i) = V_\infty(0,i,0)$ for all $i \in E$. For the infinite horizon problem, we will deal with convex and concave utility functions separately. First we analyze the concave case.

4.3.1 Concave Utility Function.

let $U : \mathbb{R}_+ \to \mathbb{R}$ be a concave utility function. We introduce one more notation $\bar{v}(t,\lambda) = U(e^{-\alpha t}\frac{\bar{c}}{\alpha} + \lambda)$. It is straight forward to see that $U(\lambda) \le V_\infty(i,\lambda,t) \le \bar{v}(t,\lambda)$. Thus $V_\infty \in B([0,\infty)\times E\times[0,\infty))$ where $B([0,\infty)\times E\times[0,\infty))$ is given by (4.2.2).

Theorem 4.3.1. *Suppose that Assumptions 4.1.2 and 4.2.1 hold. Also suppose that the derivative $U'(0)$ exists. Then the following statements hold.*

(a) V_∞ is the unique solution of $v = Tv$ in $B([0,\infty)\times E\times[0,\infty))$ for T defined in (4.2.4). Moreover, $T^nU \uparrow V_\infty$ and $T^n\bar{v} \downarrow V_\infty$ as $n\to\infty$.

(b) There exists a minimizer $f^* \in F$ of V_∞ and the stationary policy determined by f^* is an optimal policy for the infinite horizon problem.

Proof. (a) Here we first show that $V_n = T^nU \uparrow V_\infty$ as $n \to \infty$. It is known that for $U : \mathbb{R}_+ \to \mathbb{R}$ increasing and concave we have the inequality

$$U(\lambda_1+\lambda_2) \le U(\lambda_1) + U'_-(\lambda_1)\lambda_2, \quad \lambda_1,\lambda_2 \ge 0,$$

where U'_- is the left-hand side derivative of U that exists since U is concave. Moreover,

$U'_-(\lambda) \geq 0$ and is decreasing. For $(t, i, \lambda) \in [0, \infty) \times E \times [0, \infty)$ and $\pi \in \Pi$ we have,

$$\begin{aligned}
V_n(t,i,\lambda) \leq V_{n\pi}(t,i,\lambda) \leq V_{\infty\pi}(t,i,\lambda) &= \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\int_t^{\infty} e^{-\alpha u}C(X_u,A_u)du) + \lambda\big)\big] \\
&= \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du + \lambda + \int_{S_n}^{\infty} e^{-\alpha u}C(X_u,A_u)du\big)\big] \\
&\leq \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du) + \lambda\big)\big] \\
&+ \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U'_-\big(\int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du) + \lambda\big) \times \big(\int_{S_n}^{\infty} e^{-\alpha u}C(X_u,A_u)du\big)\big] \\
&\leq V_{n\pi}(t,i,\lambda) + U'_-(\lambda)\frac{\bar{c}}{\alpha}\mathbb{E}^{\pi}_{t,i,\lambda}(e^{-\alpha S_n}).
\end{aligned}$$

Now using (4.1.1), we have

$$\begin{aligned}
\mathbb{E}^{\pi}_{(t,i,\lambda)}[e^{-\alpha S_1}] &\leq \sup_{a\in A(i)} \int_0^{\infty} e^{-\alpha(t+s)}Q(ds,E|i,a) \\
&= \sup_{a\in A(i)} e^{-\alpha t}\left[\int_0^{\delta} e^{-\alpha s}Q(ds,E|i,a) + \int_{\delta}^{\infty} e^{-\alpha s}Q(ds,E|i,a)\right] \\
&\leq \sup_{a\in A(i)} e^{-\alpha t}\left[Q(\delta,E|i,a) + e^{-\alpha\delta}(1-Q(\delta,E|i,a))\right] \\
&\leq e^{-\alpha t}[(1-e^{-\alpha\delta})(1-\epsilon) + e^{-\alpha\delta}] \\
&= e^{-\alpha t}(1-\epsilon+\epsilon e^{-\alpha\delta}).
\end{aligned}$$

Thus, it can be shown by induction that

$$\mathbb{E}^{\pi}_{(t,i,\lambda)}[e^{-\alpha S_n}] \leq e^{-\alpha t}(1-\epsilon+\epsilon e^{-\alpha\delta})^n,$$

for all n. Thus we have,

$$V_n(t,i,\lambda) \leq V_{\infty\pi}(t,i,\lambda) \leq V_{n\pi}(i,\lambda,t) + \epsilon_n(t,\lambda),$$

where $\epsilon_n(\lambda,t) = U'_-(\lambda)\frac{\bar{c}}{\alpha}e^{-\alpha t}(1-\epsilon+\epsilon e^{-\alpha\delta})^n$. Thus $\lim_{n\to\infty}\epsilon_n(\lambda,t) = 0$. Taking infimum

over all policies we get,

$$V_n(t,i,\lambda) \leq V_\infty(i,\lambda,t) \leq V_n(t,i,\lambda) + \epsilon_n(t,\lambda).$$

Now as $n \to \infty$ we have $V_n = T^n U \uparrow V_\infty$ for $n \to \infty$. Now, we try to show that $V_\infty = TV_\infty$. Note that $V_n \leq V_\infty$ for all n. Using the fact that T is increasing we have $V_{n+1} = TV_n \leq TV_\infty$ for all n. Letting $n \to \infty$ implies $V_\infty \leq TV_\infty$.
For the reverse inequality we have from above $V_n + \epsilon_n \geq V_\infty$. Applying the T operator and also using its monotonicity we get,

$$\begin{aligned} T(V_n+\epsilon_n)(t,i,\lambda) = \inf_{a\in A(i)} \sum_j \int \Big(& V_n\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big) \\ & + \epsilon_n\big(t+s, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big)\Big) Q(ds, j|i, a) \\ & \leq V_{n+1} + \epsilon_{n+1}. \end{aligned}$$

Hence, we have $V_{n+1}+\epsilon_{n+1} \geq T(V_n+\epsilon_n) \geq TV_\infty$. Now letting $n \to \infty$ we obtain $V_\infty \geq TV_\infty$. So together we have $V_\infty = TV_\infty$. Now,

$$\begin{aligned} T\bar{v}(t,\lambda) &= \inf_{a\in A(i)} \sum_j \int U\big(e^{-\alpha(t+s)}\frac{\bar{c}}{\alpha} + \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{\alpha(t+s)})\big) Q(ds, j|i, a) \\ &\leq \inf_{a\in A(i)} \sum_j \int U\big(e^{-\alpha(t+s)}\frac{\bar{c}}{\alpha} + \lambda + \frac{\bar{c}}{\alpha}(e^{-\alpha t} - e^{\alpha(t+s)})\big) Q(ds, j|i, a) = \bar{v}(t,\lambda). \end{aligned}$$

Hence we have $T^n\bar{v}$ is a decreasing sequence. Moreover, we have by iteration

$$\begin{aligned} (T^n U)(t,i,\lambda) &= \inf_{\pi\in\Pi^M} \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u} C(X_u, A_u)du + \lambda\big)\big] \\ (T^n\bar{v})(t,i,\lambda) &= \inf_{\pi\in\Pi^M} \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\frac{\bar{c}}{\alpha}e^{-\alpha S_n} + \int_t^{S_n} e^{-\alpha u} C(X_u, A_u)du + \lambda\big)\big]. \end{aligned}$$

Again using the inequality $U(\lambda_1 + \lambda_2) - U(\lambda_1) \leq U'_-(\lambda_1)\lambda_2$, we have,

$$\begin{aligned}
0 &\leq (T^n\bar{v})(i,\lambda,t) - (T^nU)(i,\lambda,t) \\
&\leq \sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{(t,i,\lambda)}\big[U\big(\frac{\bar{c}}{\alpha}e^{-\alpha S_n} + \int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du + \lambda\big) \\
&- U\big(\int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du + \lambda\big)\big] \\
&\leq \epsilon_n(\lambda,t),
\end{aligned}$$

and $\epsilon_n(\lambda,t)$ converges to zero as $n \to \infty$. Hence $T^n\bar{v} \downarrow V_\infty$ as $n \to \infty$.

For the uniqueness let w be another solution of $w = Tw$ with $U(\lambda) \leq w \leq \bar{v}$. Then, by iteration we get $T^nU \leq w \leq T^n\bar{v}$, for all n. So, by taking $n \to \infty$ in the inequality and using the fact that $T^n\bar{v} \downarrow V_\infty$ and $T^nU \uparrow V_\infty$ we get the uniqueness.

(b) The existence of a minimizer follows from our Assumptions and standard measurable selection theorem. Now using the fact that $V_\infty(t,i,\lambda) \geq U(\lambda)$ we obtain

$$V_\infty = \lim_{n\to\infty} T^n_{f^*}V_\infty \geq \lim_{n\to\infty} T^n_{f^*}U = \lim_{n\to\infty} V_{n(f^*,f^*,\ldots)} = V_{f^*} \geq V_\infty,$$

where the last equation follows using dominated convergence theorem. Hence we get the optimality of the stationary policy given by f^*. □

Obviously it can be shown that for a policy $\pi = (f_0, f_1, ...) \in \Pi^M$ we have the following cost iteration: $V_{\infty\pi} = \lim_{n\to\infty}(T_{f_0}T_{f_1}...T_{f_{n-1}})U$. For a stationary policy $\pi = (f, f, ...)$ the cost iteration becomes $V_f = T_fV_f$.

The above Theorem tells us that from a computational point of view, the value function of the infinite horizon optimization problem can be approximated arbitrarily close by sandwiching between T^nU and $T^n\bar{v}$. Moreover, also the policy improvement algorithm works in this setting. For that, for $v \in B([0,\infty) \times E \times [0,\infty))$, we define the operator $(Lv)(t,i,\lambda,a) := \sum_j \int v\big(t+s, j, \lambda + \frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})\big)Q(ds,j|i,a)$. Also, for any $f \in F$ and $(t,i,\lambda) \in [0,\infty) \times E \times [0,\infty)$ we set $D(t,i,\lambda,f) := \{a \in A(i) : (Lv_f)(i,\lambda,t,a) < V_f(i,\lambda,t)\}$. Then by arguments analogous to Theorem 4 in Bäuerle and Rieder [2014], the following Theorem can be proved.

Theorem 4.3.2 (Policy Improvement)**.** *Suppose $f \in F$.*

(a) Define $h \in F$ by $h(\cdot) \in D(\cdot, f)$ if the set $D(., f)$ is non-empty and otherwise $h = f$. Then $V_h \leq V_f$ and the improvement is strict in states where $D(\cdot, f) \neq \phi$.
(b) If $D(\cdot, f) = \phi$ for all states, then $V_f = V_\infty$ and f defines an optimal policy.
(c) Suppose f_{k+1} is a minimizer of V_{f_k} for $k \geq 0$ where $f_0 = f$. Then $V_{f_{k+1}} \leq V_{f_k}$ and $\lim_{k\to\infty} V_{f_k} = V_\infty$.

4.3.2 Convex Utility Function

Now we look into the case of a convex utility function. For that $U : \mathbb{R}_+ \to \mathbb{R}$ be a convex utility function. The functions $V_{n\pi}$, V_n, $V_{\infty\pi}$, V_∞ are defined as in the previous section.

Theorem 4.3.3. *Under Assumptions 4.1.2 and 4.2.1, the conclusions of Theorem 4.3.1 hold for convex utility function as well.*

Proof. The proof of this Theorem is similar to that of Theorem 4.3.1. The main difference is that now we need to use the following property of convex function. Since $U : \mathbb{R}_+ \to \mathbb{R}$ strictly increasing and convex we have the inequality

$$U(\lambda_1 + \lambda_2) \leq U(\lambda_1) + U'_+(\lambda_1 + \lambda_2)\lambda_2, \quad \lambda_1, \lambda_2 \geq 0,$$

where U'_+ is the right-hand side derivative of U that exists since U is convex. Moreover, $U'_+(\lambda) \geq 0$ and U'_+ is increasing. For $(t, i, \lambda) \in [0, \infty) \times E \times [0, \infty)$ and $\pi \in \Pi$ we have,

$$\begin{aligned}
V_n(t,i,\lambda) \leq V_{n\pi}(t,i,\lambda) \leq V_{\infty\pi}(t,i,\lambda) &= \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\int_t^\infty e^{-\alpha u} C(X_u, A_u)du) + \lambda\big)\big] \\
&= \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u} C(X_u, A_u)du) + \lambda + \int_{S_n}^\infty e^{-\alpha u} C(X_u, A_u)du\big)\big] \\
&\leq \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\int_t^{S_n} e^{-\alpha u} C(X_u, A_u)du) + \lambda\big)\big] \\
&+ \mathbb{E}^\pi_{(t,i,\lambda)}\big[U'_+\big(\int_t^\infty e^{-\alpha u} C(X_u, A_u)du) + \lambda\big) \times \big(\int_{S_n}^\infty e^{-\alpha u} C(X_u, A_u)du\big)\big] \\
&\leq V_{n\pi}(t,i,\lambda) + U'_+(e^{-\alpha t}\frac{\bar{c}}{\alpha} + \lambda)\frac{\bar{c}}{\alpha}e^{-\alpha t}(1 - \epsilon + \epsilon e^{-\alpha\delta})^n, \\
&= V_{n\pi}(t,i,\lambda) + \delta_n(t,\lambda),
\end{aligned}$$

where $\delta_n(\lambda, t) = U'_+(e^{-\alpha t}\frac{\bar{c}}{\alpha} + \lambda)\frac{\bar{c}}{\alpha}e^{-\alpha t}(1 - \epsilon + \epsilon e^{-\alpha\delta})^n$. As $n \to \infty$, $\lim\limits_{n\to\infty} \delta_n(\lambda, t) = 0$. Taking

the infimum over all policies in the above inequality yields

$$V_n(t,i,\lambda) \leq V_\infty(t,i,\lambda) \leq V_n(t,i,\lambda) + \delta_n(t,\lambda).$$

Letting $n \to \infty$ yields $\lim_{n\to\infty} T^n U = V_\infty$. Again, using the same inequality we have

$$\begin{aligned}
0 &\leq (T^n\bar{v})(t,i,\lambda) - (T^n U)(t,i,\lambda) \\
&\leq \sup_{\pi\in\Pi} \mathbb{E}^\pi_{(t,i,\lambda)}\big[U\big(\frac{\bar{c}}{\alpha}e^{-\alpha S_n} + \int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du + \lambda\big) \\
&- U\big(\int_t^{S_n} e^{-\alpha u}C(X_u,A_u)du + \lambda\big)\big] \\
&\leq \delta_n(t,\lambda).
\end{aligned}$$

The rest of the arguments are same as Theorem 4.3.1. □

The policy improvement algorithm for convex utility functions works in exactly the same way as for the concave case. The following corollary is easy to deduce from Theorems 4.3.1 and 4.3.3.

Corollary 4.3.4. *In case $U(\lambda) = (\frac{1}{\gamma})e^{\gamma\lambda}$ with $\gamma \neq 0$, we obtain $V_\infty(t,i,\lambda) = e^{\gamma\lambda}h_\infty(t,i)$ and $J_\infty(i) = h_\infty(0,i)$. And the function h_∞ is the unique fixed point of*

$$h_\infty(t,i) = \inf_{a\in A(i)} \sum_j \int e^{\gamma\frac{C(i,a)}{\alpha}(e^{-\alpha t} - e^{-\alpha(t+s)})} h_\infty(t+s,j)Q(ds,j|i,a). \tag{4.3.1}$$

with $\frac{1}{\gamma} \leq h_\infty(t,i) \leq \frac{1}{\gamma}e^{\gamma e^{-\alpha t}\frac{\bar{c}}{\alpha}}$.

Remark 4.3.5. *Few remarks are in order.*

1. *We see from Corollaries 4.2.3 and 4.3.4 that in the case of exponential utility, the case which is classically referred to as the risk-sensitive control in literature, the value functions split, i.e. the value functions can be written as product of two functions, one a function of time and state and the other a function of the accumulated cost. Thus, for the exponential utility, it is clear from equations (4.2.6) and (4.3.1) that the minimizer does not depend on the accumulated cost and hence the optimal controls as given by*

Theorems 4.2.2, 4.3.1 and 4.3.3 will not depend on the accumulated cost and thus they belong to the original non-augmented policy sets.

2. *The dependence of the optimal policies on the jump times is not surprising. Because, in the risk-sensitive control literature it is known that in presence of discounting, for discrete-time Markov chains (see Bäuerle and Rieder [2014]), continuous-time Markov chains(see Ghosh and Saha [2014]) as well as for diffusions (see Menaldi and Robin [2005]), that optimal policies do depend on time.*

4.4 Example

We end the Chapter with a simple illustrative example. Consider a service center where customers arrive according to a Poisson process with rate 1. After completing a service and before starting a new service, the service person decides on an action which determines the service time of the next person waiting to be served. More precisely, if the service person chooses an action $a \in \{0, 1, 2, \dots, N\}$, then the service time of the next person will have distribution function $F(\cdot|a)$. Assume that $F(0|a) = 0$ for all a. Action 0 means the service person decides to take a vacation and thus there will be no service for $F(\cdot|0)$ distributed amount of time. Here the state is the number of persons waiting to be served at each decision epoch. Thus here, $E = \{0, 1, 2, \dots, \}$, $A = \{0, 1, 2, \dots, N\}$, $A(i) = \{0\}$ for $i = 0$ and $A(i) = A$ for $i \geq 1$. Also assume that if at a particular decision epoch the state is $i \in E$ and action $a \in A(i)$ is chosen, then the service person pays a cost at the rate $C(i, a)$ where C is a bounded non-negative real-valued function. The transition mechanism is given by

$$Q(dt, j|i, 0) = \begin{cases} e^{-t}\frac{t^k}{k!}F(dt|0) & \text{if } j = i + k \text{ for } k \geq 0, \\ 0 \quad \text{otherwise,} \end{cases}$$

and for $a \neq 0$,

$$Q(dt, j|i, a) = \begin{cases} e^{-t}\frac{t^k}{k!}F(dt|a) & \text{if } j = i + k - 1 \text{ for } k \geq 0, \\ 0 \quad \text{otherwise.} \end{cases}$$

Here Assumption 4.1.2 is satisfied because A is finite and $F(0|a) = 0$ for all $a \in A$. Assumption 4.2.1 is satisfied because $A(i)$ is finite for each $i \in E$. Thus the results of this Chapter apply to this example.

