Semi-Markov Decision Processes and Perfect Information Semi-Markov Stochastic Games

Kushal Guha Bakshi

THESIS
SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
TO
MATHEMATICS DEPARTMENT
JADAVPUR UNIVERSITY
KOLKATA, INDIA
700 032

February 2024

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **"Semi-Markov Decision Processes and Perfect Information Semi-Markov Stochastic Games"**, submitted by **Shri Kushal Guha Bakshi** who got his name registered on 23rd February, 2018 (Index No: 85/18/*Maths.*/25) for the award of Ph.D. (Science) degree of Jadavpur University, is absolutely based upon his own work under the supervision of **Professor Sagnik Sinha**, Department of Mathematics, Jadavpur University and that neither this thesis nor any part of it has been submitted for either any degree/diploma or any other academic award anywhere before.

21/02/2024

Professor
DEPARTMENT OF MATHEMATICS
Jadavpur University
Kolkata – 700 032 West Bengal

....
Professor Sagnik Sinha
Mathematics Department
Jadavpur University
Kolkata-700032

Dedicated to my parents

Acknowledgements

I am glad to take the opportunity to acknowledge and thank all those who helped me directly or indirectly in completing my doctoral thesis. I am immensely indebted to my thesis supervisor **Professor Sagnik Sinha**, Department of Mathematics, Jadavpur University for his encouragement, countless advice and valuable criticisms from the first day of my research period. He introduced this subject to me, taught me with patience and supported me at every steps of my student life. Without his enthusiasm and augmentation, this thesis could not even be formulated in the way it is.

I would like to express my deepest gratitude to **Prof. Arup Kumar Das**, Senior Technical Officer, ISI, Kolkata; **Dr. Prasenjit Mondal**, Assistant Professor, Government General Degree College, Ranibandh and **Dr. Rwitam Jana**, Senior Analyst, Data and Analytics, Tata Steel Limited for their helpful guidance and cooperation to carry out this work. Although no words of gratitude are enough for the help and encouragement that I received from them.

I gratefully acknowledge the Department of Mathematics, Jadavpur University for its indispensable support. I convey my sincere gratitude to the Doctoral Committee members for their valuable suggestions and comments during the period of my work. I am very much grateful to all non-teaching staff relating to my research work at Jadavpur University for giving me various facilities during my work.

I am also deeply thankful to my friends and colleagues who have provided a supportive academic community. Their camaraderie, discussions, and shared experiences have enriched my research and made this journey both intellectually stimulating and enjoyable. Special thanks to Dr. Ashesh Paul, Dr. Sandip Dalui, Mr. Salil Ghosh, Ms. Srijita Mondal and Mrs. Satakshi Ghosh for their collaborative spirit and camaraderie.

My gratitude extends to my parents, Mr. Jayanta Guha Bakshi and Mrs. Dolly Guha Bakshi, whose love, encouragement, and sacrifices have been the foundation of my academic pursuits. Their unwavering support and belief in my abilities have been a constant source of inspiration. I am truly grateful for the values they instilled in me and for being my pillars of strength.

Last but not least, I want to express my deepest appreciation to Ms. Somasree Mazumder and Mr. Arunabha Sengupta. Their unwavering support, understanding, and patience have been my refuge during the challenging times of this journey. Their belief in me has been a driving force, and I look forward to a future filled with continued collaboration and shared accomplishments.

I must be grateful to **INSPIRE-DST**(Department of Science and Technology, Govt. of India) for their financial support since my graduate study. I am deeply indebted to **CR7** and **MSD** who are the core of my inspiration and mental strength.

Jadavpur, Kolkata
February, 2024 Kushal Guha Bakshi

Preface

The thesis entitled **Semi-Markov Decision Processes and Perfect Information Semi-Markov Stochastic Games** concentrates on proving the existence of solutions for various finite discounted and undiscounted semi-Markov decision processes and perfect information semi-Markov/ Stochastic games. We also discuss some efficient algorithms to solve such special classes of semi-Markov decision processes and semi-Markov (Stochastic) games. The main results are stated as follows:

The first chapter is introductory in nature. Here we present the required definitions and introduce the notations used in this dissertation. We also present a brief survey on the literature of zero-sum two person matrix games. To introduce stochastic games, we start from the very beginning of Markov decision processes and its different payoff criterion. Finally we describe semi-Markov decision processes as well as semi-Markov games.

In the second chapter, we study Semi-Markov Decision Processes (SMDPs) With Vector Pay-offs under discounted as well as limiting ratio average payoff (undiscounted) structure and prove the existence of pure stationary/semi-stationary Pareto-optimal strategies. We also discuss efficient algorithms to compute pure stationary and pure semi-stationary Pareto-optimal strategies for both the discounted and undiscounted (limiting ratio average) SMDP models with vector rewards respectively.

In the third chapter, we study Undiscounted Perfect Information Semi-Markov Stochastic Games and prove the existence of the value and a pair of optimal pure semi-stationary strategies for both the players. An algorithm has also been provided in this chapter to solve such stochastic/semi-Markov games. Some numerical examples are added too.

The thesis ends with a chapter on Undiscounted Semi-Markov Decision Processes With Countably Infinite Action Spaces. We establish the existence of a near-optimal pure semi-stationary strategy of the decision maker in such SMDP models. The analysis here is done without putting any bounded condition on the reward structure. However, we allow strategies/policies with finite support only. We also propose an efficient algorithm to compute the value and a near-optimal pure semi-stationary strategy of the decision maker in such a semi-Markov decision process. We further develop an optimality equation of such SMDP model using a recurrence condition.