CHAPTER 5

Zero and Non-zero Sum Risk-sensitive Semi-Markov Games

In this Chapter we consider zero and non-zero sum risk-sensitive average criterion games for semi-Markov processes with a finite state space. For the zero-sum case, under suitable assumptions we show that the game has a value. We also establish the existence of a stationary saddle point equilibrium. For the non-zero sum case, under suitable assumptions we establish the existence of a stationary Nash equilibrium. This Chapter is organized as follows: In section 1, we describe the zero-sum game problem under consideration. In section 2 we introduce the optimality equations and establish its solution. In section 3, we describe the non-zero sum game problem. Section 4 establishes the existence of Nash equilibrium for the non-zero sum game. This Chapter is based on Bhabak and Saha [2023].

5.1 Zero-Sum Game Model

The risk-sensitive zero-sum semi-Markov game model that we consider here is given by

$$(S, A, B, \{A(i) \subset A, B(i) \subset B, i \in S\}, C(i,a,b), \{\rho_{(i,a,b)}(\cdot)\}, \{F_{i,a,b}\}, [p_{i,j}(a,b)]), \quad (5.1.1)$$

where,

- S is the state space, which is assumed to be finite and is endowed with the discrete

topology.

- The Borel spaces A and B are the action sets for player 1 and 2 respectively. And for each $i \in S$, $A(i) \subset A$, $B(i) \subset B$ are Borel subsets denoting the set of all admissible actions in state i for player 1 and 2 respectively.

- Define $K = \{(i, a, b) : i \in S, a \in A(i), b \in B(i)\}$ to be the set of admissible state-action pairs. Then $C : K \to \mathbb{R}$ is the immediate cost function for player 1 and immediate reward for player 2.

- For each $(i, a, b) \in K$, the mapping $\rho_{(i,a,b)} : [0, \infty) \to \mathbb{R}$ denotes the running cost function for player 1 and running reward function for player 2.

- $F_{i,a,b}$ is the sojourn time distribution function for both the players in state i under the actions a and b. It is assumed that the sojourn times are positive, so that

$$F_{i,a,b}(0) = 0, \quad (i, a, b) \in K \tag{5.1.2}$$

- Finally, $[p_{i,j}(a, b)]$ is the controlled transition law and satisfies $\sum_{j \in S} p_{i,j}(a, b) = 1$ for every $(i, a, b) \in K$.

The game evolves in the following manner. At the initial time $t = 0$, the process starts at $X_0 = i_0 \in S$. Suppose player 1 chooses an action $A_0 = a_0 \in A(i_0)$ and player 2 independently chooses an action $B_0 = b_0 \in B(i_0)$. As a result player 2 gets an immediate reward $C(i_0, a_0, b_0)$ from player 1. Player 1 also incurs a holding cost at the rate $\rho_{(i_0,a_0,b_0)}$. The process stays in state i for a random amount of time T_0 whose distribution function is given by F_{i_0,a_0,b_0} and then jumps to a new state $X_1 = i_1$ with probability $p_{i_0,i_1}(a_0, b_0)$. Immediately after the first transition, players 1 and 2 chooses actions $A_1 = a_1 \in A(i_1)$ and $B_1 = b_1 \in B(i_1)$. The same sequence of events as described above repeats itself. Let S_n be the time when the nth transition is completed, then

$$S_0 = 0 \quad \textit{and} \quad S_n = \sum_{i=0}^{n-1} T_i \quad n = 1, 2, ..., \tag{5.1.3}$$

where $T_n, n = 0, 1, 2, \ldots$ denotes the random sojourn times at the nth state. We denote the number of transitions N_t in the interval $[0, t]$ by

$$N_t = \sup\{n \in \mathbb{N} : S_n \leq t\}, \;\; t \geq 0. \tag{5.1.4}$$

Let $\mathcal{H}_n$ be the information available upto time S_n, i.e., $\mathcal{H}_0 = X_0$ and for $n \geq 1$, $\mathcal{H}_n = \{X_0, A_0, B_0, T_0, \ldots, X_{n-1}, A_{n-1}, B_{n-1}, T_{n-1}, X_n\}$, where for $n \geq 0$, X_n is the nth state, A_n and B_n are the actions of player 1 and 2 respectively at the nth transition time and T_n is the sojourn time at the nth state. For $n \geq 0$, we also define the admissible history spaces H_n by $H_0 = S$ and $H_n = K \times (0, \infty) \times H_{n-1}$ for $n = 1, 2, \ldots$. We endow these spaces with the Borel sigma-algebra. Now we introduce the concept of policies.

Definition 5.1.1. *A randomized history dependent policy or simply a policy for player 1 is a sequence $\pi^1 = \{\pi_n^1 \; : n \geq 0\}$ of stochastic kernels π_n^1 on A given H_n such that*

$$\pi_n^1(A(i_n)|h_n) = 1 \quad \forall \; h_n \in H_n, n = 0, 1, ...$$

A randomized history dependent policy for player 2 can be defined analogously.

Let Φ^1 be the set of all stochastic kernels ϕ^1 on A given S satisfying $\phi^1(A(i)|i) = 1$. A policy π^1 for player 1 is said to be stationary if there exists a stochastic kernel $\phi^1 \in \Phi^1$ such that $\pi_n^1(.|h_n) = \phi^1(.|i_n)$ for all $h_n = (i_0, a_0, b_0, s_0, \ldots, i_{n-1}, a_{n-1}, b_{n-1}, s_{n-1}, i_n) \in H_n$ and $n = 0, 1, ...$ We will identify a stationary policy π^1 with ϕ^1. Similarly stationary policies for player 2 can be defined.

For each $m = 1, 2$, Π_m and Φ^m represent the set of all randomized history dependent strategies and the set of all stationary strategies for player m, respectively. We will have the following assumptions on our model.

Assumption 5.1.2.

(i) For each $i \in S$, the set $A(i)$ and $B(i)$ are compact subsets of A and B.

(ii) For each $i, j \in S$, $(a, b) \to C(i, a, b)$ and $(a, b) \to p_{ij}(a, b)$ are continuous in $(a, b) \in A(i) \times B(i)$.

(iii) The family $\{F_{i,a,b}\}$ is supported on a compact interval and is weakly continuous, that

is, there exists $B > 0$ *such that*

$$F_{i,a,b}(B) = 1, \quad (i,a,b) \in K, \tag{5.1.5}$$

and for each $i \in S$ *and* u *bounded measurable,* $(a,b) \to \int_0^B u(s)dF_{i,a,b}(s)$ *is continuous in* $(a,b) \in A(i) \times B(i)$.

(iv) For every $i \in S$, *the mapping* $(a,b,s) \to \rho_{(i,a,b)}(s)$ *is continuous in* $(a,b,s) \in A(i) \times B(i) \times [0,B]$.

Since the spaces $A(i)$ and $B(i)$ are compact and the state space is finite, so it follows by Assumption 5.1.2 that,

$$M_\rho := \sup_{(i,a,b)\in K, s\in[0,B]} |\rho_{(i,a,b)}(s)| < \infty. \tag{5.1.6}$$

Given the initial state $X_0 = i$ and a pair of policies (π^1, π^2) , the distribution of $\{(X_n, A_n, B_n, T_n)\}$ is uniquely determined by the Tulcea theorem Arapostathis et al. [1993]. We denote such a distribution by $\mathbb{P}_i^{\pi^1,\pi^2}$, and $\mathbb{E}_i^{\pi^1,\pi^2}$ be the corresponding expectation operator. The following relations are satisfied almost surely under each distribution $\mathbb{P}_i^{\pi^1,\pi^2}$: For each $i, j \in S$, C Borel subset of A, D Borel subset of B and $n \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{P}_i^{\pi^1,\pi^2}[X_0 = i] = 1,\\
&\mathbb{P}_i^{\pi^1,\pi^2}[A_n \in C, B_n \in D|\mathcal{H}_n] = \pi_n^1(C|\mathcal{H}_n)\pi_n^2(D|\mathcal{H}_n),\\
&\mathbb{P}_i^{\pi^1,\pi^2}[T_n \le t|\mathcal{H}_n, A_n, B_n] = F_{X_n,A_n,B_n}(t),\\
&\mathbb{P}_i^{\pi^1,\pi^2}[X_{n+1} = j|\mathcal{H}_n, A_n, B_n, T_n] = p_{X_n,j}(A_n, B_n).
\end{aligned} \tag{5.1.7}$$

Now we describe the evaluation criterion for our game. The total cost incurred by player 1 and the total reward gained by player 2 up to time $t > 0$ is given by:

$$\mathcal{C}_t = \sum_{k=0}^{N_t-1} \left[C(X_k, A_k, B_k) + \int_0^{T_k} \rho_{(X_k,A_k,B_k)}(r)dr\right] + C(X_{N_t}, A_{N_t}, B_{N_t}) + \int_0^{t-S_{N_t}} \rho_{(X_{N_t},A_{N_t},B_{N_t})}(r)dr. \tag{5.1.8}$$

For risk-sensitivity parameter $\theta > 0$ and a policy pair (π^1, π^2) define,

$$J_\theta(i, \pi^1, \pi^2) := \limsup_{t\to\infty} \frac{1}{\theta t} \ln \left[\mathbb{E}_i^{\pi^1,\pi^2} \left(e^{\theta \mathcal{C}_t}\right)\right]. \tag{5.1.9}$$

Upper and lower values of the game are defined as below.

$$L(i, \theta) = \sup_{\pi^2\in\Pi_2} \inf_{\pi^1\in\Pi_1} J_\theta(i, \pi^1, \pi^2),$$

$$U(i, \theta) = \inf_{\pi^1\in\Pi_1} \sup_{\pi^2\in\Pi_2} J_\theta(i, \pi^1, \pi^2),$$

where $J_\theta(i, \pi^1, \pi^2)$ is defined in (5.1.9). $L(\cdot)$ is called the lower value of the game and $U(\cdot)$ is called the upper value of the game.

Definition 5.1.3. *If $U(i, \theta) = L(i, \theta)$ for all $i \in S$, then we say that the game has a value. If the game has a value, then the common function is referred to as the value function of the game and will be denoted by $V(\cdot)$.*

Here player 1 is interested in minimizing $J_\theta(i, \pi^1, \pi^2)$ over $\pi^1 \in \Pi_1$ for each $\pi^2 \in \Pi_2$, and player 2 wants to maximize $J_\theta(i, \pi^1, \pi^2)$ over $\pi^2 \in \Pi_2$ for each $\pi^1 \in \Pi_1$. This motivates the following definition.

Definition 5.1.4. *Suppose that the value of the game exists. A policy $\pi^{*^1} \in \Pi_1$ is said to be optimal for player 1, if for any $i \in S$,*

$$V(i, \theta) = \sup_{\pi^2\in\Pi_2} J_\theta(i, \pi^{*^1}, \pi^2), \ \ \forall i \in S.$$

Similarly, for player 2 a policy $\pi^{*^2} \in \Pi_2$ is said to be optimal, if for any $i \in S$,

$$V(i, \theta) = \inf_{\pi^1\in\Pi_1} J_\theta(i, \pi^1, \pi^{*^2}), \ \ \forall i \in S.$$

If $\pi^{*^m} \in \Pi_m$ is optimal for player $m (m = 1, 2)$, then $(\pi^{*^1}, \pi^{*^2}) \in \Pi_1 \times \Pi_2$ is called a saddle point equilibrium.

5.2 Analysis of Zero-Sum Game

For $i \in S$, let $\mathcal{P}(A(i))$ and $\mathcal{P}(B(i))$ denote the set of all probability measures on $A(i)$ and $B(i)$ respectively. The analysis of the zero-sum game crucially depends on the following equation.

$$e^{\theta h(i)} = \sup_{\varphi \in \mathcal{P}(B(i))} \inf_{\psi \in \mathcal{P}(A(i))} \Big[\int_{A(i)} \int_{B(i)} \psi(da)\varphi(db) e^{\theta C(i,a,b)} \int_0^B e^{\theta[\int_0^s \rho_{(i,a,b)}(t)dt - gs]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta h(j)} p_{ij}(a,b) \Big], \quad i \in S. \tag{5.2.1}$$

where g is a real number and $h(.)$ is a real function defined on the state space S. Using Assumption 5.1.2 and Fan's minimax theorem Fan [1952], equation (5.2.1) can also be written as:

$$e^{\theta h(i)} = \inf_{\psi \in \mathcal{P}(A(i))} \sup_{\varphi \in \mathcal{P}(B(i))} \Big[\int_{A(i)} \int_{B(i)} \psi(da)\varphi(db) e^{\theta C(i,a,b)} \int_0^B e^{\theta[\int_0^s \rho_{(i,a,b)}(t)dt - gs]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta h(j)} p_{ij}(a,b) \Big], \quad i \in S. \tag{5.2.2}$$

The importance of the above equations is illustrated by the next theorem.

Theorem 5.2.1. *Suppose Assumption 5.1.2 is satisfied and the pair $(g, h(.))$ satisfies the equation (5.2.1) and hence equation (5.2.2). Then the game has a value and is given by $g = V(i,\theta)$. Further if $\phi^{*1} \in \Phi^1$ is the outer minimizing selector of the right hand side of (5.2.2) and if $\phi^{*2} \in \Phi^2$ is the outer maximizing selector of the right hand side of (5.2.1), then (ϕ^{*1}, ϕ^{*2}) is a saddle point equilibrium.*

In order to prove Theorem 5.2.1, we need the following auxiliary lemma.

Lemma 5.2.2. *Suppose Assumption 5.1.2 holds. Then the following holds:*

(i) Given $\alpha \in (0,1)$, there exists an integer $r_\alpha > 0$ such that, for every $(i,a,b) \in K$, the inequality $\int_0^B e^{-rs} dF_{i,a,b}(s) \leq \alpha$ holds for every $r \geq r_\alpha$.

(ii) For each $\alpha \in (0,1)$, $t \geq 0$ and $n \in \mathbb{N}$, $\mathbb{P}_i^{\pi^1,\pi^2}[N_t \geq n] \leq \alpha^n e^{r_\alpha t}$ for all $i \in S$ and

$(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, *where* r_α *is as in part (ii). Thus,*

$$\mathbb{P}_i^{\pi^1,\pi^2}[N_t < \infty] = 1. \tag{5.2.3}$$

Proof. The proof is a simple generalization of Lemma 4.1 in Chávez-Rodríguez et al. [2016]. □

Proposition 5.2.3. *Suppose Assumption 5.1.2 is satisfied and the pair* $(g, h(.))$ *satisfies the equation* (5.2.1) *and hence equation* (5.2.2). *Then the following are true.*
For each $i \in S$, $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$ *and* $t > 0$*:*

$$e^{\theta h(i)} \geq \mathbb{E}_i^{\phi^{*1},\pi^2}\left[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(s)ds)-gS_{N_t+1}+h(X_{N_t+1})]}\right], \tag{5.2.4}$$

and also we have,

$$e^{\theta h(i)} \leq \mathbb{E}_i^{\pi^1,\phi^{*2}}\left[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(s)ds)-gS_{N_t+1}+h(X_{N_t+1})]}\right], \tag{5.2.5}$$

where ϕ^{*1} *and* ϕ^{*2} *are as in Theorem 5.2.1.*

Proof. From (5.2.2) we have for any $\varphi \in \mathcal{P}(B(i))$

$$e^{\theta h(i)} \geq \Big[\int_{A(i)}\int_{B(i)} \phi^{*1}(da|i)\varphi(db)e^{\theta C(i,a,b)}\int_0^B e^{\theta[\int_0^s \rho_{(i,a,b)}(t)dt-gs]}dF_{i,a,b}(s) \times \sum_{j\in S} e^{\theta h(j)}p_{ij}(a,b)\Big], \quad i \in S.$$

Thus for any $\pi^2 \in \Pi_2$ we have,

$$e^{\theta h(i)} \geq \mathbb{E}_i^{\phi^{*1},\pi^2}\left[e^{\theta[C(X_0,A_0,B_0)+\int_0^{T_0}\rho_{(X_0,A_0,B_0)}(t)dt-gT_0+h(X_1)]}\right], \quad i \in S. \tag{5.2.6}$$

More generally, via equations (5.1.7) it follows that for every $n \in \mathbb{N}$,

$$e^{\theta h(X_n)} \geq \mathbb{E}_i^{\phi^{*1},\pi^2}\left[e^{\theta[C(X_n,A_n,B_n)+\int_0^{T_n}\rho_{(X_n,A_n,B_n)}(t)dt-gT_n+h(X_{n+1})]}|\mathcal{H}_n\right], \quad i \in S. \tag{5.2.7}$$

We prove by induction that for every non-negative integer n,

$$\begin{aligned} e^{\theta h(i)} \geq & \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{N_t+1}+h(X_{N_t+1})]}1_{[N_t\leq n]}\right] \\ & + \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}+h(X_{n+1})]}1_{[N_t>n]}\right] \end{aligned} \tag{5.2.8}$$

To show this, from (5.2.6) we get,

$$\begin{aligned} e^{\theta h(i)} &\geq \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[C(X_0,A_0,B_0)+\int_0^{T_0}\rho_{(X_0,A_0,B_0)}(t)dt-gT_0+h(X_1)]}\right] \\ &= \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[C(X_0,A_0,B_0)+\int_0^{T_0}\rho_{(X_0,A_0,B_0)}(t)dt-gT_0+h(X_1)]}1_{[N_t=0]}\right] \\ &\quad + \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[C(X_0,A_0,B_0)+\int_0^{T_0}\rho_{(X_0,A_0,B_0)}(t)dt-gT_0+h(X_1)]}1_{[N_t>0]}\right]; \end{aligned}$$

since $T_1 = S_0$, hence we have the basis step for $n = 0$. Now suppose that (5.2.8) is true for a non-negative integer n. Then we have

$$\begin{aligned} & e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}+h(X_{n+1})]}1_{[N_t>n]} \\ &= e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}]}1_{[N_t\geq n+1]}e^{\theta h(X_{n+1})} \\ &\geq e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}]}1_{[N_t\geq n+1]} \\ &\times \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[C(X_{n+1},A_{n+1},B_{n+1})+\int_0^{T_{n+1}}\rho_{(X_{n+1},A_{n+1},B_{n+1})}(t)dt-gT_{n+1}+h(X_{n+2})]}|\mathcal{H}_{n+1}\right] \\ &= \mathbb{E}_i^{\phi^*1,\pi^2}\left[e^{\theta[\sum_{k=0}^{n+1}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(Xk,A_k,B_k)}(t)dt)-g[S_{n+1}+T_{n+1}]+h(X_{n+2})]}\times 1_{[N_t\geq n+1]}|\mathcal{H}_{n+1}\right] \end{aligned}$$

where (5.2.7) was used to deduce the first inequality, whereas the fact that the random variables $1_{[N_t\geq n+1]}$ and $\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt) - gS_{n+1} + h(X_{n+1})$ are $\sigma(\mathcal{H}_{n+1})$-measurable was used in the last step. Since $S_{n+2} = T_{n+1} + S_{n+1}$, by (5.1.3) it

follows that

$$\begin{aligned}
&\mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}+h(X_{n+1})]}1_{[N_t>n]}\Big]\\
&\geq \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n+1}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+2}+h(X_{n+2})]}1_{[N_t\geq n+1]}\Big]\\
&= \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n+1}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+2}+h(X_{n+2})]}1_{[N_t=n+1]}\Big]\\
&+ \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n+1}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+2}+h(X_{n+2})]}1_{[N_t>n+1]}\Big]\\
&= \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{N_t+1}+h(X_{N_t+1})]}1_{[N_t=n+1]}\Big]\\
&+ \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n+1}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+2}+h(X_{n+2})]}1_{[N_t>n+1]}\Big].
\end{aligned}$$

so, together with the induction hypothesis it follows that (5.2.8) is also valid for $n+1$. Thus the induction argument is complete. Then Monotone convergence theorem, together with (5.2.3) gives,

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{N_t+1}+h(X_{N_t+1})]}1_{[N_t\leq n]}\Big]\\
&= \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{N_t+1}+h(X_{N_t+1})]}\Big]. \qquad (5.2.9)
\end{aligned}$$

Now using Assumption 5.1.2 and Lemma 5.2.2 we get that

$$\mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta[\sum_{k=0}^{n}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(t)dt)-gS_{n+1}+h(X_{n+1})]}1_{[N_t>n]}\Big]\to 0 \quad as\ n\to\infty.$$

Now taking $n\to\infty$ on both the sides of (5.2.8) and using the last convergence and (5.2.9) we get the desired inequality (5.2.4).
The other inequality (5.2.5) also follows analogously starting from (5.2.1). □

Proof of Theorem 5.2.1 We have, $S_{N_t}\leq t<S_{N_t+1}=T_{N_t}+S_{N_t}$, for every $t>0$, and thus

$$0\leq t-S_{N_t}\leq T_{N_t}\leq B \ \ and \ \ S_{N_t+1}-t\leq T_{N_t}\leq B. \qquad (5.2.10)$$

Now from (5.1.8) we have

$$\sum_{k=0}^{N_t}\Big[C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(r)dr\Big]-gS_{N_t+1}$$
$$=(\mathcal{C}_t-tg)+\int_{t-S_{N_t}}^{T_{N_t}}\rho_{(X_{N_t},A_{N_t},B_{N_t})}(r)dr-(S_{N_t+1}-t)g.$$

and together with the equality (5.1.6) and (5.2.10) it follows that

$$\Big|\sum_{k=0}^{N_t}\Big[C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(r)dr\Big]-gS_{N_t+1}-(\mathcal{C}_t-tg)\Big|\leq B(M_\rho+|g|). \quad (5.2.11)$$

Using (5.2.5) we get that $e^{-2\theta||h||}\leq \mathbb{E}_i^{\pi^1,\phi^{*2}}\Big[e^{\theta\Big[\sum_{k=0}^{N_t}(C(X_k,A_k,B_k)+\int_0^{T_k}\rho_{(X_k,A_k,B_k)}(r)dr)\Big]-gS_{N_t+1}}\Big]$. Using (5.2.11), we have

$$e^{-2\theta||h||}\leq \mathbb{E}_i^{\pi^1,\phi^{*2}}\Big[e^{\theta[\mathcal{C}_t-tg+B(M_\rho+|g|)]}\Big],$$

so that $e^{-2\theta||h||-\theta B(M_\rho+|g|)+\theta tg}\leq \mathbb{E}_i^{\pi^1,\phi^{*2}}[e^{\theta\mathcal{C}_t}]$. Taking logarithm on both sides, dividing by θt and then taking limit $t\to\infty$ we get,

$$g\leq J_\theta(i,\pi^1,\phi^{*2}),\quad \forall i\in S.$$

For the other inequality consider inequality (5.2.4). Then proceeding similarly as above we have the following inequality,

$$e^{2\theta||h||+\theta B(M_\rho+|g|)+\theta tg}\geq \mathbb{E}_i^{\phi^{*1},\pi^2}\Big[e^{\theta\mathcal{C}_t}\Big].$$

Again taking logarithm on both sides, dividing by θt and then taking limit $t\to\infty$ we get,

$$g\geq J_\theta(i,\phi^{*1},\pi^2),\quad \forall i\in S.$$

Since (π^1, π^2) was arbitrary, we get

$$g \leq \inf_{\pi^1 \in \Pi_1} J_\theta(i, \pi^1, \phi^{*2}) \leq L(i, \theta) \leq U(i, \theta) \leq \sup_{\pi^2 \in \Pi_2} J_\theta(i, \phi^{*1}, \pi^2) \leq g.$$

Hence we have the desired conclusions.

In view of Theorem 5.2.1, in order to establish the existence of the value of the game and saddle point equilibrium, it suffices to show the existence of solution of the optimality equation (5.2.1). For that we impose one more assumption on our model.

Assumption 5.2.4. *Under each stationary policy, the embedded discrete-time Markov chain* $\{X_n\}$ *is irreducible.*

In order to establish the existence of solution of (5.2.1), we first consider risk-sensitive average criterion game problem for the discrete time process $\{X_n\}$. For that we consider policies (π^1, π^2), where for each positive integer n, the kernels (π^1_n, π^2_n) depends only on $X_0, A_0, B_0, X_1, ..., X_{n-1}, A_{n-1}, B_{n-1}, X_n$. Given a bounded continuous function D on K, define the discrete-time average at $i \in S$ under (π^1, π^2) by

$$V_{\theta,D}(i, \pi^1, \pi^2) := \limsup_{n \to \infty} \frac{1}{\theta n} ln\big(\mathbb{E}_i^{\pi^1, \pi^2}\big[e^{\theta \sum_{k=0}^{n-1} D(X_k, A_k, B_k)}\big]\big) \tag{5.2.12}$$

and θ-optimal discrete time average value function, if it exists, is given by

$$V^*_{\theta,D}(i) := \inf_{\pi^1} \sup_{\pi^2} V_{\theta,D}(i, \pi^1, \pi^2) = \sup_{\pi^2} \inf_{\pi^1} V_{\theta,D}(i, \pi^1, \pi^2) \tag{5.2.13}$$

It is easy to see that the value function $V^*_{\theta,D}(\cdot)$ satisfies the following.

$$V^*_{\theta,D}(\cdot) \leq V^*_{\theta,D_1}(\cdot) \ \ if \ \ D \leq D_1 \ \ and \ \ V^*_{\theta,c+D}(\cdot) = c + V^*_{\theta,D}(\cdot) \tag{5.2.14}$$

where $c \in \mathbb{R}$. Since $D \leq D_1 + ||D - D_1||$ it follows that $V^*_{\theta,D}(\cdot) \leq V^*_{\theta,D_1}(\cdot) + ||D - D_1||$. Similarly, by interchanging the roles of D and D_1 this yields that

$$||V^*_{\theta,D}(\cdot) - V^*_{\theta,D_1}(\cdot)|| \leq ||D - D_1||. \tag{5.2.15}$$

Observing that $V^*_{\theta,0} = 0$, the monotonicity property in (5.2.14) yields that, for bounded continuous functions D, D_1,

$$V^*_{\theta,D} \leq 0 \leq V^*_{\theta,D_1}, \text{ when } D \leq 0 \leq D_1. \tag{5.2.16}$$

We have the following theorem.

Theorem 5.2.5. *Under Assumptions 5.1.2 and 5.2.4, we have the following:*

(i) For each bounded continuous function D on K there exist $\mu_D \in \mathbb{R}$ and $h_D : S \to \mathbb{R}$ such that

$$\begin{aligned} e^{\theta[\mu_D + h_D(i)]} &= \sup_{\varphi \in \mathcal{P}(B(i))} \inf_{\psi \in \mathcal{P}(A(i))} \Big[\int_{A(i)} \int_{B(i)} \psi(da)\varphi(db) e^{\theta D(i,a,b)} \sum_{j \in S} p_{i,j}(a,b) e^{\theta h(j)} \Big] \\ &= \inf_{\psi \in \mathcal{P}(A(i))} \sup_{\varphi \in \mathcal{P}(B(i))} \Big[\int_{A(i)} \int_{B(i)} \psi(da)\varphi(db) e^{\theta D(i,a,b)} \sum_{j \in S} p_{i,j}(a,b) e^{\theta h(j)} \Big] \\ &\text{and} \quad \mu_D = V^*_{\theta,D}(i), \quad i \in S. \end{aligned} \tag{5.2.17}$$

(ii) For bounded continuous functions D, D_1,

$$|\mu_D - \mu_{D_1}| \leq ||D - D_1||. \tag{5.2.18}$$

Proof. The proof of (i) follows by putting together arguments and results from the existing literature on risk-sensitive control of discrete-time Markov chains. We just outline the steps and cite appropriate references.

Step 1: Using a standard contraction argument similar to the proof of Theorem 3.1(a) of Wei and Chen [2019b] it can be shown that for each $\beta \in (0,1)$ there exists function $V_\beta(\cdot)$ on

S satisfying

$$e^{\theta V_\beta(i)} = \sup_{\varphi\in\mathcal{P}(B(i))}\inf_{\psi\in\mathcal{P}(A(i))}\Big[\int_{A(i)}\int_{B(i)}\psi(da)\varphi(db)e^{\theta D(i,a,b)}\sum_{j\in S}p_{i,j}(a,b)e^{\theta\beta V_\beta(j)}\Big]$$
$$= \inf_{\psi\in\mathcal{P}(A(i))}\sup_{\varphi\in\mathcal{P}(B(i))}\Big[\int_{A(i)}\int_{B(i)}\psi(da)\varphi(db)e^{\theta D(i,a,b)}\sum_{j\in S}p_{i,j}(a,b)e^{\theta\beta V_\beta(j)}\Big],\quad i\in S. \tag{5.2.19}$$

Also it is true that $||V_\beta|| \le \frac{||D||}{1-\beta}$.

Step 2: Fix a sequence $\beta_n \uparrow 1$. For $n \ge 1$, define

$$z_{\beta_n} = \sup_{i\in S} V_{\beta_n}(i),\quad w_{\beta_n}(i) = V_{\beta_n}(i) - z_{\beta_n},\quad g_{\beta_n} = (1-\beta_n)z_{\beta_n}. \tag{5.2.20}$$

From (5.2.19) and (5.2.20) we get

$$e^{\theta w_{\beta_n}(i)+\theta g_{\beta_n}} = \sup_{\varphi\in\mathcal{P}(B(i))}\inf_{\psi\in\mathcal{P}(A(i))}\Big[\int_{A(i)}\int_{B(i)}\psi(da)\varphi(db)e^{\theta D(i,a,b)}\sum_{j\in S}p_{i,j}(a,b)e^{\theta\beta_n w_{\beta_n}(j)}\Big] \tag{5.2.21}$$
$$= \inf_{\psi\in\mathcal{P}(A(i))}\sup_{\varphi\in\mathcal{P}(B(i))}\Big[\int_{A(i)}\int_{B(i)}\psi(da)\varphi(db)e^{\theta D(i,a,b)}\sum_{j\in S}p_{i,j}(a,b)e^{\theta\beta_n w_{\beta_n}(j)}\Big],\quad i\in S. \tag{5.2.22}$$

Now arguing as in Proposition 3.1 in Wei and Chen [2019b], it can be shown that there exists a subsequence of β_n, which we relabel as β_n and function $h_D(i)$ and constant μ_D such that $h_D(i) = \lim_{n\to\infty} w_{\beta_n}(i)$ and $\mu_D = \lim_{n\to\infty} g_{\beta_n}$.

Step 3: Now taking limit in (5.2.21) and using Step 2 we get (5.2.17).

Step 4: The fact that $\mu_D = V^*_{\theta,D}(i)$ follows as in Lemma 2.3 in Cavazos-Cadena and Hernández-Hernández [2019].

The proof of (ii) is straightforward from part (i) and (5.2.15).

□

Lemma 5.2.6. *Suppose that Assumption 5.1.2 is valid and for each $g \in \mathbb{R}$ define the function*

$D_g : K \to \mathbb{R}$ *by*

$$D_g(i,a,b) = C(i,a,b) + \frac{1}{\theta} ln\Big(\int_0^B e^{\theta[\int_0^s \rho_{(i,a,b)}(t)dt - gs]} dF_{i,a,b}(s)\Big). \tag{5.2.23}$$

The following assertions hold.

(i) D_g *is bounded continuous on* K *for each* $g \in \mathbb{R}$.

(ii) $||D_g - D_{g_1}|| \leq B|g - g_1|$, $g, g_1 \in \mathbb{R}$.

(iii) There exist $g^- \geq 0$ *such that* $D_{g^-} \leq 0$.

(iv) $D_{g_+} \geq 0$ *for some* $g^+ \leq 0$.

Proof. The proof is a straight forward generalization of Lemma 6.1 in Chávez-Rodríguez et al. [2016]. □

We finally have the existence theorem.

Theorem 5.2.7. ***(Existence of solutions)*** *Under Assumptions 5.1.2 and 5.2.4, there exists* $g \in \mathbb{R}$ *and* $h : S \to \mathbb{R}$ *such that the optimality equation (5.2.1) is satisfied.*

Proof. For each g consider D_g given by (5.2.23). Combining Lemma 5.2.6 and Theorem 5.2.5 we get that μ_{D_g} is continuous in g. So again using Lemma 5.2.6 and intermediate value property we get the existence of a g such that $\mu_{D_g} = 0$. Hence we have the result from Theorem 5.2.5. □

5.3 Non-zero Sum Game Model

In the non-zero sum game model we assume that there is no immediate cost and individual players have there own running cost functions. For $m = 1, 2$, we denote the running cost function for player m by ρ^m. Here the evolution of the game is similar, except for the fact that upon taking their individual actions both players incur a holding cost upto the next transition. The definition of the policies is same as the zero-sum case. Thus, the total cost upto a positive time t for player 1 is given by:

$$\mathcal{C}_t^1 = \sum_{k=0}^{N_t-1} \int_0^{T_k} \rho^1_{(X_k,A_k,B_k)}(r)dr + \int_0^{t-S_{N_t}} \rho^1_{(X_{N_t},A_{N_t},B_{N_t})}(r)dr, \tag{5.3.1}$$

while for player 2 it is given by:

$$\mathcal{C}_t^2 = \sum_{k=0}^{N_t-1} \int_0^{T_k} \rho^2_{(X_k,A_k,B_k)}(r)dr + \int_0^{t-S_{N_t}} \rho^2_{(X_{N_t},A_{N_t},B_{N_t})}(r)dr. \tag{5.3.2}$$

Here the objective of each player is to minimize their own average costs.

Definition 5.3.1. *Fix a pair of policies $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$. For $m = 1, 2$, define the value function for player 1 as $V_\theta^1(\pi^2) = \inf_{\pi^1} J_\theta^1(i, \pi^1, \pi^2)$, where J_θ^1 is given by (5.1.9), with $\mathcal{C}_t$ replaced by $\mathcal{C}_t^1$. Similarly, the value function for player 2 is given by $V_\theta^2(\pi^1) = \inf_{\pi^2} J_\theta^2(i, \pi^1, \pi^2)$, where J_θ^2 is given by (5.1.9), with $\mathcal{C}_t$ replaced by $\mathcal{C}_t^2$.*

Definition 5.3.2. *(Nash equilibrium) A pair of policies $(\pi^{*^1}, \pi^{*^2}) \in \Pi_1 \times \Pi_2$ is called a Nash equilibrium for the non-zero sum game if*

$$J_\theta^1(i, \pi^{*^1}, \pi^{*^2}) \leq J_\theta^1(i, \pi^1, \pi^{*^2}) \quad \text{and}$$
$$J_\theta^2(i, \pi^{*^1}, \pi^{*^2}) \leq J_\theta^2(i, \pi^{*^1}, \pi^2),$$

for all $i \in S$ and $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$.