Abbreviations

LP	Linear Programming
MOLP	Multi Objective Linear Programming
MDP	Markov Decision Process
SMDP	Semi-Markov Decision Process
SG	Stochastic Game
SMG	Semi-Markov Game
PISMG	Perfect Information Semi-Markov Game
PISG	Perfect Information Stochastic Game

Glossary of Notation

Although the chapterwise special notations are specified in the respective chapter, still we mention below some most frequently used notations:

Spaces

$\mathbb{R}^n$	real n-dimensional space
$\mathbb{R}$	the real line
$\mathbb{R}^{m\times k}$	the space of real $m \times k$ matrices
$\mathbb{R}^z_+$	real z-dimensional space including the point $+\infty$
$\mathbb{R}^z_-$	real z-dimensional space including the point $-\infty$
$\mathbb{N}$	set of natural numbers
S	the state space of the player/ decision maker
A and B	action space of player-I and player-II respectively
$\mathbb{P}(D)$	the family of probability distributions on a finite set D

Vectors

x^t	the transpose of a vector x
e_m	an m-dimensional vector of all ones
$x^t y$	the standard inner product of vectors in $\mathbb{R}^n$
$x \geq y$	$x_i \geq y_i,\ i = 1,...,n$
$x > y$	$x_i > y_i,\ i = 1,...,n$

Sets

$A(s)$	action set of player-I in the state $s \in S$
$B(s)$	action set of player-II in the state $s \in S$
K	set of all admissible triplets $\{(s,i,j) \mid s \in S, i \in A(s), j \in B(s)\}$
$\|\alpha\|$	cardinality of a finite set α

Matrices

$A = ((a_{ij}))$	a matrix with real entries a_{ij}
$det(M)$	the determinant of a square matrix M
M^{-1}	the inverse of a matrix M
A^t	the transpose of a matrix A
I_n	the identity matrix of order n

$val(A)$	minimax value of a matrix game A

History and Strategy Spaces

$hist_n$	the space of all histories upto n-th decision epoch
Π_1 and Π_2	spaces of all behavioural strategies of player-I and player-II
F_1^s and F_2^s	spaces of all stationary strategies of player-I and player-II
F_1^{sp} and F_2^{sp}	spaces of all pure stationary strategies of player-I and player-II
ξ_1^s and ξ_2^s	spaces of all semi-stationary strategies of player-I and player-II
ξ_1^{sp} and ξ_2^{sp}	spaces of all pure semi-stationary strategies of player-I and player-II

Miscellaneous Symbols

For SGs and SMGs: For $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$

$\mathbb{P}_{\pi_1\pi_2}(. \mid X_1 = s)$	the probability measure determined by (π_1, π_2) and initial state $s \in S$.
$\mathbb{E}_{\pi_1\pi_2}(. \mid X_1 = s)$	the expectation operator determined by (π_1, π_2) and initial state $s \in S$.
$v_\beta(s, \pi_1, \pi_2)$	the total expected β discounted pay-off of player-I in the state $s \in S$ in the zero-sum case.
$v_\beta(\pi_1, \pi_2)$	the β discounted pay-off vector $[v_\beta(s, \pi_1, \pi_2)]_{s=1}^{z}$
$v_\beta(s)$	the β discounted value in the state $s \in S$.
$\phi(s, \pi_1, \pi_2)$	the expected undiscounted (limiting ratio average) pay-off player-I in the state $s \in S$ in the zero-sum case.
$\phi(s)$	the undiscounted (limiting ratio average) value in the state $s \in S$.
$Q(f_1, f_2)$	transition probability matrix for $(f_1, f_2) \in F_1^s \times F_2^s$
$Q^*(f_1, f_2)$	Cesaro limiting matrix of $Q(f_1, f_2)$

For SMDPs with vector rewards:

$V_{\beta}^{\pi}(r,s)$	the total expected β discounted pay-off (real valued) of the decision maker in the state $s \in S$ using the strategy $\pi \in \Pi$.
$V_{\beta}^{\pi}(\bar{r},s)$	the β discounted value profile of the decision maker for the $l \times 1$ reward vector $\bar{r} = (r_1, \cdots, r_l)$.
$Count_{\beta}^{s^{'},\pi}(s)$	the counter of the state $s^{'} \in S$
$\phi(r,s,\pi)$	the expected undiscounted (limiting ratio average) pay-off of the decision maker in the state $s \in S$.
$\phi(\bar{r},s,\pi)$	the expected undiscounted (limiting ratio average) pay-off profile of the decision maker.

For SMDPs with countably infinite action space:

Γ_{∞}	the undiscounted (limiting ratio average) SMDP with countably infinite action space.
Γ_n	the $n(\in \mathbb{N})$-truncated SMDP corresponding to Γ_{∞}.
$v_{\infty}(s)$	the undiscounted (limiting ratio average) value of Γ_{∞}
$v_n(s)$	the undiscounted (limiting ratio average) value of Γ_n

Numbering

For internal referencing Section j in Chapter i is denoted by $i.j$ and $i.j.k$ is used to refer Item k of Section j in Chapter i. For example, the triple 2.3.5 refers to Item 5 in Section 3 of Chapter 2. All items (e.g., Lemma, Theorem, Example, Remark etc.) are identified in this fashion. For equation, $(i.j.k)$ is used to refer Equation k in Section j in Chapter i. We use brackets [ ] for a bibliographical reference.

List of Research Papers

- **K. G. Bakshi**. Semi-Markov Decision Processes with Vector Pay-Offs (2022). Proceedings of the Seventh International Conference on Mathematics and Computing, Advances in Intelligent Systems and Computing. DOI: https://doi.org/10.1007/978-981-16-6890-6_76. **Published**

- **S. Sinha, K. G. Bakshi**. On Zero-Sum Two Person Perfect Information Semi-Markov Games (2023). Proceedings of the Ninth International Conference on Mathematics and Computing, Lecture Notes in Networks and Systems 697. DOI: https://doi.org/10.1007/978-981-99-3080-7_23. **Published**

- **K. G. Bakshi, S. Sinha**. On Zero-Sum Two Person Perfect Information Stochastic Games (2023): Accepted for publication in the **Journal of Calcutta Mathematical Society.**

- **K. G. Bakshi, S. Sinha. Undiscounted Semi-Markov Decision Processes With Vector Pay-Offs: Communicated**

- **K. G. Bakshi, S. Sinha. Semi-Markov Decision Processes With Countably Infinite Action Space: Communicated**

Table of contents

Chapter 1

Background and Preliminaries

1.1 Theory Of Games

A theory of **Games of strategy** was introduced in mathematics between 1928 [47] and 1944. John Von Neumann is known as the pioneer of the theory of games. In 1944, his work with Morgenstern was a seminal publication of the book: 'Theory Of Games And Economic Behaviour' [48]. A game is simply a set of descriptive rules. A play of the game includes every particular instance in which the game is played from beginning to end. The participants are called the players. A game consists of a sequence of moves of the players, while a play comprises a sequence of choices made by them. The decisive step in the mathematical treatment of games is the normalisation achieved by introduction of pure strategies(actions). A pure strategy is a plan formulated by a player prior to the start of a play, which covers all of the possible decisions which he may face during any play permitted by the rules of the game. Von Neumann first considered games with a finite number of pure strategies i.e., finite games.