5.4 Analysis of Non-Zero Sum Game

We wish to establish the existence of Nash equilibrium for the non-zero sum game. To that end we, just like in the zero-sum case first consider a discrete time non-zero sum game given by the embedded Markov chain. Given two bounded continuous functions D_1 and D_2 on K, we define for $i \in S$, under (π^1, π^2), the discrete-time cost functional for player $m, m = 1, 2$ by

$$V_{\theta,D_m}(i, \pi^1, \pi^2) := \limsup_{n\to\infty} \frac{1}{\theta n} ln\big(\mathbb{E}_i^{\pi^1,\pi^2}\big[e^{\theta \sum_{k=0}^{n-1} D_m(X_k,A_k,B_k)}\big]\big). \tag{5.4.1}$$

We have the following discrete-time theorem.

Theorem 5.4.1. *Suppose that Assumptions 5.1.2 and 5.2.4 are satisfied. Fix a pair of stationary strategies (ϕ^1, ϕ^2). Then there exist functions y^{ϕ^1} and y^{ϕ^2} on S and constants μ^{ϕ^1}*

and μ^{ϕ^2} *such that the following are true.*
(i)

$$e^{\theta y^{\phi^2}(i)+\theta\mu^{\phi^2}} = \inf_{\psi\in\mathcal{P}(A(i))}\left\{\int_{A(i)}\int_{B(i)} e^{\theta D_1(i,a,b)}\times\sum_{j\in S} e^{\theta y^{\phi^2}(j)}p_{i,j}(a,b)\phi^2(db|i)\psi(da)\right\}\ \forall i\in S, \tag{5.4.2}$$

and $\mu^{\phi^2} = \inf_{\pi^1} V_{\theta,D_1}(i,\pi^1,\phi^2)$.
(ii)

$$e^{\theta y^{\phi^1}(i)+\theta\mu^{\phi^1}} = \inf_{\varphi\in\mathcal{P}(B(i))}\left\{\int_{A(i)}\int_{B(i)} e^{\theta D_2(i,a,b)}\times\sum_{j\in S} e^{\theta y^{\phi^1}(j)}p_{i,j}(a,b)\varphi(db)\phi^1(da|i)\right\}\ \forall i\in S, \tag{5.4.3}$$

and $\mu^{\phi^1} = \inf_{\pi^2} V_{\theta,D_2}(i,\phi^1,\pi^2)$.

Proof. The proof again follows by putting together arguments and result from the existing literature. So like in the zero-sum case we outline the steps and cite appropriate references. **Step 1:** Let $\alpha\in(0,1)$. Then again using a contraction argument similar to the proof of Theorem 3.1(a) of Wei and Chen [2019b] the following can be shown.

(a) For each fixed $\phi^2\in\Phi^2$, there exists a function $w^{\phi^2,\alpha}$ such that

$$e^{\theta w^{\phi^2,\alpha}(i)} = \inf_{\psi\in\mathcal{P}(A(i))}\left[\int_{A(i)}\int_{B(i)} e^{\theta D_1(i,a,b)}\times\sum_{j\in S} e^{\theta\alpha w^{\phi^2,\alpha}(j)}p_{i,j}(a,b)\phi^2(db|i)\psi(da)\right], \tag{5.4.4}$$

for all $i\in S$.

(b) For each fixed $\phi^1\in\Phi^1$, there exists a function $w^{\phi^1,\alpha}$ on S such that

$$e^{\theta w^{\phi^1,\alpha}(i)} = \inf_{\varphi\in\mathcal{P}(B(i))}\left[\int_{A(i)}\int_{B(i)} e^{\theta D_2(i,a,b)}\times\sum_{j\in S} e^{\theta\alpha w^{\phi^1,\alpha}(j)}p_{i,j}(a,b)\varphi(db)\phi^1(da|i)\right], \tag{5.4.5}$$

for all $i \in S$.

Step 2: Fix an arbitrary sequence $\{\alpha_n\} \in (0,1)$ satisfying $\alpha_n \uparrow 1$, as $n \to \infty$. For each $n \geq 1$ set

$$\gamma^{\phi^2}_{\alpha_n} = \sup_{i \in S} w^{\phi^2,\alpha_n}(i), \qquad \gamma^{\phi^1}_{\alpha_n} = \sup_{i \in S} w^{\phi^1,\alpha_n}(i),$$
$$\mu^{\phi^2}_{\alpha_n} = (1-\alpha_n)\gamma^{\phi^2}_{\alpha_n}, \qquad \mu^{\phi^1}_{\alpha_n} = (1-\alpha_n)\gamma^{\phi^1}_{\alpha_n},$$
$$v^{\phi^2}_{\alpha_n}(i) = w^{\phi^2,\alpha_n}(i) - \gamma^{\phi^2}_{\alpha_n}, \qquad v^{2}_{\alpha_n}(i) = w^{\phi^1,\alpha_n}(i) - \gamma^{\phi^1}_{\alpha_n}.$$

Now arguing as in Proposition 3.1 in Wei and Chen [2019b], it can be shown that there exists functions y^{ϕ^1} and y^{ϕ^2} and constants μ^{ϕ^1} and μ^{ϕ^2} such that along a subsequence $y^{\phi^m}(i) = \lim_{n\to\infty} v^{\phi^m}_{\alpha_n}(i)$ and $\mu^{\phi^m} = \lim_{n\to\infty} \mu^{\phi^m}_{\alpha_n}$, for $m = 1, 2$. From Step 1, using boundedness of $D_m, m = 1, 2$, it is easy to see that $\mu^{\phi^m}_{\alpha_n}$ is bounded and hence existence of a convergent subsequence is trivial. It follows from the definition that $v^{\phi^m}_{\alpha_n}(i) \leq 0$ for each i. The finiteness of the state space and irreducibility ensures the lower bound.

Step 3: First we rewrite equations (5.4.4) and (5.4.5) in terms of the quantities defined in Step 2. Then taking limit $n \to \infty$ and using Step 2, we obtain equations (5.4.2) and (5.4.3) respectively.

Step 4: The interpretations of μ^{ϕ^1} and μ^{ϕ^2} follows by similar arguments as in Theorem 4.1 of Wei and Chen [2019b]. □

In order to establish the existence of a Nash equilibrium we need the following additional assumption.

Assumption 5.4.2. *Fix a state $i^* \in S$. Define $\tau^* = \inf\{n \geq 1 : X_n = i^*\}$. We assume that there exist $R > 1$ and $0 < M < \infty$ such that*

$$\sup_{\phi^1 \in \Phi^1} \sup_{\phi^2 \in \Phi^2} \sup_{i \in S} \mathbb{E}^{\phi^1,\phi^2}_i \left[R^{\tau^*}\right] \leq M.$$

For this R, we further assume that θ is such that

$$e^{2\theta B M_\rho} \leq R,$$

where $M_\rho = \max\{M_{\rho^1}, M_{\rho^2}\}$ where M_{ρ^i} is as in (5.1.6) with ρ replaced by ρ^i.

By Proposition 3 in Basu and Ghosh [2018], it follows that aperiodicity of the embedded discrete time Markov chain $\{X_n\}$ under each pair of stationary policies is a sufficient condition for the first part of Assumption 5.4.2. The remaining assumptions in the cited proposition follows by finiteness of state space, continuity of $p_{ij}(\cdot)$ and Assumption 5.2.4. Next we obtain the following theorem as a consequence of the previous theorem.

Theorem 5.4.3. *Let Assumptions 5.1.2, 5.2.4 and 5.4.2 hold. Fix $(\phi^1, \phi^2) \in \Phi^1 \times \Phi^2$. Then there exist constants g^{ϕ^1}, g^{ϕ^2}, real valued functions h^{ϕ^1}, h^{ϕ^2} on S with $h^{\phi^1}(i^*) = h^{\phi^2}(i^*) = 0$, such that the following are true.*
(i)

$$e^{\theta h^{\phi^2}(i)} = \inf_{\psi \in \mathcal{P}(A(i))} \left\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2} s]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta h^{\phi^2}(j)} p_{i,j}(a,b) \phi^2(db|i) \psi(da) \right\}, \qquad \forall i \in S. \tag{5.4.6}$$

(ii)

$$e^{\theta h^{\phi^1}(i)} = \inf_{\varphi \in \mathcal{P}(B(i))} \left\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^2_{(i,a,b)}(t)dt - g^{\phi^1} s]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta h^{\phi^1}(j)} p_{i,j}(a,b) \phi^1(da|i) \varphi(db) \right\}, \qquad \forall i \in S. \tag{5.4.7}$$

(iii) $g^{\phi^2} = \inf_{\pi^1 \in \Pi_1} J^1_\theta(i, \pi^1, \phi^2)$ *for all* i *and* $g^{\phi^1} = \inf_{\pi^2 \in \Pi_2} J^2_\theta(i, \phi^1, \pi^2)$ *for all* i.
(iv) For $(i,a,b) \in K$, *let* $D_1^{g^{\phi^2}}(i,a,b) = \frac{1}{\theta} ln \left(\int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2} s]} dF_{i,a,b}(s) \right)$ *and* $D_2^{g^{\phi^1}}(i,a,b) = \frac{1}{\theta} ln \left(\int_0^B e^{\theta[\int_0^s \rho^2_{(i,a,b)}(t)dt - g^{\phi^1} s]} dF_{i,a,b}(s) \right)$. *Then* h^{ϕ^1} *and* h^{ϕ^2} *have the following representations.*

$$h^{\phi^2}(i) = \inf_{\phi^1 \in \Phi^1} \frac{1}{\theta} ln \mathbb{E}_i^{\phi^1, \phi^2} \left[e^{\theta \sum_{k=0}^{\tau^* - 1} D_1^{g^{\phi^2}}(X_k, A_k, B_k)} \right], \ \forall i \in S \setminus \{i^*\}.$$

$$h^{\phi^1}(i) = \inf_{\phi^2 \in \Phi^2} \frac{1}{\theta} ln \mathbb{E}_i^{\phi^1,\phi^2} \left[e^{\theta \sum_{k=0}^{\tau^*-1} D_2^{g^{\phi^1}}(X_k, A_k, B_k)} \right], \ \forall i \in S \setminus \{i^*\}.$$

Proof. The proof of (i) and (ii) follows from Theorem 5.4.1 by a similar trick as in Theorem 5.2.7 of the zero-sum game section. Proof of (iii) follows by arguments similar to Theorem 5.2.1. Finally, the proof of (iv) follows by arguments similar to Lemma 8.1 in Wei and Chen [2019b]. Note that in the proofs of (i), (ii) and (iii) we do not require Assumption 5.4.2. It is in the proof of (iv) where we need Assumption 5.4.2. □

Now, fix any $(\phi^1, \phi^2) \in \Phi^1 \times \Phi^2$. Define

$$\begin{aligned}
\Delta(\phi^2) = \Big\{ \phi^{*^1} \in \Phi^1 : & \text{ for each } i \in S, \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2}s]} dF_{i,a,b}(s) \\
& \times \sum_{j \in S} e^{\theta h^{\phi^2}(j)} p_{i,j}(a,b) \phi^2(db|i) \phi^{*^1}(da|i) \\
= & \inf_{\psi \in \mathcal{P}(A(i))} \Big\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2}s]} dF_{i,a,b}(s) \\
& \times \sum_{j \in S} e^{\theta h^{\phi^2}(j)} p_{i,j}(a,b) \phi^2(db|i) \psi(da) \Big\} \Big\},
\end{aligned}$$

and

$$\begin{aligned}
\Delta(\phi^1) = \Big\{ \phi^{*^2} \in \Phi^2 : & \text{ for each } i \in S, \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^2_{(i,a,b)}(t)dt - g^{\phi^1}s]} dF_{i,a,b}(s) \\
& \times \sum_{j \in S} e^{\theta h^{\phi^1}(j)} p_{i,j}(a,b) \phi^{*^2}(db|i) \phi^1(da|i) \\
= & \inf_{\varphi \in \mathcal{P}(B(i))} \Big\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^2_{(i,a,b)}(t)dt - g^{\phi^1}s]} dF_{i,a,b}(s) \\
& \times \sum_{j \in S} e^{\theta h^{\phi^1}(j)} p_{i,j}(a,b) \varphi(db) \phi^1(da|i) \Big\} \Big\}.
\end{aligned}$$

It follows from our assumptions that the sets $\Delta(\phi^2)$ and $\Delta(\phi^1)$ are non-empty.

Lemma 5.4.4. *Suppose that Assumptions 5.1.2 and 5.2.4 are true. For each* $(\phi^1, \phi^2) \in$

$\Phi^1 \times \Phi^2$, $\Delta(\phi^2) \times \Delta(\phi^1)$ *is convex and compact with respect to the weak topology.*

Proof. We first show that $\Delta(\phi^2)$ is convex. For that let $\tilde{\phi}^1, \tilde{\psi}^1 \in \Delta(\phi^2)$ and $\lambda \in [0,1]$, define: $\phi^1_\beta(\cdot|i) = \lambda\tilde{\phi}^1(\cdot|i) + (1-\lambda)\tilde{\psi}^1(.|i)$ for all $i \in S$. By writing down the expression of ϕ^1_β one easily gets that $\phi^1_\beta \in \Delta(\phi^2)$. Thus $\Delta(\phi^2)$ is convex. By analogous argument $\Delta(\phi^1)$ is also convex, which together implies that $\Delta(\phi^2) \times \Delta(\phi^1)$ is convex.
By the compactness of $\Phi^1 \times \Phi^2$ and the fact that $\Delta(\phi^2) \times \Delta(\phi^1)$ is a subset of $\Phi^1 \times \Phi^2$, its enough to show that $\Delta(\phi^2) \times \Delta(\phi^1)$ is a closed subset. First we show that $\Delta(\phi^2)$ is a closed subset of the compact space Φ^1. Let $\{\phi^{*^1}_n\} \subset \Delta(\phi^2)$ be an arbitrary sequence converging to $\phi^{*^1} \in \Phi^1$, and $G(i,a) := \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2}s]} dF_{i,a,b}(s) \times \sum_{j\in S} e^{\theta h^{\phi^2}(j)} p_{i,j}(a,b)\phi^2(db|i)$ for $i \in S$ and $a \in A(i)$. By Assumption 5.1.2, we have that for each $i \in S$, $G(i,.)$ is a bounded continuous function on $A(i)$. Thus by definition of weak topology we obtain

$$\int_{A(i)} G(i,a)\phi^{*^1}_n(da|i) \to \int_{A(i)} G(i,a)\phi^{*^1}(da|i).$$

as $n \to \infty$. Since $\{\phi^{*^1}_n\} \subset \Delta(\phi^2)$

$$\int_{A(i)} G(i,a)\phi^{*^1}_n(da|i) = \inf_{\mu\in\mathcal{P}(A(i))} \left\{ \int_{A(i)} G(i,a)\mu(da) \right\}$$

for all $n = 1, 2, \ldots$. Hence, we have $\phi^{*^1} \in \Delta(\phi^2)$. Thus, $\Delta(\phi^2)$ is closed. Similarly, $\Delta(\phi^1)$ is closed. So combining we get $\Delta(\phi^2) \times \Delta(\phi^1)$ is convex and compact.

Lemma 5.4.5. *Suppose that Assumptions 5.1.2, 5.2.4 and 5.4.2 hold. For each $i \in S$, the functions $\phi^1 \to h^{\phi^1}(i)$ and $\phi^2 \to h^{\phi^2}(i)$ are continuous in $\phi^1 \in \Phi^1$ and $\phi^2 \in \Phi^2$ respectively. Continuity also holds for the functions $\phi^1 \to g^{\phi^1}$ and $\phi^2 \to g^{\phi^2}$.*

Proof. By (iii) of Theorem 5.4.3, we have $|g^{\phi^1}| \le M_\rho$ and $|g^{\phi^1}| \le M_\rho$. We also have $||D_1^{\phi^2}|| \le 2BM_\rho$ and $||D_2^{\phi^1}|| \le 2BM_\rho$. Thus by Assumption 5.4.2, we have $h^{\phi^m}(i) \le \frac{1}{\theta} lnM$ for $m = 1, 2$ and for all $i \in S$. Now Assumption 5.4.2 also implies that $\sup_{\pi^1\in\Phi^1} \sup_{\phi^2\in\Phi^2} \sup_{i\in S} \mathbb{E}_i^{\phi^1,\phi^2}\tau^* \le K$, for some K. So by Jensen's inequality we have $h^{\phi^m}(i) \ge -2KBM_\rho$ for $m = 1, 2$ and for all $i \in S$. Now suppose $\phi^2_n \to \phi^2$. Let us consider subsequences $\{g^{\phi^2_{n_k}}\}$, $\{h^{\phi^2_{n_k}}(i)\}$. We will get a further subsequence such that $g^{\phi^2_{n_l}} \to g^*$ for some constant g^* and $h^{\phi^2_{n_l}}(j) \to u(j)$ for

all $j \in S$ for some function u on S. We have,

$$e^{\theta h^{\phi^2_{n_l}}(i)} = \inf_{\psi \in \mathcal{P}(A(i))} \left\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2_{n_l}} s]} dF_{i,a,b}(s) \right.$$
$$\left. \times \sum_{j \in S} e^{\theta h^{\phi^2_{n_l}}(j)} p_{i,j}(a,b) \phi^2_{n_l}(db|i) \psi(da) \right\} \qquad \forall i \in S. \tag{5.4.8}$$

Now by our assumptions, definition of weak convergence and extended Fatou's lemma (Lemma 8.3.7 in Hernández-Lerma and Lasserre [1999]), we obtain by taking limit $l \to \infty$ in the above equation,

$$e^{\theta u(i)} = \inf_{\psi \in \mathcal{P}(A(i))} \left\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^* s]} dF_{i,a,b}(s) \right.$$
$$\left. \times \sum_{j \in S} e^{\theta u(j)} p_{i,j}(a,b) \phi^2(db|i) \psi(da) \right\} \qquad \forall i \in S. \tag{5.4.9}$$

Thus again arguing as in Theorem 5.4.3, we will get that $g^* = \inf_{\pi^1 \in \Pi_1} J^1_\theta(i, \pi^1, \phi^2) = g^{\phi^2}$ and $u(i) = \inf_{\phi^1 \in \Phi^1} \frac{1}{\theta} ln \mathbb{E}^{\phi^1, \phi^2}_i \left[e^{\theta \sum_{k=0}^{\tau^*-1} D_1^{g^{\phi^2}}(X_k, A_k, B_k)} \right] = h^{\phi^2}(i)\ \forall i \in S \setminus \{i^*\}$. Since every subsequence has a further subsequence which converges to the same limit, we are done. □

Now we state the main theorem of this section.

Theorem 5.4.6. *Suppose that Assumptions 5.1.2, 5.2.4 and 5.4.2 hold. There exists constants g^{*^1}, g^{*^2}, real valued functions y^{*^1}, y^{*^2} on S and a pair of stationary policies $(\phi^{*^1}, \phi^{*^2}) \in \Phi^1 \times \Phi^2$ such that*

$$e^{\theta y^{*^1}(i)} = \inf_{\psi \in \mathcal{P}(A(i))} \left\{ \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{*^1} s]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta y^{*^1}(j)} p_{i,j}(a,b) \phi^{*^2}(db|i) \psi(da) \right\}$$
$$= \int_{A(i)} \int_{B(i)} \int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{*^1} s]} dF_{i,a,b}(s) \times \sum_{j \in S} e^{\theta y^{*^1}(j)} p_{i,j}(a,b) \phi^{*^2}(db|i) \phi^{*^1}(da|i), \tag{5.4.10}$$

and

$$e^{\theta y^{*^2}(i)} = \inf_{\varphi\in\mathcal{P}(B(i))}\left\{\int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho^2_{(i,a,b)}(t)dt - g^{*^2}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta y^{*^2}(j)}p_{i,j}(a,b)\varphi(db)\phi^{*^1}(da|i)\right\}$$
$$= \int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho_{(i,a,b)}(t)dt - g^{*^2}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta y^{*^2}(j)}p_{i,j}(a,b)\phi^{*^2}(db|i)\phi^{*^1}(da|i), \quad (5.4.11)$$

for all $i\in S$. Moreover, the pair of policies $(\phi^{*^1},\phi^{*^2})\in\Phi^1\times\Phi^2$ is a Nash-equilibrium and we have $J_m(i,\phi^{*^1},\phi^{*^2}) = g^{*^m}$ for all $i\in S$ and $m=1,2$.

Proof. Let $2^{\Phi^1\times\Phi^2}$ be the power set of $\Phi^1\times\Phi^2$ and define the multi function $\Psi:\Phi^1\times\Phi^2\to 2^{\Phi^1\times\Phi^2}$ by $\Psi((\phi^1,\phi^2)) = \Delta(\phi^2)\times\Delta(\phi^1)$. Next we show that Ψ has a closed graph. Let $\{(\phi^1_n,\phi^2_n)\}\subset\Phi^1\times\Phi^2$ and $\{(\phi^{*^1}_n,\phi^{*^2}_n)\}\subset\Phi^1\times\Phi^2$ be arbitrary sequences with $\{(\phi^{*^1}_n,\phi^{*^2}_n)\}\in\Psi((\phi^1_n,\phi^2_n))$ and $\{(\phi^1_n,\phi^2_n)\}$ and $\{(\phi^{*^1}_n,\phi^{*^2}_n)\}$ converges to $(\bar\phi^1,\bar\phi^2)$ and $(\bar\phi^{*^1},\bar\phi^{*^2})$, respectively. Then by the definition of $\Delta(\phi^2_n)$, we have

$$\inf_{\psi\in\mathcal{P}(A(i))}\left\{\int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2_n}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta h^{\phi^2_n}(j)}p_{i,j}(a,b)\phi^2_n(db|i)\psi(da)\right\}$$
$$= \int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\phi^2_n}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta h^{\phi^2_n}(j)}p_{i,j}(a,b)\phi^2_n(db|i)\phi^{*^1}_n(da|i). \quad (5.4.12)$$

Now using our assumptions, Lemma 5.4.5 and extended Fatou's lemma (Lemma 8.3.7 in Hernández-Lerma and Lasserre [1999]) we obtain by taking limit $n\to\infty$ in (5.4.12),

$$\inf_{\psi\in\mathcal{P}(A(i))}\left\{\int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\bar\phi^2}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta h^{\bar\phi^2}(j)}p_{i,j}(a,b)\bar\phi^2(db|i)\psi(da)\right\}$$
$$= \int_{A(i)}\int_{B(i)}\int_0^B e^{\theta[\int_0^s \rho^1_{(i,a,b)}(t)dt - g^{\bar\phi^2}s]}dF_{i,a,b}(s)\times\sum_{j\in S}e^{\theta h^{\bar\phi^2}(j)}p_{i,j}(a,b)\bar\phi^2(db|i)\bar\phi^{*^1}(da|i).$$

for all $i\in S$, which implies $\bar\phi^{*^1}\in\Delta(\bar\phi^2)$. Using similar arguments as above, we can also show that $\bar\phi^{*^2}\in\Delta(\bar\phi^1)$. Hence, the multi function Ψ has a closed graph. Therefore by Fan's fixed point theorem Fan [1952] we have the existence of $(\phi^{*^1},\phi^{*^2})\in\Phi^1\times\Phi^2$ such that $((\phi^{*^1},\phi^{*^2}))\in\Delta(\phi^{*^2})\times\Delta(\phi^{*^1})$. Now using Theorem 5.4.3 we obtain solution to the coupled system of equations (5.4.10) and (5.4.11).

Now for the Nash equilibrium part, it follows from (5.4.10) and arguments similar to Theorem 5.2.1, that

$$g^{*^1} = J_1(i, \phi^{*^1}, \phi^{*^2}) = V_\theta^1(\phi^{*^2}).$$

Analogously, starting from (5.4.11) it can be shown that

$$g^{*^2} = J_2(i, \phi^{*^1}, \phi^{*^2}) = V_\theta^2(\phi^{*^1}).$$

Hence we are done. □

Remark 5.4.7. *In this Chapter we have studied both zero-sum and non-zero sum risk-sensitive average criterion games for semi-Markov process. Here we assume that the state space is finite and the sojourn time distributions are supported within a fixed compact interval. So it remains an open problem to extend the setting to more general state space and sojourn time distributions. Note that such a problem is also open for the control case as well, because in Chávez-Rodríguez et al. [2016] where the control problem is studied similar assumptions are made and crucially used in the analysis. Also, in the non-zero sum case we assume that the immediate cost is zero. The reason behind that is, in presence of immediate cost, in the expression for cost accumulated upto time t, the component corresponding to immediate cost involves a summation whose upper limit is a random variable equal to the number of transitions upto time t. So in presence of that we are unable to establish uniform(over stationary policies) bounds as in the first two lines of the proof of Lemma 5.4.5. This lemma is crucially used in proving the existence of Nash equilibrium.*

CHAPTER 6

Partially Observable Discrete-time Discounted Markov Games with General Utility

In this Chapter, we investigate partially observable zero sum games where the state process is a discrete time Markov chain. We consider a general utility function in the optimization criterion. We show the existence of value for both finite and infinite horizon games and also establish the existence of optimal polices. The main step involves converting the partially observable game into a completely observable game which also keeps track of the total discounted accumulated reward/cost. This Chapter is organized as follows. In Section 1, we describe our game model. Section 2 deals with finite horizon problem. Finally in Section 3 we investigate the infinite horizon problem as a limit of finite horizon problem. This Chapter is based on Bhabak and Saha [2022a].

6.1 Zero-Sum Game Model

The partially observable risk-sensitive zero-sum Markov game model that we are interested in can be represented by the tuple

$$(X, Y, A, B, \{A(x) \subset A, B(x) \subset B, x \in X\}, C(x, y, a, b), Q), \tag{6.1.1}$$

where each individual component has the following interpretation.

- X, Y are the Borel spaces, X represents the observable state space and Y is the non-observable space. We endow these spaces with Borel sigma-algebras.

- The Borel spaces A and B are the action sets for player 1 and 2 respectively. And for each $x \in X$, $A(x) \subset A$, $B(x) \subset B$ are Borel subsets denoting the set of all admissible actions in state $x \in X$ for player 1 and 2 respectively.

- Define $K = \{(x, a, b) : x \in X, a \in A(x), b \in B(x)\}$ to be the set of admissible state-action pairs, which is assumed to be a measurable subset of $X \times A \times B$. Then $C : K \times Y \to \mathbb{R}$ is the immediate reward function for player 1 and immediate cost function for player 2 respectively.

- For each $(x, y, a, b) \in K \times Y$, the mapping $Q(C|x, y, a, b)$ is the transition probability that the next pair is in $C \in \mathcal{B}(X \times Y)$, where $\mathcal{B}(X \times Y)$ is the collection of all Borel subsets of $X \times Y$, given the current state is (x, y) and actions $(a, b) \in A(x) \times B(x)$ are chosen by the players.

Also we have the discount factor $\beta \in (0, 1)$. In what follows, we assume that the transition kernel Q has a measurable density q with respect to some σ-finite measures λ and ν, i.e.,

$$Q(B|x, y, a, b) = \int_B q(x^{'}, y^{'}|x, y, a, b)\lambda(dx^{'})\nu(dy^{'}), \quad B \in \mathcal{B}(X \times Y). \tag{6.1.2}$$

We also introduce the marginal transition kernel density by

$$q^X(x^{'}|x, y, a, b) := \int_Y q(x^{'}, y^{'}|x, y, a, b)\nu(dy^{'}).$$

We assume that the distribution Q_0 of Y_0, the initial (unobservable) state is known to the players. The risk-sensitive partially observed zero sum game evolves in the following way:

- At the 0th decision epoch, based on the initial observation x_0, the players choose actions $a_0 \in A(x_0)$ and $b_0 \in B(x_0)$ simultaneously and independent of each other.

- As a consequence of these chosen actions player 1 gets a reward $C(x_0, y_0, a_0, b_0)$ and player 2 incurs a cost $C(x_0, y_0, a_0, b_0)$, where y_0 is the initial unobservable state. Note that the reward/cost depends on the unobservable component as well and therefore is itself unobservable.

- At the next time epoch the system moves to the next state (x_1, y_1) according to the transition law $Q(.|x_0, y_0, a_0, b_0)$. If the observable state at time 1 is x_1, then based on this observation, the previous observation x_0 and the previous pair of actions (a_0, b_0), players 1 and 2 choose actions $a_1 \in A(x_1)$ and $b_1 \in B(x_1)$ respectively. This yields a reward $C(x_1, y_1, a_1, b_1)$ for player 1 and a cost $C(x_1, y_1, a_1, b_1)$ for player 2. Now the sequence of events as described above repeats itself.

Let $\mathcal{H}_n$ be the admissible observable history available upto time n, i.e., $\mathcal{H}_0 = X$ and for $n \geq 1$, $\mathcal{H}_n = \mathcal{H}_{n-1} \times A \times B \times X$. Thus a typical element of $\mathcal{H}_n$ is given by $h_n = (x_0, a_0, b_0, x_1, a_1, b_1, \ldots, x_{n-1}, a_{n-1}, b_{n-1}, x_n)$, where for each n, x_n represents the observable state at the *nth* decision epoch, a_n is the action chosen by player 1 at the *nth* decision epoch and b_n is action chosen by player 2 at the *nth* decision epoch. We endow these spaces with the Borel sigma-algebra. Throughout whenever we talk about measurability we mean measurability with respect to the Borel sigma-algebra. Next we introduce the concept of decision rules and policies.

Definition 6.1.1. *Let $P(A)$ and $P(B)$ denote the set of all probability measures on A and B respectively.*

(a) A measurable mapping $f_n : \mathcal{H}_n \to P(A)$ with the property $f_n(h_n)(A(x_n)) = 1$ for $h_n \in \mathcal{H}_n$ is called a decision rule at stage n for player 1. Similarly, a measurable mapping $g_n : \mathcal{H}_n \to P(B)$ with the property $g_n(h_n)(B(x_n)) = 1$ for $h_n \in \mathcal{H}_n$ is called a decision rule at stage n for player 2.

(b) A sequence $\pi = (f_0, f_1, ...)$, and $\sigma = (g_0, g_1, ...)$, where f_n, g_n's are decision rules at stage n for all n, is called a pair of policies for player 1 and 2 respectively.

6.2 Finite Horizon Problem

For a fixed policy $\pi = (f_0, f_1, ...)$ and $\sigma = (g_0, g_1, ...)$, fixed (observable) initial state $x \in X$, the initial distribution Q_0 together with the transition kernel Q, we obtain by a theorem of Ionescu Tulcea a probability measure $\mathbb{P}_{xy}^{\pi\sigma}$ on $(X \times Y)^\infty$ endowed with the product σ-algebra. More precisely $\mathbb{P}_{xy}^{\pi\sigma}$ is the probability measure under policy (π, σ) given $X_0 = x$ and $Y_0 = y$. On this probability space, for $\omega = (x_0, y_0, x_1, y_n, \ldots, x_n, y_n, \ldots)$ we define the

random variables X_n and Y_n via the canonical projections:

$$X_n(\omega) = x_n, \quad Y_n(\omega) = y_n.$$

The action sequence for player 1 is given by $\{A_n\}$ and that of player 2 given by $\{B_n\}$. Under the policies, $\pi = (f_0, f_1, ...)$ and $\sigma = (g_0, g_1, ...)$, the distribution of A_n is given by $f_n(X_0, A_0, B_0, \ldots, X_n)$ and the distribution of B_n is given by $g_n(X_0, A_0, B_0, \ldots, X_n)$.

In this section we look into the N stage optimization problem. For defining the optimality criterion, consider a utility function $U : \mathbb{R}_+ \to \mathbb{R}$ which is assumed to be continuous and strictly increasing. The discounted cost/reward generated over N stages is given by:

$$C_N = \sum_{k=0}^{N-1} \beta^k C(X_k, Y_k, A_k, B_k). \tag{6.2.1}$$

For a fixed initial observable state x and given policies π and σ, the optimization criterion that we are interested in is given by:

$$J_N(x, \pi, \sigma) := \int_Y \mathbb{E}_{xy}^{\pi\sigma}[U(C_N)]Q_0(dy). \tag{6.2.2}$$

Here player 1 tries to maximize $J_N(x, \pi, \sigma)$ over all policies π, for each σ. Analogously, player 2 tries to minimize $J_N(x, \pi, \sigma)$ over all policies σ, for each π. This leads to the following definitions of optimal policies and value of the game.

Definition 6.2.1. *A strategy π^* is said to be optimal for player 1 in the partially observed model if*

$$J_N(x, \pi^*, \sigma) \geq \inf_{\sigma'} \sup_{\pi} J_N(x, \pi, \sigma'), \quad for\ any\ \ \sigma.$$

The quantity $\inf_{\sigma'} \sup_{\pi} J_N(x, \pi, \sigma')$ is referred to as the upper value of the partially observed game.
Similarly, a strategy σ^* is said to be optimal for player 2 in the partially observed model if

$$J_N(x,\pi,\sigma^*) \leq \sup_{\pi'} \inf_{\sigma} J_N(x,\pi',\sigma), \quad \textit{for any } \pi.$$

The quantity $\sup_{\pi'} \inf_{\sigma} J_N(x,\pi',\sigma)$ *is referred as the lower value of the partially observed game.*
Hence, a pair optimal strategies (π^*,σ^*) *satisfies*

$$J_N(x,\pi,\sigma^*) \leq J_N(x,\pi^*,\sigma^*) \leq J_N(x,\pi^*,\sigma)$$

for any π,σ*. Thus,* (π^*,σ^*) *constitutes a saddle point equilibrium. The partially observed game is said to have a value if*

$$J_N(x) = \sup_{\pi} \inf_{\sigma} J_N(x,\pi,\sigma) = \inf_{\sigma} \sup_{\pi} J_N(x,\pi,\sigma).$$

Note that if both the players have optimal strategies then the partially observed game has a value. Our aim is to show that the game model under consideration has a value and also there exists a saddle point equilibrium. Towards that end, we assume that the following assumptions are in force throughout the Chapter.

Assumption 6.2.2.

(i) For each $x \in X$*, the sets* $A(x)$ *and* $B(x)$ *are compact subsets of* A *and* B*. Also the mappings* $x \to A(x)$ *and* $x \to B(x)$ *are continuous.*

(ii) $(x,y,a,b) \to C(x,y,a,b)$ *is continuous.*

(iii) $(x,y,x',y',a,b) \to q(x',y'|x,y,a,b)$ *is bounded and continuous.*

(iv) C *is also bounded, i.e., there exists constants* $0 < \underline{c} \leq C(x,y,a,b) \leq \bar{c}$*.*

In literature partially observable risk-neutral games are treated by first converting it to an equivalent completely observable game. Following that approach, in this risk-sensitive approach also we first convert our partially observed game problem into a complete observed

game problem. We show the existence of the value function and optimal strategies in case of the equivalent completely observable model and then revert back to our partially observed model.