1.1.1 Zero-Sum Two Person Games (Matrix Games)

These games are played by two players and what one player wins, other player loses. If player-I chooses his i-th pure strategy $(i = 1, 2, \cdots, m)$ and player-II chooses his j-th pure strategy $(j = 1, 2, \cdots, n)[m, n \in \mathbb{N}]$, then the zero-sum two person game can be described by the $m \times n$ matrix $A_{m \times n} = a(i, j)$, where $a(i, j)$ is the payment given to player-I by player-II. Mixed strategies (lotteries on the set of pure strategies) for player-I and II are denoted by the m-tuple $x = (x_1, x_2, \cdots, x_m)$ and n-tuple $y = (y_1, y_2, \cdots, y_n)$ respectively with $x_i \geq 0$ $(\forall i = 1, 2, \cdots, m)$, $y_j \geq 0$, $(\forall j = 1, 2, \cdots, n)$ and $\sum_{i=1}^{m} x_i = 1$, $\sum_{j=1}^{n} y_j = 1$. The expected payment to the player-I by player-II is denoted by the bilinear function $\phi(x,y)$, where $\phi(x,y)$ denotes the payment to player-I by player-II when player-I and II plays their mixed strategies x and y respectively. Let X and Y be the set of mixed strategies available to player-I and II respectively. The bilinear function is defined in the cartesian product of X and Y and it is calculated as:

$$\phi(x,y) = \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} x_i a_{ij} y_j = x^T A y$$

The minimax theorem asserts that

$$\max_x \min_y \phi(x,y) = \min_y \max_x \phi(x,y)$$

The unique minimax value of ϕ is called the value of the game and it is denoted by $val(A)$ (i.e., $val(A) = \phi(x^*,y^*)$). The optimal mixed strategy pair of player-I and I is denoted by (x^*,y^*), satisfying

$$\phi(x^*,y) \geq \phi(x^*,y^*) \geq \phi(x,y^*),\ \forall(x,y) \in X \times Y.$$

1.2 Introduction To Stochastic Games

Shapley (1953) ([42]) introduced 'Stochastic games' in his paper, which are also called Markov games. If two players play a matrix game repeatedly over an infinite time horizon and the average payoff is considered, then the value of this infinitely repeated game coincides with the value of the one shot game (by Folk Theorem [11]). Shapley introduced the idea of not playing the same matrix game everyday (i.e., in every stage of the game), but playing one among finitely many matrix games, with a motion among them governed by the present game and the actions chosen there in such a manner that the game is certain to stop in finite time almost surely. Then the payoffs of the players can be formulated as the ratio of two bilinear forms. Minimax theorem for such games was established by von Neumann [35] and an elementary proof was subsequently given by Loomis [26]. Shapley's stochastic games are called 'Stopping Stochastic Games' nowadays. Thus stopping stochastic games were born naturally from matrix games. Since 1953, the theory of stochastic games has been extended in various directions. Stochastic games with zero stop probabilities (where stop probabilities are zero everywhere) have also been studied under discounted as well as undiscounted pay-offs (Gillette [12], Hoffman-Karp [13]).

1.2.1 Finite Zero-Sum Two Person Stochastic Games

A zero-sum two person finite stochastic game is described by a collection of five objects $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, r >$, where $S = \{1, 2, \cdots, z\}$ is the finite non-empty state space and $A(s) = \{1, 2, \cdots, m_s\}, B(s) = \{1, 2, \cdots, n_s\}$ are respectively the non-empty sets of admissible actions of the players I and II respectively in the state s. Let us denote $K = \{(s, i, j) : s \in S, i \in A(s), j \in B(s)\}$ to be the set of admissible triplets. For each $(s, i, j) \in K$, we denote $q(. \mid s, i, j)$ to be the transition law of the game. Finally r is the real valued functions on K, which represents the immediate (expected) reward for the player-I (whereas -r is the reward for the player-II). Let us consider player-I as the maximiser and player-II as the minimiser in the zero-sum two person stochastic game.

The Stochastic game over infinite time is played as follows. At the 1-st decision epoch, the game starts at $s \in S$ and the players I and II simultaneously and independently choose actions $i \in A(s)$ and $j \in B(s)$ respectively. Consequently player-I and II get immediate rewards $r(s, i, j)$ and $-r(s, i, j)$ respectively and the game moves to the state $s^{'}$ with probability $q(s^{'} \mid s, i, j)$. After reaching the state $s^{'}$ on the next decision epoch, the game is repeated over infinite time with the state s replaced by $s^{'}$. Shapley extended the idea of defining SGs where $\sum_{s' \in S} q(s^{'} \mid s, i, j) < 1$ for all $(s, i, j) \in K$ and the play terminates with probability $1 - \sum_{s' \in S} q(s^{'} \mid s, i, j) < 1$. Such games are called 'stopping SGs'. The 'non-stopping SGs' are those where $\sum_{s' \in S} q(s^{'} \mid s, i, j) = 1$ for all $(s, i, j) \in K$, i.e., the play never terminates.

By a strategy (behavioural) π_1 of the player-I, we mean a sequence $\{(\pi_1)_n(. \mid hist_n)\}_{n=1}^{\infty}$, where $(\pi_1)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to n-th decision epoch (where $hist_n$=$(s_1, a_1, b_1, s_2, a_2, b_2 \cdots, s_{n-1}, a_{n-1}, b_{n-1}, s_n)$ for $n \geq 1$, $hist_1 = (s_1)$ and $(s_k, a_k, j_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi_1)_n(. \mid hist_n)$ on $A(s_n)$. Behavioural strategy π_2 for player-II can be defined analogously. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π_1 and Π_2 to be the sets of strategy (behavioural) spaces of the players I and II respectively. A strategy $\pi_1 = \{\pi_{1n}\}_{n=1}^{\infty}$ is called a stationary strategy if $\exists$ a map $f : S \to \mathbb{P}(A) = \{\mathbb{P}(A(s)) : s \in S\}$, where $\mathbb{P}(A(s))$ is the set of probability distribution on $A(s)$ such that $\pi_{1n} = f$ for all n and $f(s) \in \mathbb{P}(A(s))$. A stationary strategy for player-I is defined as z tuple $f = (f(1), f(2), \cdots, f(z))$, where each $f(s)$ is the probability distribution on $A(s)$ given by $f(s) = (f(s, 1), f(s, 2), \cdots, f(s, m_s))$. $f(s, i)$ denotes the probability of choosing action i in the state s by player-I. By similar manner, one can define a stationary strategy g for player-II as $g = (g(1), g(2), \cdots, g(z))$ where each $g(s)$ is the probability distribution on $B(s)$. Let us denote F_1^s and F_2^s to be the set of stationary strategies for player-I and II respectively.