Now in the unobserved model, the state y and the cost accumulated so far cannot be observed because it depends on y. Thus we need to estimate them. For that we consider the following set of probability measures on $Y \times \mathbb{R}_+$:

$$P_b(Y \times \mathbb{R}_+) := \{\mu \text{ is a probability measure on the Borel } \sigma\text{- algebra } \mathcal{B}(Y \times \mathbb{R}_+) \text{ such that there exists a constant } K = K(\mu) > 0 \text{ with } \mu(Y \times [0, K]) = 1\}.$$

The elements of the above set will essay the role of the conditional distribution of the hidden state component and accumulated cost. The precise interpretation will be seen in Theorem 6.2.3. In order to estimate the unobserved state component and accumulated cost we define the following updating operator $\Phi : X \times A \times B \times X \times P_b(Y \times \mathbb{R}_+) \times \mathbb{R}_+ \to P_b(Y \times \mathbb{R}_+)$ given by

$$\Phi(x, a, b, x^{'}, \mu, z)(B) := \frac{\int_Y \int_{\mathbb{R}_+} \big(\int_B q(x^{'}, y^{'}|x, y, a, b)\nu(dy^{'})\delta_{s+zC(x,y,a,b)}(ds^{'})\big)\mu(dy, ds)}{\int_Y q^X(x^{'}|x, y, a, b)\mu^Y(dy)}, \tag{6.2.3}$$

where $B \in \mathcal{B}(Y \times \mathbb{R}_+)$ and $\mu^Y(dy) = \mu(dy, \mathbb{R}_+)$ is the Y-marginal distribution of μ. Going forward we will also use the notation $\mu^S(ds) := \mu(Y, ds)$, which will denote the S-marginal. We define the updating operator only when the denominator is positive. For $h_n = (x_0, a_0, b_0, ..., x_n)$ and $B \in \mathcal{B}(Y \times \mathbb{R}_+)$ we now define the sequence of probability measures

$$\begin{aligned} &\mu_0(C|h_0) := (Q_0 \times \delta_0)(C), \\ &\mu_{n+1}(C|h_n, a, b, x^{'}) = \Phi(x_n, a, b, x^{'}, \mu_n(.|h_n), \beta^n)(C). \end{aligned} \tag{6.2.4}$$

The next theorem provides the interpretation of the above defined sequence of probability measures (μ_n) as the sequence of conditional distributions. For that purpose we first define the sequence of random variables:

$$S_0 := 0, \;\; S_n := \sum_{k=0}^{n-1} \beta^k C(X_k, Y_k, A_k, B_k), \;\; n \in \mathbb{N}$$

We then have the following result which generalizes Theorem 1 of Bäuerle and Rieder [2017a] to the game setting.

Theorem 6.2.3. *Suppose that (μ_n) is given by the recursion (6.2.4). For $n \geq 0$, for a given initial observable state $x \in X$ and given policies $\pi = (f_0, f_1, \ldots)$ and $\sigma = (g_0, g_1, \ldots)$ of the respective players we have*

$$\mathbb{P}_x^{\pi\sigma}((Y_n, S_n) \in C | X_0, A_0, B_0, ..., X_n) = \mu_n(B | X_0, A_0, B_0, ..., X_n), \;\; \mathbb{P}_x^{\pi\sigma} - a.s., \;\; for \;\; C \in \mathcal{B}(Y \times \mathbb{R}_+),$$

where $\mathbb{P}_x^{\pi\sigma}(.) := \int \mathbb{P}_{xy}^{\pi\sigma}(.) Q_0(dy)$.

Proof. We first show that

$$\mathbb{E}_x^{\pi\sigma}[v(X_0, A_0, B_0, X_1, ..., X_n, Y_n, S_n)] = \mathbb{E}_x^{\pi\sigma}[v'(X_0, A_0, B_0, X_1, ..., X_n)] \tag{6.2.5}$$

for all bounded and measurable $v : \mathcal{H}_n \times Y \times \mathbb{R}_+ \to \mathbb{R}$ and

$$v'(h_n) := \int_Y \int_{\mathbb{R}_+} v(h_n, y_n, s_n) \mu_n(dy_n, ds_n | h_n).$$

We use induction on n. The basis step is true as for $n = 0$ both sides reduce to $\int v(x, y, 0) Q_0(dy)$. Now suppose that the statement is true for $n - 1$. we simply write f_n, g_n in place of $f_n(h_n), g_n(h_n)$. For a given observable history h_{n-1}, the left hand side of (6.2.5) becomes:

$$\mathbb{E}_x^{\pi\sigma}[v(h_{n-1}, A_{n-1}, B_{n-1}, X_n, Y_n, S_n)] = \int_B \int_A \int_Y \int_{\mathbb{R}_+} \mu_{n-1}(dy_{n-1}, ds_{n-1}|h_{n-1})\times$$
$$\int_Y \int_X \nu(dy_n)\lambda(dx_n)q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1})\times$$
$$\int_{\mathbb{R}_+} \delta_{s_{n-1}+\beta^{n-1}C(x_{n-1},y_{n-1},a_{n-1},b_{n-1})}(ds_n)v(h_{n-1}, a_{n-1}, b_{n-1}, x_n, y_n, s_n)f_{n-1}(da_{n-1})g_{n-1}(db_{n-1})$$
$$= \int_B \int_A \int_Y \int_{\mathbb{R}_+} \mu_{n-1}(dy_{n-1}, ds_{n-1}|h_{n-1}) \int_Y \int_X \nu(dy_n)\lambda(dx_n)q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1})\times$$
$$v(h_{n-1}, a_{n-1}, b_{n-1}, x_n, y_n, s_{n-1} + \beta^{n-1}C(x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1}))f_{n-1}(da_{n-1})g_{n-1}(db_{n-1}).$$

While the right hand side becomes:

$$\mathbb{E}_x^{\pi\sigma}[v^{'}(h_{n-1}, A_{n-1}, B_{n-1}, X_n)]$$
$$= \int_B \int_A \int_Y \int_{\mathbb{R}_+} \mu_{n-1}(dy_{n-1}, ds_{n-1}|h_{n-1})\times$$
$$\int_X \lambda(dx_n)q^X(x_n|x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1})v^{'}(h_{n-1}, a_{n-1}, b_{n-1}, x_n)f_{n-1}(da_{n-1})g_{n-1}(db_{n-1})$$
$$= \int_B \int_A \int_Y \mu_{n-1}^Y(dy_{n-1}|h_{n-1}) \times \int_X \lambda(dx_n)q^X(x_n|x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1})\times$$
$$\int_Y \int_{\mathbb{R}_+} \mu_n(dy_n, ds_n|h_n)v(h_{n-1}, a_{n-1}, b_{n-1}, x_n, y_n, s_n)f_{n-1}(da_{n-1})g_{n-1}(db_{n-1})$$
$$= \int_B \int_A \int_Y \int_X \nu(dy_n)\lambda(dx_n) \int_Y \int_{\mathbb{R}_+} \mu_{n-1}(dy_{n-1}, ds_{n-1}|h_{n-1})q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1}, g_{n-1})\times$$
$$\int_{\mathbb{R}_+} \delta_{s_{n-1}+\beta^{n-1}C(x_{n-1},y_{n-1},f_{n-1},g_{n-1})}(ds_n)v(h_{n-1}, f_{n-1}, g_{n-1}, x_n, y_n, s_n)f_{n-1}(da_{n-1})g_{n-1}(db_{n-1})$$
$$= \int_B \int_A \int_Y \int_{\mathbb{R}_+} \mu_{n-1}(dy_{n-1}, ds_{n-1}|h_{n-1}) \int_Y \int_X \nu(dy_n)\lambda(dx_n)q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1})\times$$
$$v(h_{n-1}, a_{n-1}, b_{n-1}, x_n, y_n, s_{n-1} + \beta^{n-1}C(x_{n-1}, y_{n-1}, a_{n-1}, b_{n-1}))f_{n-1}(da_{n-1})g_{n-1}(db_{n-1}),$$

where we use the recursion for μ_n in the third equation and use Fubini's theorem, to cancel out the normalizing constant of μ_n. Hence we are done by induction.
Now, in particular, if we take $v = 1_{C\times D}$ with $C \in \mathcal{B}(Y \times \mathbb{R}_+)$ and $D \subset X \times A \times B \times ... \times X$

a measurable subset of histories until time n then we get from (6.2.5),

$$\mathbb{P}_x^{\pi\sigma}((Y_n, S_n) \in C, (X_0, A_0, B_0, ..., X_n) \in D) = \mathbb{E}_x^{\pi\sigma}[\mu_n(C|X_0, A_0, B_0, ..., X_n)1_D(X_0, A_0, B_0, ..., X_n)]$$

This establishes the fact that $\mu_n(\cdot|X_0, A_0, B_0, ..., X_n)$ is a conditional $\mathbb{P}_x^{\pi\sigma}$-distribution of (Y_n, S_n) given the history $(X_0, A_0, B_0, ..., X_n)$. □

Now we again look at the optimization problem (6.2.2). Motivated by the previous result we define for $x \in X$, $\mu \in P_b(Y \times \mathbb{R}_+)$, $z \in (0, 1]$ and $n = 1, 2, ..., N$:

$$V_{n\pi\sigma}(x, \mu, z) := \int_Y \int_{\mathbb{R}_+} \mathbb{E}_{xy}^{\pi\sigma}\Big[U\big(s + z\sum_{k=0}^{n-1} \beta^k C(X_k, Y_k, A_k, B_k)\big)\Big]\mu(dy, ds). \tag{6.2.6}$$

We solve our optimization problem by using a state augmentation technique. For that purpose we define, for a probability measure $\mu \in P(Y)$:

$$Q^X(C|x, \mu, a, b) := \int_B \int_Y q^X(x^{'}|x, y, a, b)\mu(dy)\lambda(dx^{'}), \quad C \in \mathcal{B}(X).$$

We now consider the completely observable model with new state space $E = X \times P_b(Y \times \mathbb{R}_+) \times (0, 1]$. The action spaces for player 1 and player 2 are same as the partially observable model. One stage cost/reward is 0 and the terminal cost/reward function is $V_0(x, \mu, z) := \int_Y \int_{\mathbb{R}_+} U(s)\mu(dy, ds)$. Since for all $\mu \in P_b(Y \times \mathbb{R}_+)$ the support of μ in the s-component is bounded, the expectation is well defined. The transition law for the new model is given by $\tilde{Q}(.|x, \mu, z, a, b)$, which for $(x, \mu, z, a, b) \in E \times A \times B$, and a Borel measurable subset $C \subset E$ is defined by

$$\tilde{Q}(C|x, \mu, z, a, b) := \int_X 1_B\big((x^{'}, \Phi(x, a, b, x^{'}, \mu, z), \beta z)\big)Q^X(dx^{'}|x, \mu^Y, a, b).$$

The decision rules for player 1 in the newly defined model are given by measurable mappings $f : E \to P(A)$ such that $f(x, \mu, z)(A(x)) = 1$. Similarly, the decision rules for player 2 are given by measurable mappings $g : E \to P(B)$ such that $g(x, \mu, z)(B(x)) = 1$. We denote by F_1 the set of all decision rules for player 1 and F_2 denotes the same for player 2. For player

1, we denote by Π_1^M the set of all Markov policies $\pi = (f_0, f_1, ...)$ with $f_n \in F_1$ for all $n \geq 0$. Π_2^M represents the same for player 2.

Let $\mathcal{C}(E) = \{v : E \to \mathbb{R}, v \text{ is continuous and } v \geq V_0\}$. Note that we consider the topology of weak convergence on $P_b(Y \times \mathbb{R}_+)$. For $v \in \mathcal{C}(E)$, $(\zeta, \eta) \in P(A(x)) \times P(B(x))$ and $(f, g) \in F_1 \times F_2$, we consider the following operators:

$$(T_{fg}v)(x,\mu,z) := \int_B \int_A \int_X v\big(x', \Phi(x,a,b,x',\mu,z), \beta z\big) Q^X(dx'|x,\mu^Y,a,b) f(x,\mu,z)(da) g(x,\mu,z)(db).$$

$$(Lv)(x,\mu,z,\zeta,\eta) := \int_B \int_A \int_X v\big(x', \Phi(x,a,b,x',\mu,z), \beta z\big) Q^X(dx'|x,\mu^Y,a,b) \zeta(da)\eta(db).$$

$$Tv(x,\mu,z) := \inf_g \sup_f T_{fg}v(x,\mu,z) = \inf_\zeta \sup_\eta (Lv)(x,\mu,z,\zeta,\eta), \quad (x,\mu,z) \in E. \tag{6.2.7}$$

Next we have the following theorem:

Theorem 6.2.4. *(a) Let $\pi = (f_0, f_1, ..., f_{N-1})$ and $\sigma = (g_0, g_1, ..., g_{N-1})$ be two policies for player 1 and 2 respectively. Then it holds that for all $n = 1, 2, ..., N$,*

$$V_{n\pi\sigma}(x,\mu,z) = T_{f_0 g_0} T_{f_1 g_1} ... T_{f_{n-1} g_{n-1}} V_0.$$

(b) For all $n = 1, 2, ..., N$ let $V_n = \inf_\pi \sup_\sigma V_{n\pi\sigma}$. Then $V_n \in \mathcal{C}(E)$ and

$$V_n(x,\mu,z) = TV_{n-1}(x,\mu,z).$$

(c) For $n = 1, 2, ..., N$ there exists measurable functions $(\gamma_n^*, \delta_n^*) \in F_1 \times F_2$ such that

$$LV_{n-1}(x,\mu,z,\zeta,\delta_n^*(x,\mu,z)) \leq LV_{n-1}(x,\mu,z,\gamma_n^*(x,\mu,z),\delta_n^*(x,\mu,z)) \leq LV_{n-1}(x,\mu,z,\gamma_n^*(x,\mu,z),\eta),$$

for all $(\zeta,\eta)\in P(A(x))\times P(B(x))$ *and* $(x,\mu,z)\in E$. *Then* $V_N(\cdot,Q_0\times\delta_0,1)$ *is the value of the* N *stage partially observable stochastic game and* $(\pi^*,\sigma^*)=(f_n^*,g_n^*)_{n=0,1,...,N-1}$ *with* $f_n^*(h_n)=\gamma_{N-n}^*(x_n,\mu(\cdot|h_n),\beta^n)$ *and* $g_n^*=\delta_{N-n}^*(x_n,\mu(\cdot|h_n),\beta^n)$ *are optimal policies for player 1 and 2 respectively.*

Proof. (a) We establish the above iteration by induction on n. For $n=1$ we have,

$$T_{f_0,g_0}V_0(x,\mu,z)=\int_B\int_A\int_X V_0\big(x^{'},\Phi(x,a,b,x^{'},\mu,z),\beta z\big)Q^X(dx^{'}|x,\mu^Y,a,b)f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_B\int_A\int_Y\int_{\mathbb{R}_+}\int_X\int_{\mathbb{R}_+}U(s^{'})\delta_{s+zC(x,y,a,b)}(ds^{'})q^X(x^{'}|x,y,a,b)\lambda(dx^{'})\mu(dy,ds)f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_Y\int_{\mathbb{R}_+}\mathbb{E}_{xy}^{\pi\sigma}\big[U\big(s+zC(x,y,A_0,B_0)\big)\big]\mu(dy,ds)$$
$$=V_{1\pi\sigma}(x,\mu,z).$$

Now suppose that the statement is true for V_n. Let $\bar\pi=(f_1,f_2,...)$ and $\bar\sigma=(g_1,g_2,...)$ denote the 1-shifted policies. Then we get,

$$(T_{f_0g_0}T_{f_1g_1}...T_{f_ng_n}V_0)(x,\mu,z)$$
$$=\int_B\int_A\int_X V_{n\bar\pi\bar\sigma}\big(x^{'},\Phi(x,a,b,x^{'},\mu,z),\beta z\big)Q^X(dx^{'}|x,\mu^Y,a,b)f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_B\int_A\int_X\int_Y\int_{\mathbb{R}_+}\mathbb{E}_{x^{'},y^{'}}^{\bar\pi\bar\sigma}\big[U(s^{'}+z\sum_{k=0}^{n-1}\beta^{k+1}C(X_k,Y_k,A_k,B_k))\big]$$
$$\Phi(x,a,b,x^{'},\mu,z)(dy^{'},ds^{'})Q^X(dx^{'}|x,\mu^Y,a,b)f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_B\int_A\int_X\int_Y\int_{\mathbb{R}_+}\mathbb{E}^{\pi,\sigma}\big[U(s^{'}+z\sum_{k=1}^{n}\beta^kC(X_k,Y_k,A_k,B_k))|X_1=x^{'},Y_1=y^{'}\big]$$
$$\cdot\int_Y\int_{\mathbb{R}_+}q(x^{'},y^{'}|x,y,a,b)\delta_{s+zC(x,y,a,b)}(ds^{'})\mu(dy,ds)\nu(dy^{'})\lambda(dx^{'})f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_B\int_A\int_Y\int_Y\int_X\int_{\mathbb{R}_+}\mathbb{E}^{\pi,\sigma}\big[U(s+zC(x,y,a,b)+z\sum_{k=1}^{n}\beta^kC(X_k,Y_k,A_k,B_k))|X_1=x^{'},Y_1=y^{'}\big]$$
$$q(x^{'},y^{'}|x,y,a,b)\mu(dy,ds)\nu(dy^{'})\lambda(dx^{'})f_0(x,\mu,z)(da)g_0(x,\mu,z)(db)$$
$$=\int_Y\int_{\mathbb{R}_+}\mathbb{E}^{\pi,\sigma}\big[U(s+z\sum_{k=0}^{n}\beta^kC(X_k,Y_k,A_k,B_k))\big]\mu(dy,ds)$$
$$=V_{n+1\pi\sigma}(x,\mu,z);$$

Hence we have the desired conclusion by induction.

(b and c) We show by induction on n that
(i) $V_n = TV_{n-1} \in \mathcal{C}(E)$
(ii) $T_{\gamma_n\delta_n^*}T_{\gamma_{n-1}\delta_{n-1}^*}...T_{\gamma_1\delta_1^*}V_0 \leq V_n$, for any measurable $\gamma_1, \gamma_2, ..., \gamma_n : E \to P(A(x))$.
(iii) $T_{\gamma_n^*\delta_n}T_{\gamma_{n-1}^*\delta_{n-1}}...T_{\gamma_1^*\delta_1}V_0 \geq V_n$, for any measurable $\delta_1, \delta_2, ..., \delta_n : E \to P(B(x))$.
(iv) $T_{\gamma_n^*\delta_n^*}T_{\gamma_{n-1}^*\delta_{n-1}^*}...T_{\gamma_1^*\delta_1^*}V_0 = V_n$.