A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F_1^{sp} and F_2^{sp} to be the set of pure stationary strategies of the players I and II respectively.

1.2.2 Discounted Pay-Off Criterion In Zero-Sum Two Person Stochastic Games

The notion of discounted pay-off in stochastic game is applied directly from dynamic programming. We define the discounted pay-off of the players as follows:

Definition 1.1. *Let us consider a pair of behavioural strategies $(\pi_1, \pi_2) \in (\Pi_1 \times \Pi_2)$. For a finite zero-sum two person stochastic game we define the β-discounted pay-off for player-I from player-II as*

$$v_\beta(s, \pi_1, \pi_2) = \mathbb{E}_{\pi_1, \pi_2}[\sum_{m=1}^{\infty} \beta^{m-1} r(X_m, A_m, B_m) \mid X_1 = s] \tag{1.1}$$

Where $r(X_m, A_m, B_m)$ is the expected reward for the player-I by player-II [where X_m is the state on the mth decision epoch and A_m, B_m are the action chosen by player-I and II respectively on the m-th decision epoch]. The existence of value and a pair of optimal pure stationary strategies for the players in a β-discounted zero-sum two person stochastic games were established by Shapley [42] and Blackwell [2].

Definition 1.2. *A pair of stationary strategies $(f_1^*, f_2^*) \in F_1^s \times F_2^s$ is called optimal if for all $s \in S$:*

$$v_\beta(s, f_1^*, f_2) \geq v_\beta(s, f_1^*, f_2^*) \geq v_\beta(s, f_1, f_2^*) \forall (f_1, f_2) \in F_1^s \times F_2^s \tag{1.2}$$

Shapley proved the uniqueness of the vector $[v_\beta(s, f_1^*, f_2^*)]_{z \times 1}$ as a function of the initial state $s \in S$. We denote this value vector by $v_\beta = [v_\beta(s)]_{z \times 1}$.

Theorem 1.2.1. *The value vector v_β of the β-discounted stochastic game is the unique solution of the system of equations:*

$$v_\beta(s) = val[r(s,i,j) + \beta \sum_{s' \in S} q(s^{'} \mid s,i,j) v_\beta(s^{'})] \forall s \in S. \tag{1.3}$$

The above set of equations is also known as Shapley equation for stochastic games. Furthermore, a pair of pure stationary strategies (f_1^*, f_2^*) is called a pair of optimal strategies for the discounted stochastic games if and only if for all $s \in S$, $(f_1^*(s), f_2^*(s))$ is a pair of optimal strategies of the matrix game in the right hand side of equation (1.3).

1.2.3 Undiscounted Zero-Sum Two Person Stochastic Games

Let $(X_1, A_1, B_1, X_2, A_2, B_2, \cdots)$ be a co-ordinate sequence in $S \times (A \times B \times S)^\infty$. Given behavioural strategy pair $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi_1\pi_2}(. \mid X_1 = s)$ (hence an expectation $\mathbb{E}_{\pi_1\pi_2}(. \mid X_1 = s)$) on the product σ- field of $S \times (A \times B \times S)^\infty$ by Kolmogorov's extension theorem. For a pair of strategies $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$ for the players I and II respectively, the limiting average (undiscounted) pay-off for player-I from player-II, starting from a state $s \in S$ is defined by:

$$\phi(s, \pi_1, \pi_2) = \liminf_{n \to \infty} \frac{1}{n} \mathbb{E}_{\pi_1\pi_2} \sum_{m=1}^{n} [r(X_m, A_m, B_m) \mid X_1 = s] \tag{1.4}$$

Alternatively, for any pair of stationary strategies $(f_1, f_2) \in F_1^s \times F_2^s$ of player-I and II, we write the undiscounted pay-off for player-I from player-II as:

$$\phi(s, f_1, f_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} r^m(s, f_1, f_2) \tag{1.5}$$

(as we know the limit exists for stationary strategy pair (f_1, f_2)) for all $s \in S$. Where $r^m(s, f_1, f_2)$ is the expected reward for player-I at the m-th decision epoch, when player-I chooses f_1 and player-II chooses f_2 respectively and the initial state is s.

Definition 1.3. *For a pair of stationary strategies $(f_1, f_2) \in F_1^s \times F_2^s$, we define the transition probability matrix by:*

$$Q(f_1,f_2) = [q(s^{'} \mid s, f_1(s), f_2(s))]_{s,s^{'}=1}^{z},$$

where $q(s^{'} \mid s, f_1(s), f_2(s)) = \sum_{i \in A(s)} \sum_{j \in B(s)} q(s^{'} \mid s, i, j) f_1(s,i) f_2(s,j)$ is the probability is that the system jumps to the state $s^{'}$ from given state s when the players play the stationary strategies f_1 and f_2 respectively.

Lemma 1(Kemeney and Snell, 1976, [21]) Let Q be any $z \times z$ Markov matrix, then the sequence $\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} Q^m(f_1,f_2)$ converges as $n \to \infty$ to a Markov matrix Q^* (the Cesaro limiting matrix) such that $QQ^* = Q^*Q = Q^*Q^* = Q^*$.

For each $(f_1,f_2) \in F_1^s \times F_2^s$, we define $r(f_1,f_2) = [r(s,f_1,f_2)]_{z\times 1}$ as the expected reward, where for each $s \in S$,

$$r(s,f_1,f_2) = \sum_{i\in A(s)} \sum_{j \in B(s)} r(s,i,j) f_1(s,i) f_2(s,j).$$

Now we have the following result:

Proposition 1 For each pair of pure stationary strategies $(f_1,f_2) \in F_1^{sp} \times F_2^{sp}$,

$$\phi(s,f_1,f_2) = [Q^*(f_1,f_2) r(f_1,f_2)](s) \forall s \in S.$$

Where $Q^*(f_1,f_2)$ is the Cesaro limiting matrix of $Q(f_1,f_2)$.

Definition 1.4. *A zero-sum two person undiscounted stochastic game is said to have a value vector $\phi = [\phi(s)]_{z\times 1}$ if $\sup_{\pi_1\in\Pi_1} \inf_{\pi_2 \in \Pi_2} \phi(s,\pi_1,\pi_2) = \phi(s) = \inf_{\pi_2\in\Pi_2} \sup_{\pi_1 \in \Pi_1}$ $\phi(s,\pi_1,\pi_2)$ for all $s \in S$. A pair of strategies $(\pi_1^*, \pi_2^*) \in \Pi_1, \times \Pi_2$ is said to be an optimal strategy pair for the players if $\phi(s,\pi_1^*,\pi_2) \geq \phi(s) \geq \phi(s,\pi_1,\pi_2^*)$ for all $s \in S$ and all $(\pi_1,\pi_2) \in \Pi_1 \times \Pi_2$.*

Previously Mertens and Neyman [31] proved the existence of a value in a zero-sum two person undiscounted stochastic game but optimal strategies need not exist. But ϵ-optimal strategies may exist for both the players, i.e., given $\epsilon > 0$ there exists a pair of strategies $(\pi_{1\epsilon}, \pi_{2\epsilon}) \in \Pi_1 \times \Pi_2$ such that

$$\phi(s,\pi_{1\epsilon},\pi_2) \geq \phi(s) - \epsilon \text{ and } \phi(s,\pi_1,\pi_{2\epsilon}) \leq \phi(s) + \epsilon \text{ for all } s \in S \text{ and all } (\pi_1,\pi_2) \in \Pi_1 \times \Pi_2$$

Mertens and Neyman [31] have also showed the relationship between the discounted and undiscounted value of a zero-sum two person SG, which is:

$$\lim_{\beta\uparrow 1}(1-\beta) v_\beta(s) = \phi(s) \text{ for all } s \in S.$$

1.3 Markov Decision Processes

Markov decision process (MDP) is a special case of dynamic programming. It is named after Andrey Markov. In the mid of twentieth century, the two topics, i.e., Stochastic games and Markov decision processes were studied extensively. Further formulation was done by Bellman [1]. Analysis of Manne[30], Blackwell [2], Derman [4] were also studied in literature. An MDP is nothing but an SG in which player-I (or player-II) is a dummy.

Definition 1.5. *A finite MDP is defined by an ordered quadruple $< S, A = A(s), q, r >$, where S is the finite state space$(= \{1, 2, \cdots, z\})$, $A(s) = \{1, 2, \cdots, m_s\}$ is the set of finite actions available for the decision maker in the state $s \in S$, $q(. \mid s, a)$ is the transition probability to move to a new state from the state s while choosing the action $a \in A(s)$ and r is a function from $S \times A \to \mathbb{R}$, where $r(s, a)$ denotes the immediate reward for the action $a \in A(s)$ taken by the decision maker. Note that for any $s^{'} \in S$, $q(s^{'} \mid s, a) \geq 0$ and $\sum_{s^{'} \in S} q(s^{'} \mid s, a) = 1$ $\forall a \in A(s)$, $s \in S$.*

The process (i.e., MDP) proceeds over infinite time horizon ($t = 1, 2, \cdots$) as follows:
(i) At the first decision epoch (i.e.,$t = 1$) the system starts from an initial state $s \in S$.
(ii) Now the decision maker chooses an action a from the available set of actions $A(s)$.
(iii) As a result, the decision maker gets an immediate reward $r(s, a)$. and the system moves to a new state $s^{'} \in S$ with transition probability $q(s^{'} \mid s, a)$.
(iv) After the transition to the next day, the process is repeated over and over again with s replaced with $s^{'}$.

Definition 1.6. *By a strategy (behavioural) π of the decision maker, we mean a sequence $\{(\pi)_n(. \mid hist_n)\}_{n=1}^{\infty}$, where $(\pi)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n$=$(s_1, a_1, \cdots, s_{n-1}, a_{n-1}, s_n)$ for $n \geq 1$, $hist_1 = (s_1)$ and $(s_k, a_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi)_n(. \mid hist_n)$ on $A(s_n)$. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π to be the set of strategy (behavioural) spaces of decision maker. A strategy $g = \{g_n\}_{n=1}^{\infty}$ is called Markov if g_n depends on the history h_n, through the current state s_n and the decision epoch number n. A strategy $\pi = \{\pi_n\}_{n=1}^{\infty}$ is called a stationary strategy if $\exists$ a map $f : S \to \mathbb{P}(A) = \{\mathbb{P}(A(s)) : s \in S\}$, where $\mathbb{P}(A(s))$ is the set of probability distribution on $A(s)$ such that $\pi_n = f$ for all n and $f(s) \in \mathbb{P}(A(s))$. A stationary strategy for the decision maker is defined as z*

tuple $f = (f(1), f(2), \cdots, f(z))$*, where each* $f(s)$ *is the probability distribution on* $A(s)$ *given by* $f(s) = (f(s,1), f(s,2), \cdots, f(s,m_s))$*.* $f(s,a)$ *denotes the probability of choosing action* a *in the state* s *by the decision maker. Let us denote* F^s *to be the set of stationary strategies for the decision maker.*
A stationary strategy is called pure if any player selects a particular action with probability 1 *while visiting a state* s*. We denote* F^{sp} *to be the set of pure stationary strategies of the decision maker.*

1.3.1 Discounted Pay-Off Criterion In An MDP

Suppose $0 < \beta < 1$ is the discount factor we consider in the MDP model. Then the total discounted pay-off of the decision maker for a behavioural strategy $\pi \in \Pi$ over the infinite horizon will be

$$v_\beta(s,\pi) = \mathbb{E}_\pi[\sum_{m=1}^{\infty} \beta^{m-1} r(X_m, A_m) \mid X_1 = s] \tag{1.6}$$

Let us take $M = \max_{(s,a)} \mid r(s,a) \mid$. This leads

$$\mid v_\beta(s,\pi) \mid \leq M \textstyle\sum_{m=1}^{\infty} \beta^{m-1} = \frac{M}{(1-\beta)}. \ \forall s \in S, \pi \in \Pi.$$