Under our assumptions $V_0 \in \mathcal{C}(E)$. For $n = 1$, it follows from definition that

$$V_1 = \inf_g \sup_f T_{fg}V_0 = TV_0.$$

Under our assumptions the existence of (γ_1^*, δ_1^*) follows from a classical measurable selection theorem Brown and Purves [1973] and Fan's minimax theorem Fan [1953]. Now $(ii) - (iv)$ follows from the definition of (γ_1^*, δ_1^*). Now suppose the statement is true for $n-1$. Since $V_{n-1} \in \mathcal{C}(E)$, we again have the existence of (γ_n^*, δ_n^*). Using the induction hypothesis and monotonicity of the T we obtain

$$T_{\gamma_n\delta_n^*}T_{\gamma_{n-1}\delta_{n-1}^*}...T_{\gamma_1\delta_1^*}V_0 \leq T_{\gamma_n\delta_n^*}V_{n-1} \leq T_{\gamma_n^*\delta_n^*}V_{n-1} = T_{\gamma_n^*\delta_n^*}T_{\gamma_{n-1}^*\delta_{n-1}^*}...T_{\gamma_1^*\delta_1^*}V_0$$

for any measurable $\gamma_1, \gamma_2, ..., \gamma_n : E \to P(A(x))$. On the other hand

$$T_{\gamma_n^*\delta_n^*}V_{n-1} \leq T_{\gamma_n^*\delta_n}V_{n-1} \leq T_{\gamma_n^*\delta_n}T_{\gamma_{n-1}^*\delta_{n-1}}...T_{\gamma_1^*\delta_1}V_0$$

for any measurable $\delta_1, \delta_2, ..., \delta_n : E \to P(B(x))$. Thus combining with part (a), we have shown that $((\gamma_N^*, \gamma_{N-1}^*, \ldots, \gamma_1^*), (\delta_N^*, \delta_{N-1}^*, \ldots, \delta_1^*))$ is a saddle point equilibrium for the N stage completely observable game problem with optimization criterion given by (6.2.6). The rest of the conclusions now follow from the relation between the partially observable game and completely observable game. □

6.3 Infinite Horizon Problem

In case of finite horizon problem we have the existence of the value of the game and also we have shown the existence of the optimal strategies. Now we consider the case of infinite horizon, i.e for a given pair of policies (π, σ) and $x \in X$ we are interested in the optimization problem:

$$J_\infty(x, \pi, \sigma) := \int_Y \mathbb{E}_{xy}^{\pi\sigma}\big[U\big(\sum_{k=0}^{\infty} \beta^k C(X_k, Y_k, A_k, B_k)\big)\big] Q_0(dy), \quad x \in E.$$

The upper value, lower value, optimal strategies and value of the game have the same definitions as in Definition 6.2.1 with N replaced by ∞. For the infinite horizon problem we will assume that the utility function U is either concave or convex. We need to deal with concave and convex utility functions separately.

Concave utility function: We first investigate the case of concave utility function. Just as in the finite horizon case we will consider the equivalent completely observable game. To that end we have the following notations.

$$V_{\infty\pi\sigma}(x, \mu, z) := \int_Y \int_{\mathbb{R}_+} \mathbb{E}_{xy}^{\pi\sigma}\big[U\big(s + z\sum_{k=0}^{\infty} \beta^k C(X_k, Y_k, A_k, B_k)\big)\big] \mu(dy, ds),$$

$$V_\infty(x, \mu, z) := \inf_\sigma \sup_\pi V_{\infty\pi\sigma}(x, \mu, z). \tag{6.3.1}$$

We denote

$$\bar{a}(\mu, z) := \int_{\mathbb{R}_+} U(s + \frac{z\bar{c}}{1-\beta}) \mu^S(ds),$$

$$\underline{a}(\mu, z) := \int_{\mathbb{R}_+} U(s + \frac{z\underline{c}}{1-\beta}) \mu^S(ds), \quad (\mu, z) \in P_b(Y \times \mathbb{R}_+) \times (0, 1],$$

where $\underline{c}$ and $\bar{c}$ are as in Assumption 6.2.2. Then we have the main theorem of this section:

Theorem 6.3.1. *(a) V_∞ is the unique solution of $v = Tv$ in $\mathcal{C}(E)$ with $\underline{a}(\mu, z) \le v(x, \mu, z) \le \bar{a}(\mu, z)$ where T is as defined in (6.2.7). Moreover, $T^n V_0 \uparrow V_\infty$, $T^n \underline{a} \uparrow V_\infty$ and $T^n \bar{a} \downarrow V_\infty$ as $n \to \infty$.*

(b) There exist measurable functions $(\gamma^*, \delta^*) \in F_1 \times F_2$ *such that*

$$T_{\gamma\delta^*}V_\infty(x,\mu,z) \le T_{\gamma^*\delta^*}V_\infty(x,\mu,z) \le T_{\gamma^*\delta}V_\infty(x,\mu,z), \tag{6.3.2}$$

for all $(\gamma,\delta) \in F_1 \times F_2$ *and* $(x,\mu,z) \in E$. *Then* $V_\infty(x, Q_0 \times \delta_0, 1)$ *is the value of the partially observable infinite horizon stochastic game. Moreover,* $(\pi^*,\sigma^*) = ((f_0^*, f_1^*, \ldots), (g_0^*, g_1^*, \ldots))$ *with* $f_n^*(h_n) := \gamma^*(x_n, \mu_n(\cdot|h_n), \beta^n)$ *and* $g_n^*(h_n) := \delta^*(x_n, \mu_n(\cdot|h_n), \beta^n)$ *are optimal policies for player 1 and 2 respectively.*

Proof. (a) Here we first show that $V_n = T^nV_0 \uparrow V_\infty$. Since $U : \mathbb{R}_+ \to \mathbb{R}$ is increasing and concave, it satisfies the inequality

$$U(s_1+s_2) \le U(s_1) + U_-'(s_1)s_2, \quad s_1, s_2 \ge 0,$$

where U_-' is the left hand side derivative of U that exists since U is concave. Further more, $U_-'(s) \ge 0$ and U_-' is non increasing. For $(x,\mu,z) \in E$ it holds, $V_n(x,\mu,z) \le V_\infty(x,\mu,z)$. Using this, we get

$$\begin{aligned}
&V_{n\pi\sigma}(x,\mu,z) \le V_{\infty\pi\sigma}(x,\mu,z)\\
&= \int_Y \int_{\mathbb{R}_+} \mathbb{E}_{xy}^{\pi\sigma}\big[U\big(s + z\sum_{k=0}^{\infty}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds),\\
&\le \int_Y \int_{\mathbb{R}_+} \mathbb{E}_{xy}^{\pi\sigma}\big[U\big(s + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds),\\
&+ \int_Y \int_{\mathbb{R}_+} \mathbb{E}_{xy}^{\pi\sigma}\big[U_-'\big(s + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big) z\sum_{m=n}^{\infty}\beta^k C(X_m,Y_m,A_m,B_m)\big]\mu(dy,ds),\\
&\le V_{n\pi\sigma}(x,\mu,z) + \beta^n \frac{z\bar{c}}{1-\beta}\int_{\mathbb{R}_+} U_-'(s+z\underline{c})\mu^S(ds)\\
&\le V_{n\pi\sigma}(x,\mu,z) + \beta^n \frac{z\bar{c}}{1-\beta}U_-'(z\underline{c}) =: V_{n\pi\sigma}(x,\mu,z) + \epsilon_n(z),
\end{aligned} \tag{6.3.3}$$

where $\epsilon_n(z)$ is defined by the last equation. Clearly, $\lim_{n\to\infty}\epsilon_n(z) = 0$. Now from (6.3.3) we get $V_n(x,\mu,z) \le V_\infty(x,\mu,z) \le V_n(x,\mu,z) + \epsilon(z)$. Taking limit we get $V_n = T^nV_0 \uparrow V_\infty$. Now,

$V_{n+1} = TV_n \leq TV_\infty$. Taking limit $n \to \infty$ we get $V_\infty \leq TV_\infty$. Again, $V_\infty \leq V_n + \epsilon_n$. Now applying operator T on both sides we have $TV_\infty \leq T(V_n + \epsilon_n) = V_{n+1} + \epsilon_{n+1}$. Thus again letting $n \to \infty$ we obtain $TV_\infty \leq V_\infty$. Hence combining two inequalities we have $V_\infty = TV_\infty$. Next, we obtain

$$\begin{aligned}(T\bar{a})(x,\mu,z) &= \inf_{\eta}\sup_{\zeta}\int_B\int_A\int_X\int_{\mathbb{R}_+} U(s^{'} + \frac{z\beta\bar{c}}{1-\beta})\Phi^s(x,a,b,x^{'},\mu,z)(ds^{'})Q^X(dx^{'}|x,\mu^Y,a,b)\zeta(da)\eta(db)\\ &\leq \int_{\mathbb{R}_+} U(s + z\bar{c} + \frac{z\beta\bar{c}}{1-\beta})\mu^S(ds)\\ &= \int_{\mathbb{R}_+} U(s + \frac{z\bar{c}}{1-\beta})\mu^S(ds) = \bar{a}(\mu,z).\end{aligned}$$

By the similar reasoning it can be shown that $T\underline{a} \geq \underline{a}$. Thus we have $T^n\bar{a} \downarrow$ and $T^n\underline{a} \uparrow$ and the limits exist. Moreover, by iteration we have,

$$\begin{aligned}(T^n\underline{a})(x,\mu,z) &= \inf_{\sigma}\sup_{\pi}\int_Y\int_{\mathbb{R}_+}\mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + \frac{z\beta^n\underline{c}}{1-\beta} + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds)\\ &\geq (T^nV_0)(x,\mu,z).\\ (T^n\bar{a})(x,\mu,z) &= \inf_{\sigma}\sup_{\pi}\int_Y\int_{\mathbb{R}_+}\mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + \frac{z\beta^n\bar{c}}{1-\beta} + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds)\end{aligned}$$

Using the fact $U(s_1 + s_2) \leq U(s_1) + U^{'}_{-}(s_1)s_2$, we obtain

$$\begin{aligned}0 &\leq (T^n\bar{a})(x,\mu,z) - (T^n\underline{a})(x,\mu,z) \leq (T^n\bar{a})(x,\mu,z) - (T^nV_0)(x,\mu,z)\\ &\leq \sup_{\sigma}\sup_{\pi}\int_Y\int_{\mathbb{R}_+}\mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + \frac{z\beta^n\bar{c}}{1-\beta} + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\\ &- U\big(s + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds)\\ &\leq \epsilon_n(z).\end{aligned}$$

Since $\lim_{n\to\infty}\epsilon_n = 0$, we obtain $T^n\underline{a} \uparrow V_\infty$ and $T^n\bar{a} \downarrow V_\infty$ as $n \to \infty$. Also since V_∞ is both increasing as well as decreasing limit of continuous functions, it is also continuous.

Now for the uniqueness purpose if possible let there be another solution $v \in \mathcal{C}(E)$ of $v = Tv$ with $\underline{a} \leq v \leq \bar{a}$. This then implies that $T^n\underline{a} \leq v \leq T^n\bar{a}$ for all n. Taking the limit $n \to \infty$ we have the uniqueness of the solution.
(b) The existence of (γ^*, δ^*) follows again from the measurable selection theorem and the minimax theorem. By monotonicity and the fact that $V_0 \leq V_\infty \leq \bar{a}$ we obtain that $\lim_{n\to\infty} T^n_{\gamma^*,\delta^*}V_0 = \lim_{n\to\infty} T^n_{\gamma^*,\delta^*}V_\infty = V_{\infty\pi_{\gamma^*}\sigma_{\delta^*}}$, where $\pi_{\gamma^*} = (\gamma^*, \gamma^*, \ldots)$ and $\sigma_{\delta^*} = (\delta^*, \delta^*, \ldots)$. By the definition of (γ^*, δ^*) we obtain for any $(\gamma, \delta) \in F_1 \times F_2$,

$$T_{\gamma\delta^*}V_\infty \leq T_{\gamma^*\delta^*}V_\infty \leq T_{\gamma^*\delta}V_\infty.$$

The property of (γ^*, δ^*) also implies that $V_\infty = TV_\infty = T_{\gamma^*\delta^*}V_\infty$. Hence we can also write for any $(\gamma, \delta) \in F_1 \times F_2$,

$$T_{\gamma\delta^*}V_\infty \leq V_\infty \leq T_{\gamma^*\delta}V_\infty.$$

By iterating this inequality n-times we get

$$T_{\gamma_1\delta^*}T_{\gamma_2\delta^*}...T_{\gamma_n\delta^*}V_\infty \leq V_\infty \leq T_{\gamma^*\delta_1}T_{\gamma^*\delta_2}...T_{\gamma^*\delta_n}V_\infty$$

for arbitrary $\gamma_1, \gamma_2, ..., \gamma_n$ and $\delta_1, \delta_2, ..., \delta_n$. Letting $n \to \infty$ we get,

$$V_{\infty\pi\sigma_{\delta^*}} \leq V_{\infty\pi_{\gamma^*}\sigma_{\delta^*}} = V_\infty \leq V_{\infty\pi_{\gamma^*}\sigma}$$

for all policies π and σ. The rest of the conclusions are now straight forward.

□

Convex Utility function: Now we consider the case of convex utility function U.

Theorem 6.3.2. *Theorem 6.3.1 also holds for convex U.*

Proof. For the convex case the proof is along the same lines as in Theorem 6.3.1. The only difference is that here we will use another inequality. Note that for $U : \mathbb{R}_+ \to \mathbb{R}$ increasing and convex we have the inequality

$$U(s_1 + s_2) \leq U(s_1) + U'_+(s_1 + s_2)s_2, \quad s_1, s_2 \geq 0,$$

where U'_+ is the right-hand side derivative of U that exists since U is convex. Moreover,

$U'_+(s) \geq 0$ and U'_+ is increasing. Thus, we obtain for $(x, \mu, z) \in E$,

$$
\begin{aligned}
&V_{n\pi\sigma}(x,\mu,z) \leq V_{\infty\pi\sigma}(x,\mu,z) \\
&= \int_Y \int_{\mathbb{R}_+} \mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + z\sum_{k=0}^{\infty}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds), \\
&\leq \int_Y \int_{\mathbb{R}_+} \mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds), \\
&+ \int_Y \int_{\mathbb{R}_+} \mathbb{E}^{\pi\sigma}_{xy}\big[U'_+\big(s + z\sum_{k=0}^{\infty}\beta^k C(X_k,Y_k,A_k,B_k)\big) z\sum_{m=n}^{\infty}\beta^k C(X_m,Y_m,A_m,B_m)\big]\mu(dy,ds), \\
&\leq V_{n\pi\sigma}(x,\mu,z) + \beta^n \frac{z\bar{c}}{1-\beta}\int_{\mathbb{R}_+} U'_+(s+\frac{z\bar{c}}{1-\beta})\mu^S(ds)
\end{aligned}
$$

Now let's denote $\delta_n(\mu,z) := \beta^n \frac{z\bar{c}}{1-\beta}\int_{\mathbb{R}_+} U'_+(s+\frac{z\bar{c}}{1-\beta})\mu^S(ds)$. Then we have $\lim_{n\to\infty}\delta_n(\mu,z) = 0$. So we end up with

$$V_n(x,\mu,z) \leq V_\infty(x,\mu,z) \leq V_n(x,\mu,z) + \delta_n(\mu,z).$$

Letting $n \to \infty$ yields $T^n V_0 \to V_\infty$.
We use the same inequality to get,

$$
\begin{aligned}
&0 \leq (T^n\bar{a})(x,\mu,z) - (T^n\underline{a})(x,\mu,z) \leq (T^n\bar{a})(x,\mu,z) - (T^n V_0)(x,\mu,z) \\
&\leq \sup_\sigma \sup_\pi \int_Y \int_{\mathbb{R}_+} \mathbb{E}^{\pi\sigma}_{xy}\big[U\big(s + \frac{z\beta^n\bar{c}}{1-\beta} + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big) \\
&- U\big(s + z\sum_{k=0}^{n-1}\beta^k C(X_k,Y_k,A_k,B_k)\big)\big]\mu(dy,ds) \\
&\leq \beta^n\frac{z\bar{c}}{1-\beta}\int_{\mathbb{R}_+} U'_+(s+\frac{z\bar{c}}{1-\beta})\mu^S(ds) = \delta_n(\mu,z)
\end{aligned}
$$

and the right hand side converges to 0 as $n \to \infty$. □

Few remarks are in order.

Remark 6.3.3. *(1) An important sub case of the model that we have considered here is when the reward/cost does not depend on the unobservable component, i.e., $C(x,y,a,b) = C'(x,a,b)$ for some function $C'(\cdot)$. In this case the accumulated reward/cost is no more unobservable and thereby we need not estimate it. Thus in that case it can be shown along similar lines as in Proposition 1 of Bäuerle and Rieder [2017a], that we can take the state space of the completely observable model as $X \times P(Y) \times \mathbb{R}_+ \times (0,1]$ and the updating operator as*

$$\Phi(x,a,b,x^{'},\mu,z)(B) := \frac{\int_Y \left(\int_B q(x^{'},y^{'}|x,y,a,b)\nu(dy^{'})\right)\mu(dy)}{\int_Y q^X(x'|x,y,a,b)\mu(dy)},$$

where B is a Borel subset of Y and $\mu \in P(Y)$.