Previously Blackwell [2] and Maitra [29] considered Markov strategies in Markov decision models. Let $g = \{g_n\}_{n=1}^{\infty}$ be a Markov strategy, then the transition matrix $Q(g_n)$ is defined for the transition from nth day to $(n+1)$th day and $r(g_n)$ be the expected reward vector for the nth day. Thus the n-th step transition matrix is defined by $Q_n(g) = Q(g_1)Q(g_2)\cdots Q(g_n)$. Thus the total β- discounted pay-off using the Markov strategy g will be the $z \times 1$ column vector:

$$v_\beta(g) = \sum_{n=0}^{\infty} \beta^n Q_n(g) r(g_{n+1}), \tag{1.7}$$

where $Q_0(g) = I$ is the identity matrix of dimension $z \times z$. Now, if we consider the stationary strategy f, then $g_n = f$, $\forall n \in \{1, 2 \cdots, \infty\}$. Thus we obtain:

$$v_\beta(f) = \sum_{n=0}^{\infty} \beta^n Q^n(f) r(f). \tag{1.8}$$

For each $0 < \beta < 1$, the matrix $\beta Q(f)$ is a sub-stochastic matrix, with all the rows having a fixed row-sum < 1. Thus $\lim_{n\to\infty} \beta^n Q^n(f) = 0$. Furthermore:

$$\lim_{n\to\infty}(I-\beta Q(f))[I+\beta Q(f)+\beta^2Q^2(f)+\cdots+\beta^{n-1}Q^{n-1}(f)]$$
$$=\lim_{n\to\infty}[I-\beta^nQ^n(f)]=I$$

Since $det[I-\beta^nQ^n(f)]\neq 0$ for sufficiently large n, we conclude $(I-\beta Q(f))$ is a non-singular matrix and

$$(I-\beta Q(f))^{-1}=\sum_{n=0}^{\infty}\beta^nQ^n(f) \tag{1.9}$$

Substituting(1.9) to (1.8) we get the value vector for a stationary strategy $f\in F^s$:

$$v_\beta(f)=(I-\beta Q(f))^{-1}r(f) \tag{1.10}$$

Definition 1.7. *A strategy $\pi^*\in\Pi$ is called an optimal strategy for a β-discounted MDP if*

$$v_\beta(s,\pi^*)\geq v_\beta(s,\pi),\ \forall\pi\in\Pi \text{ and } \forall s\in S.$$

Bellman [1] and Blackwell [2] proved the existence of optimal stationary strategy in a discounted MDP. An efficient linear programming (LP) algorithm to compute an optimal stationary strategy for a discounted MDP is discussed below:

1.3.2 Linear Programming For Solving Discounted MDPs

Let us consider a discounted MDP with discount factor $\beta\in[0,1)$. Then the optimal control problem for the discounted MDP is:

$$\max v_\beta(f)$$

subject to the constraints:

$$f\in F^s.$$

For a stationary strategy f^* which achieves the maximum of the objective function is called the optimal stationary strategy and the corresponding value vector

$v_\beta(f^*) = \max v_\beta(f)$ is called the discounted value of the MDP. Multiplying (1.10) by $(I - \beta Q(f))$ we get:

$$v_\beta(f) = r(f) + \beta Q(f) v_\beta(f) \tag{1.11}$$

Now any optimal stationary strategy f^* must satisfy (1.11). Thus we get for any optimal stationary strategy f^*

$$v_\beta(f^*) = r(f^*) + \beta Q(f^*) v_\beta(f^*) \tag{1.12}$$

Moreover, the value vector v_β must satisfy the optimality equation, i.e.,

$$v_\beta(s) = \max_{a \in A(s)} \{r(s,a) + \beta \sum_{s^{'} \in S} q(s^{'} \mid s,a) v_\beta(s^{'})\}, s \in S \tag{1.13}$$

Thus the value vector is the optimal solution of the following LPP:

LP 1.1

$$\min \textstyle\sum_{s=1}^{z} \frac{1}{z} v(s)$$

subject to the constraints:

$$v(s) \geq r(s,a) + \beta \textstyle\sum_{s^{'} \in S} q(s^{'} \mid s,a) v(s^{'}),\ a \in A(s),\ s \in S.$$

The co-efficients $\frac{1}{z}$ in the objective function can be interpreted as the equal probability that the process starts in any given state. Now by associating each constraint to a variable x_{sa} we get the following dual LPP:

DLP 1.1

$$\max \textstyle\sum_{s=1}^{z} \sum_{a \in A(s)} r(s,a) x_{sa}$$

subject to the constraints:

$$\begin{gathered} \textstyle\sum_{s \in S} \sum_{a \in A(s)} [\delta(s,s^{'}) - \beta q(s^{'} \mid s,a)] x_{sa} = \frac{1}{z},\ s^{'} \in S \\ x_{sa} \geq 0 \end{gathered}$$

where $\delta(s,s^{'})$ is the Kronecker delta function. Now we state the main result associating the linear programming problem to the discounted Markov decision process.

Theorem 1.3.1. *(a) The dual-primal linear programming problem, i.e,. LP* 1.1 *and DLP* 1.1 *possesses finite optimal solutions.*
(b)If $v_0 = (v_0(1), \cdots, v_0(z))$ is the optimal solution of the LP 1.1 *then $v_\beta = v_0$ is the value vector of the discounted MDP.*
(c)Let $x_0 = \{x_0^{sa} \mid s \in S, a \in A(s)\}$ be an optimal solution of the DLP 1.1. *Define $x_0^s = \sum_{a \in A(s)} x_0^{sa}$ for each $s \in S$. Then $x_0^s > 0$ and the strategy $f^*(s,a)$ defined by*

$$f^*(s,a) = \frac{x_0^{sa}}{x_0^s},\ a \in A(s),\ s \in S.$$

is the optimal strategy of the LP.

Corollary 1.3.1.1. *(a)The value vector of the MDP is the unique solution of the optimality equation* (1.13)
(b)Let $s \in S$ and $a_s^* \in A(s)$ is the action which achieves the maximum value in the optimality equation (1.13)*, i.e.,*

$$\begin{aligned} v_\beta(s) &= \{r(s,a_s^*) + \beta \sum_{s' \in S} q(s' \mid s, a_s^*) v_\beta(s')\} \qquad (1.14) \\ &= \max_{a \in A(s)} \{r(s,a) + \beta \sum_{s' \in S} q(s' \mid s, a) v_\beta(s')\} \end{aligned}$$

where $v_\beta(s)$ is the solution of (1.13) *and the stationary strategy $f^* \in F^s$ is defined by:*

$$f^*(s,a) = \begin{cases} 1 & \textit{if } a = a_s^* \\ 0 & \textit{otherwise} \end{cases}$$

for each $s \in S$. Then f^* is the pure optimal stationary strategy of the β-discounted MDP.

1.3.3 Limiting Average Pay-Off Criterion In An MDP

Let us consider a behavioural strategy $\pi \in \Pi$, then the undiscounted (limiting average pay-off) of the decision maker is defined by:

$$\phi(s,\pi) = \liminf_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_\pi[r(X_m, A_m) \mid X_1 = s],\ \forall s \in S.$$

Definition 1.8. *A behavioural strategy π^* is called an optimal strategy in the undiscounted MDP if*

$$\phi(s,\pi^*) \geq \phi(s,\pi),\ \forall \pi \in \Pi \ \mathit{and}\ \forall s \in S.$$

This optimal strategies are also often called Derman [4] optimal strategies, as they are obtained by Derman's average reward criterion [4]. Define $\phi(s) = \sup_{\pi\in\Pi} \phi(s,\pi)$, $\forall s \in S$. Then $\phi(s)$ is called the limiting average value of the MDP, with initial state $s \in S$.

Definition 1.9. ***(Chain structures in MDPs)*** *An MDP is said to be communicating if for every $s, s^{'} \in S$, there exists a strategy $f \in F^s$, such that a state $s^{'}$ is accessible from state s in the Markov matrix $Q(f)$.*
A state is called absorbing if the transition probability to any other state is zero for any stationary strategy f of the decision maker. An absorbing MDP is an MDP in which all states, but one are absorbing for any stationary strategy $f \in F^s$ of the decision maker.
An MDP is irreducible (also called completely ergodic) if the Markov chain $Q(f)$ is irreducible for every stationary strategy $f \in F^s$.
An MDP is said to be unichain if the embedded Markov chain for each strategy is unichain, i.e., if the Markov chain $Q(f)$ has at most one closed irreducible recurrent class plus a possibly empty set of transient states for all strategy $f \in F^s$.
An MDP is called a Multichain if it is not a unichain MDP, i.e., there exists a strategy $f \in F^s$ for which the Markov chain $Q(f)$ has at least two ergodic classes.

Now as the state space S is closed (i.e., any state $s^{'} \notin S$ can not be reached from any state $s \in S$), it follows that every Markov chain has at least one irreducible class. Let $S_1, S_2, \cdots, S_L$ be all irreducible classes of a finite state Markov chain. Then the stochastic matrix Q can be written as:

$$Q = \begin{bmatrix} Q_1 & 0 & \cdots & 0 & 0 \\ 0 & Q_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & Q_L & 0 \\ Q_{(L+1)1} & Q_{(L+1)2} & \cdots & Q_{(L+1)L} & Q_{L+1} \end{bmatrix}$$

where for each $l \in \{1, 2, \cdots, L\}$, Q_l is a square matrix corresponding to the l-th irreducible class S_l of the Markov chain. The states belonging to irreducible classes are called recurrent states and remaining class of states S_{L+1} corresponding to the bottom block are transient states. The matrix $Q_{(L+1)l}$ represents the probabilities with which the system vanishes from the transient states into l-th irreducible class. Now we concentrate on the following results by Blackwell [2] to establish a relationship between Derman [4] and Blackwell [2] optimality.

Lemma 1.3.2. *Let Q be a Markov matrix of dimension $z \times z$.*
(a) Then the sequence $\frac{I+Q+Q^2+\cdots+Q^n}{n+1}$ converges to a Markov matrix Q^* as $n \to \infty$ such that:

$$QQ^* = Q^*Q = Q^*Q^* = Q^*$$

(b) Q^* has the following form:

$$Q^* = \begin{bmatrix} Q_1^* & 0 & \cdots & 0 & 0 \\ 0 & Q_2^* & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & Q_L^* & 0 \\ Q_{(L+1)1}^* & Q_{(L+1)2}^* & \cdots & Q_{(L+1)L}^* & 0 \end{bmatrix}$$

where for each $l \in \{1,2,\cdots,L\}$, Q_l^* has identical rows, each equaling q_l^*, the unique invariant distribution associated to Q_l^*, i.e., the unique solution to $q_l^* Q_l = q_l^*$, $q_l^* \geq 0$ and $\sum_{s \in S_l} q_l^*(s) = 1$, and where

$$Q_{(L+1)l}^* = (I - Q_{L+1})^{-1} Q_{(L+1)l} Q_l^*,\ l \in \{1,2,\cdots,L\}.$$

(c) $rank(I-Q) + rank(Q^*) = z$
(d) For every column vector 'c' of dimension $z \times 1$, the system has a unique solution:

$$Qx = x \text{ and } Q^*x = Q^*c$$

(d) Define $H(\beta) = \sum_{n=0}^{\infty} \beta^n (Q^n - Q^*)$ and $H = (I - Q + Q^*) - Q^*$. Then the matrix $(I - (Q - Q^*))$ is non-singular and $H(\beta) \to H$ as $\beta \to 1$. Also,

$$H(\beta)Q^* = Q^*H(\beta) = HQ^* = Q^*H = 0$$

and

$$(I-Q)H = H(I-Q) = I - Q^*$$

These results were also established by Kemeny and Snell [21]. For the unichain case, we have the following results:

Lemma 1.3.3. *(Puterman [38]) Let Q be a $z \times z$ Markov matrix of a unichain Markov chain. Then Q^* has the following properties:*

(a) Each row of Q^* is identical and equal to the vector $q^* = (q^*(1), q^*(2) \cdots, q^*(z))$, which is the invariant distribution of Q.
(b) Each component of q^* is strictly positive for the recurrent states and is zero for the transient states.

Given a stationary strategy $f \in F^s$, the limiting average pay-off can be written as:

$$\phi(s,f) = \liminf_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{m=1}^{n} r^m(s,f) \text{ for all } s \in S,$$

where $r^m(s,f)$ is the expected pay-off of the decision maker on the m-th decision epoch, when he/she plays stationary strategy $f \in F^s$ and the initial state is s.