(2) An important utility function is given by the function $U(x) = \frac{1}{\theta}e^{\theta x}$, where $\theta > 0$ is a fixed parameter. In this case again it is not necessary to keep track of the accumulated cost. It is enough to consider $X \times P(Y) \times (0,1]$ as the state space of the completely observable model. Again arguing similar to Theorem 3 in Bäuerle and Rieder [2017a], it can be shown that the value of the N stage game problem is given by $J(x) = \alpha_N(x, Q_0, \theta)$ where $\alpha_0(x,\mu,\theta z) = \frac{1}{\theta}$ and for $n = 1,2,\ldots,N$,

$$\alpha_{n+1}(x,\mu,\theta z) = \inf_{\eta \in P(B(x))} \sup_{\zeta \in P(A(x))} \int_B \int_A \left[\alpha_n(x', \Phi_e(x,a,b,x',\mu,z), \beta\theta z)\hat{Q}^X(dx'|x,\mu,a,b,\theta z)\right] \zeta(da)\eta(db),$$

where $(x,\mu,z) \in X \times P(Y) \times (0,1]$ and for B_1, Borel subset of X and B_2, Borel subset of Y,

$$\hat{Q}^X(B_1|x,\mu,a,b,z) = \int_{B_1}\int_Y e^{zC(x,y,a,b)} q^X(x'|x,y,a,b)\mu(dy)\lambda(dx'),$$

$$\Phi_e(x,a,b,x',\mu,z)(B_2) = \frac{\int_{B_2}\int_Y e^{zC(x,y,a,b)} q(x',y'|x,y,a,b)\mu(dy)\nu(dy')}{\int_Y\int_Y e^{zC(x,y,a,b)} q(x',y'|x,y,a,b)\mu(dy)\nu(dy')}.$$

CHAPTER 7

Future Directions

We now conclude the thesis by enumerating a few problems for future investigation, that arises naturally out of this thesis. Here, we have considered only zero-sum stochastic games under probability criterion. While in the discrete time setup, both zero and non-zero sum games with probability criterion have been studied. In Huang and Guo [2020] the authors study non-zero sum game problem for discrete time Markov chain under probability criterion. They assumed that the state space is countable and also the one stage cost function assumes only rational values, because of this the extended state space remains countable. Analysis of non-zero sum games with countable state space is substantially simpler as compared to the general state space case. In the continuous time setup that we consider, even if we assume that the cost rate assumes only rational values, since the cost accumulated from one jump to the other will get multiplied by the holding time, the state space will no longer remain countable. This is the reason why analysis of non-zero sum games with probability criterion, in the continuous time setup becomes a very challenging problem. Thus it will be interesting to investigate even the discrete time setup without the assumption that the one stage cost function assumes only rational values. Once that is possible the obvious challenge will be then to deal with the continuous time problems.

In the risk-sensitive setup we consider both zero and non-zero sum games for semi-Markov processes. Our analysis is contingent on two crucial assumptions, namely, finiteness of state space and holding time distributions having fixed finite support. The one controller counterpart studied in Chávez-Rodríguez et al. [2016] also has similar assumptions. Such assumptions can be restrictive from applications point of view. Thus, it will be worth

studying both single and multi controller problems without these assumptions.

In the partially observable setup we consider zero-sum games with finite and infinite horizon discounted cost. Thus future problems include studying non-zero sum games. Again here, even if we start with a countable state space for the original problem, the equivalent completely observable problem will have a general state space. Thus the analysis of such a problem will be quite involved. Also it will be interesting to look at average cost problems.

Bibliography

Arapostathis, A. and Biswas, A. (2018). Infinite horizon risk-sensitive control of diffusions without any blanket stability assumptions. *Stochastic Process. Appl.*, 128(5):1485–1524.

Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.*, 31(2):282–344.

Basu, A. and Ghosh, M. K. (2012). Zero-sum risk-sensitive stochastic differential games. *Math. Oper. Res.*, 37(3):437–449.

Basu, A. and Ghosh, M. K. (2014). Zero-sum risk-sensitive stochastic games on a countable state space. *Stochastic Process. Appl.*, 124(1):961–983.

Basu, A. and Ghosh, M. K. (2018). Nonzero-sum risk-sensitive stochastic games on a countable state space. *Math. Oper. Res.*, 43(2):516–532.

Bäuerle, N. and Rieder, U. (2011). *Markov decision processes with applications to finance*. Universitext. Springer, Heidelberg.

Bäuerle, N. and Rieder, U. (2014). More risk-sensitive Markov decision processes. *Math. Oper. Res.*, 39(1):105–120.

Bäuerle, N. and Rieder, U. (2017a). Partially observable risk-sensitive Markov decision processes. *Math. Oper. Res.*, 42(4):1180–1196.

Bäuerle, N. and Rieder, U. (2017b). Zero-sum risk-sensitive stochastic games. *Stochastic Process. Appl.*, 127(2):622–642.

Bhabak, A., Pal, C., and Saha, S. (2022). Zero-sum semi-Markov games with a probability criterion. *Stochastics*, 94(3):415–431.

Bhabak, A. and Saha, S. (2021). Continuous-time zero-sum games with probability criterion. *Stoch. Anal. Appl.*, 39(6):1130–1143.

Bhabak, A. and Saha, S. (2022a). Partially observable discrete-time discounted markov games with general utility. *arXiv:2211.07888*.

Bhabak, A. and Saha, S. (2022b). Risk-sensitive semi-markov decision problems with discounted cost and general utilities. *Statist. Probab. Lett.*, 184:Paper No. 109408, 9.

Bhabak, A. and Saha, S. (2023). Zero and non-zero sum risk-sensitive semi-markov games. *Stochastic Analysis and Applications*, 41(1):134–151.

Bhattacharya, R. N. and Majumdar, M. (1989). Controlled semi-Markov models—the discounted case. *J. Statist. Plann. Inference*, 21(2):223–242.

Biswas, A., Borkar, V. S., and Suresh Kumar, K. (2010). Risk-sensitive control with near monotone cost. *Appl. Math. Optim.*, 62(2):145–163.

Biswas, A. and Pradhan, S. (2022). Ergodic risk-sensitive control of Markov processes on countable state space revisited. *ESAIM Control Optim. Calc. Var.*, 28:Paper No. 26, 50.

Biswas, A. and Saha, S. (2020). Zero-sum stochastic differential games with risk-sensitive cost. *Appl. Math. Optim.*, 81(1):113–140.

Borkar, V. S. and Meyn, S. P. (2002). Risk-sensitive optimal control for Markov decision processes with monotone cost. *Math. Oper. Res.*, 27(1):192–209.

Bouakiz, M. and Kebir, Y. (1995). Target-level criterion in Markov decision processes. *J. Optim. Theory Appl.*, 86(1):1–15.

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Ann. Statist.*, 1:902–912.

Cavazos-Cadena, R. and Hernández-Hernández, D. (2019). The vanishing discount approach in a class of zero-sum finite games with risk-sensitive average criterion. *SIAM J. Control Optim.*, 57(1):219–240.

Chávez-Rodríguez, S., Cavazos-Cadena, R., and Cruz-Suárez, H. (2016). Controlled semi-Markov chains with risk-sensitive average cost criterion. *J. Optim. Theory Appl.*, 170(2):670–686.

Chung, K. J. and Sobel, M. J. (1987). Discounted MDPs: distribution functions and exponential utility maximization. *SIAM J. Control Optim.*, 25(1):49–62.

Di Masi, G. B. and Stettner, L. (1999). Risk sensitive control of discrete time partially observed Markov processes with infinite horizon. *Stochastics Stochastics Rep.*, 67(3-4):309–322.

Fan, K. (1952). Fixed-point and minimax theorems in locally convex topological linear spaces. *Proc. Nat. Acad. Sci. U.S.A.*, 38:121–126.

Fan, K. (1953). Minimax theorems. *Proc. Nat. Acad. Sci. U.S.A.*, 39:42–47.

Federgruen, A., Hordijk, A., and Tijms, H. C. (1979). Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion. *Stochastic Process. Appl.*, 9(2):223–235.

Fleming, W. H. and McEneaney, W. M. (1995). Risk-sensitive control on an infinite time horizon. *SIAM J. Control Optim.*, 33(6):1881–1915.

Ghosh, M. K. and Bagchi, A. (1998). Stochastic games with average payoff criterion. *Appl. Math. Optim.*, 38(3):283–301.

Ghosh, M. K. and Goswami, A. (2006). Partially observable semi-Markov games with discounted payoff. *Stoch. Anal. Appl.*, 24(5):1035–1059.

Ghosh, M. K. and Goswami, A. (2008). Partially observed semi-Markov zero-sum games with average payoff. *J. Math. Anal. Appl.*, 345(1):26–39.

Ghosh, M. K., Kumar, K. S., and Pal, C. (2016). Zero-sum risk-sensitive stochastic games for continuous time Markov chains. *Stoch. Anal. Appl.*, 34(5):835–851.

Ghosh, M. K., McDonald, D., and Sinha, S. (2004). Zero-sum stochastic games with partial information. *J. Optim. Theory Appl.*, 121(1):99–118.

Ghosh, M. K. and Pradhan, S. (2020). Zero-sum risk-sensitive stochastic differential games with reflecting diffusions in the orthant. *ESAIM Control Optim. Calc. Var.*, 26:Paper No. 114, 33.

Ghosh, M. K. and Saha, S. (2014). Risk-sensitive control of continuous time Markov chains. *Stochastics*, 86(4):655–675.

Ghosh, M. K., Suresh Kumar, K., Pal, C., and Pradhan, S. (2021). Nonzero-sum risk-sensitive stochastic differential games with discounted costs. *Stoch. Anal. Appl.*, 39(2):306–326.

Golui, S., Pal, C., and Saha, S. (2022). Continuous-time zero-sum games for Markov decision processes with discounted risk-sensitive cost criterion. *Dyn. Games Appl.*, 12(2):485–512.

Guo, X. and Hernández-Lerma, O. (2003). Zero-sum games for continuous-time Markov chains with unbounded transition and average payoff rates. *J. Appl. Probab.*, 40(2):327–345.

Guo, X. and Hernández-Lerma, O. (2005). Zero-sum continuous-time Markov games with unbounded transition and discounted payoff rates. *Bernoulli*, 11(6):1009–1029.

Guo, X. and Hernández-Lerma, O. (2007). Zero-sum games for continuous-time jump Markov processes in Polish spaces: discounted payoffs. *Adv. in Appl. Probab.*, 39(3):645–668.

Guo, X. and Hernández-Lerma, O. (2009). *Continuous-time Markov decision processes*, volume 62 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin. Theory and applications.

Guo, X. and Liao, Z.-W. (2019). Risk-sensitive discounted continuous-time Markov decision processes with unbounded rates. *SIAM J. Control Optim.*, 57(6):3857–3883.

Guo, X. and Piunovskiy, A. (2011). Discounted continuous-time Markov decision processes with constraints: unbounded transition and loss rates. *Math. Oper. Res.*, 36(1):105–132.

Guo, X. and Zhang, J. (2019). Risk-sensitive continuous-time Markov decision processes with unbounded rates and Borel spaces. *Discrete Event Dyn. Syst.*, 29(4):445–471.

Hernández-Hernández, D. and Marcus, S. I. (1996). Risk sensitive control of Markov processes in countable state space. *Systems Control Lett.*, 29(3):147–155.

Hernández-Lerma, O. and Lasserre, J. B. (1996). *Discrete-time Markov control processes*, volume 30 of *Applications of Mathematics (New York)*. Springer-Verlag, New York. Basic optimality criteria.

Hernández-Lerma, O. and Lasserre, J. B. (1999). *Further topics on discrete-time Markov control processes*, volume 42 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Technology Press of M.I.T., Cambridge, Mass.; John Wiley & Sons, Inc., New York-London.

Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management Sci.*, 18:356–369.

Huang, X. and Guo, X. (2020). Nonzero-sum stochastic games with probability criteria. *Dyn. Games Appl.*, 10(2):509–527.

Huang, X., Guo, X., and Peng, J. (2017). A probability criterion for zero-sum stochastic games. *J. Dyn. Games*, 4(4):369–383.

Huang, Y., Guo, X., and Li, Z. (2013). Minimum risk probability for finite horizon semi-Markov decision processes. *J. Math. Anal. Appl.*, 402(1):378–391.

Huang, Y., Guo, X., and Song, X. (2011). Performance analysis for controlled semi-Markov systems with application to maintenance. *J. Optim. Theory Appl.*, 150(2):395–415.

Huang, Y., Lian, Z., and Guo, X. (2018). Risk-sensitive semi-Markov decision processes with general utilities and multiple criteria. *Adv. in Appl. Probab.*, 50(3):783–804.

Huo, H. and Guo, X. (2020). Risk probability minimization problems for continuous-time Markov decision processes on finite horizon. *IEEE Trans. Automat. Control*, 65(7):3199–3206.

Huo, H. and Wen, X. (2021). Risk probability optimization problem for finite horizon continuous time Markov decision processes with loss rate. *Kybernetika (Prague)*, 57(2):272–294.

Huo, H., Zou, X., and Guo, X. (2017). The risk probability criterion for discounted continuous-time Markov decision processes. *Discrete Event Dyn. Syst.*, 27(4):675–699.

James, M. R., Baras, J. S., and Elliott, R. J. (1994). Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems. *IEEE Trans. Automat. Control*, 39(4):780–792.

Jaśkiewicz, A. (2002). Zero-sum semi-Markov games. *SIAM J. Control Optim.*, 41(3):723–739.

Jaśkiewicz, A. (2007). Average optimality for risk-sensitive control with general state space. *Ann. Appl. Probab.*, 17(2):654–675.

Kira, A., Ueno, T., and Fujita, T. (2012). Threshold probability of non-terminal type in finite horizon Markov decision processes. *J. Math. Anal. Appl.*, 386(1):461–472.

Lal, A. K. and Sinha, S. (1992). Zero-sum two-person semi-Markov games. *J. Appl. Probab.*, 29(1):56–72.

Lippman, S. A. (1973). Semi-Markov decision processes with unbounded rewards. *Management Sci.*, 19:717–731.

Luque-Vásquez, F. (2002). Zero-sum semi-Markov games in Borel spaces: discounted and average payoff. *Bol. Soc. Mat. Mexicana (3)*, 8(2):227–241.

Menaldi, J.-L. and Robin, M. (2005). Remarks on risk-sensitive control problems. *Appl. Math. Optim.*, 52(3):297–310.

Nagai, H. (1996). Bellman equations of risk-sensitive control. *SIAM J. Control Optim.*, 34(1):74–101.

Nowak, A. S. (1985). Measurable selection theorems for minimax stochastic optimization problems. *SIAM J. Control Optim.*, 23(3):466–476.

Parthasarathy, T. and Sinha, S. (1989). Existence of stationary equilibrium strategies in non-zero-sum discounted stochastic games with uncountable state space and state-independent transitions. *Internat. J. Game Theory*, 18(2):189–194.

Piunovskiy, A. and Zhang, Y. ([2020] ©2020). *Continuous-time Markov decision processes—Borel space models and general control strategies*, volume 97 of *Probability Theory and Stochastic Modelling*. Springer, Cham. With a foreword by Albert Nikolaevich Shiryaev.

Saha, S. (2014). Zero-sum stochastic games with partial information and average payoff. *J. Optim. Theory Appl.*, 160(1):344–354.

Sakaguchi, M. and Ohtsubo, Y. (2010). Optimal threshold probability and expectation in semi-Markov decision processes. *Appl. Math. Comput.*, 216(10):2947–2958.

Sakaguchi, M. and Ohtsubo, Y. (2013). Markov decision processes associated with two threshold probability criteria. *J. Control Theory Appl.*, 11(4):548–557.

Shapley, L. S. (1953). Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.*, 39:1095–1100.

Suresh Kumar, K. and Pal, C. (2013). Risk-sensitive control of pure jump process on countable space with near monotone cost. *Appl. Math. Optim.*, 68(3):311–331.

Vrieze, K. (1989). Zero-sum stochastic games. A survey. *CWI Quarterly*, 2(2):147–170.

Wei, Q. (2018). Zero-sum games for continuous-time Markov jump processes with risk-sensitive finite-horizon cost criterion. *Oper. Res. Lett.*, 46(1):69–75.

Wei, Q. and Chen, X. (2019a). Risk-sensitive average continuous-time Markov decision processes with unbounded rates. *Optimization*, 68(4):773–800.

Wei, Q. and Chen, X. (2019b). Risk-sensitive average equilibria for discrete-time stochastic games. *Dyn. Games Appl.*, 9(2):521–549.

Wei, Q. and Chen, X. (2021). Nonzero-sum risk-sensitive average stochastic games: the case of unbounded costs. *Dyn. Games Appl.*, 11(4):835–862.

White, D. J. (1993). Minimizing a threshold probability in discounted markov decision processes. *Journal of Mathematical Analysis and Applications*, 173(2):634 – 646.

Whittle, P. (1990). *Risk-sensitive optimal control*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Ltd., Chichester.

Wu, C. and Lin, Y. (1999). Minimizing risk models in Markov decision processes with policies depending on target values. *J. Math. Anal. Appl.*, 231(1):47–67.

List of published and communicated papers

Based on the work in this thesis, the following research articles are published or communicated.

1. Arnab Bhabak, Subhamay Saha (2021) Continuous-time zero-sum games with probability criterion, Stochastic Analysis and Applications, 39:6, 1130-1143, DOI: 10.1080/07362994.2021.1871627
2. Arnab Bhabak, Chandan Pal and Subhamay Saha (2022) Zero-sum semi-Markov games with a probability criterion, Stochastics, 94:3, 415-431, DOI: 10.1080/17442508.2021.1957891.
3. Arnab Bhabak, Subhamay Saha (2022) Risk-sensitive semi-Markov decision problems with discounted cost and general utilities, Statistics and Probability Letters, Volume 184, 109408, DOI: 10.1016/j.spl.2022.109408.
4. Arnab Bhabak, Subhamay Saha (2023) Zero and non-zero sum risk-sensitive Semi-Markov games, Stochastic Analysis and Applications, 41:1, 134-151, DOI: 10.1080/07362994.2021.1993447
5. Arnab Bhabak, Subhamay saha (2022) Partially Observable Discrete-time Discounted Markov Games with General Utility, arXiv:2211.07888(communicated).