We have

$$r^1(s,f) = r(s,f) = [r(f)](s)$$

and for $m = 1, 2, 3, \cdots$

$$r^m(s,f) = \textstyle\sum_{s' \in S} r(s^{'}, f) q^{m-1}(s^{'} \mid s, f) = [Q^{m-1}(f) r(f)](s).$$

Let $\phi(f) = [\phi(s,f)]_{z\times 1}$ be limiting average pay-off vector. We can express it as:

$$\phi(f) = \lim_{N\to\infty} \left[\tfrac{1}{N+1} \textstyle\sum_{n=0}^{N} Q^n(f) r(f)\right] = \left[\lim_{N\to\infty} \tfrac{1}{N+1} \textstyle\sum_{n=0}^{N} Q^n(f)\right] r(f)$$

Thus using lemma 1.3.2, we can write $\phi(f) = Q^*(f) r(f)$.

Now the following theorem establishes the relationship between Blackwell [2] and Derman[4] optimality.

Theorem 1.3.4. *(Derman [4]) Let us consider $f \in F^s$ to be a stationary strategy chosen by the decision maker. Let $Q^*(f)$ to be the Cesaro limiting matrix associated with $Q(f)$. Then,*

$$\textit{(i) } v_\beta(f) = \tfrac{x(f)}{(1-\beta)} + y(f) + \epsilon(\beta, f)$$

where $x(f)$ is the unique solution of

$$(I - Q(f))x = 0 \textit{ and } Q^*(f)x = Q^*(f) r(f),$$

and $y(f)$ is the unique solution of

$$(I - Q(f))y = r(f) - x(f) \textit{ and } Q^*(f)y = 0,$$

and $\epsilon(\beta, f) \to 0$ *as* $\beta \uparrow 1$.

$$\textit{(ii)} \lim_{\beta\uparrow 1}(1-\beta)v_\beta(f) = \phi(f).$$

The above theorem gives the relationship between discounted and undiscounted pay-offs when the values of $\beta \uparrow 1$. Derman [4] proved that there exists a pure stationary strategy f^* which is limiting average optimal as well as β-discounted optimal in the MDP for values of β near 1.

Theorem 1.3.5. *(i) There exists a value of β, namely β_0 such that* $\forall\beta \in [\beta_0, 1)$

$$v_\beta(f^*) = \max_{\pi\in\Pi} v_\beta(\pi)$$

(ii) With f^* described above, we have

$$\phi(f^*) = \max_{\pi\in\Pi} \phi(\pi).$$

1.3.4 Linear Programming Formulation For Limiting Average MDPs

Hordijk and Kallenberg [15] proved that a general limiting average pay-off MDP can be solved by a single linear programming problem. Let us consider the following linear programming problem with the variables $x = (x(1), x(2), \cdots, x(z))$ and $y = (y(1), y(2), \cdots, y(z))$.

LP 1.2

$$\min \sum_{s=1}^{z} d_s x(s)$$

subject to:

$$\begin{gathered}\text{(a) } g(s) \geq \sum_{s'=1}^{z} q(s^{'} \mid s, a) g(s^{'}),\ s \in S,\ a \in A(s)\\ \text{(b) } g(s) + h(s) \geq r(s,a) + \sum_{s'=1}^{z} q(s^{'} \mid s,a) h(s^{'}),\ s \in S,\ a \in A(s).\end{gathered}$$

where $d_s > 0, s \in S$ are numbers such that $\sum_{s\in S} d_s = 1$. The dual of the above linear programming problem is given as follows:

DLP 1.2

$$\max \textstyle\sum_{s=1}^{z}\sum_{a\in A(s)} r(s,a)x_{sa}$$

with respect to the constraints:

$$\begin{gathered}
(\mathrm{a})\textstyle\sum_{s=1}^{z}\sum_{a\in A(s)}(\delta(s,s^{'})-q(s^{'}\mid s,a))x_{sa}=0,\ s^{'}\in S\\
(\mathrm{b})\textstyle\sum_{a\in A(s)} x_{s^{'}a}+\sum_{s=1}^{z}\sum_{a\in A(s)}(\delta(s,s^{'})-q(s^{'}\mid s,a))y_{sa}=d_s\\
(\mathrm{c})x_{sa},y_{sa}\geq 0,\ s\in S,\ a\in A(s).
\end{gathered}$$

Theorem 1.3.6. *(Hordijk and Kallenberg [15]) Suppose* (x,y) *an optimal solution of the DLP* 1.2*. Then the stationary strategy* $f^{*}\in F^{s}$ *defined by*

$$f^{*}(s)=a_s\in A(s), such\ that \begin{cases} x_{sa_s}>0 & s\in\hat{S}=\{s\mid x_{sa}>0\}\\ y_{sa_s}>0 & s\notin\hat{S}\end{cases}$$

is limiting average optimal. If (g,h) *is an optimal solution of LP* 1.2*, the* g *is called the value vector of the undiscounted MDP.*

1.4 Semi-Markov Games

The notion of stochastic games can be generalised in a setup, where the transition time between successive states are dependent on the current state, next state, action chosen on current state and an arbitrary probability distribution. The game ceases to be strictly Markovian, but it retains enough of it's Markovian property to be called a 'Semi-Markov game'. Such games have been widely studied in literature (e.g. Lal-Sinha [22], Luque-Vasquez [27], Vega-Omaya [46]). In simple words, semi-Markov game is an extension of semi-Markov decision process(SMDP)(which is an one player game) to more than one player. So, some basic ideas of an SMDP can be extended in a straightforward way to semi-Markov games. However, the essential points have an independent development. For instance, to show the existence of optimal strategies for the players we must analyse the minimax and maximin equations and ensure the interchange of minimum and maximum in the corresponding optimality equation. These games have been studied under discounted as well as undiscounted (limiting ratio average) pay-offs in the literature.

Definition 1.10. *A zero-sum two-person finite SMG is described by a collection of objects $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, P, r >$, where $S = \{1, 2, \cdots, z\}$ is the finite non-empty state space and $A(s) = \{1, 2, \cdots, m_s\}, B(s) = \{1, 2, \cdots, n_s\}$ are respectively the non-empty sets of admissible actions of the players I and II respectively in the state s. Let us denote $K = \{(s, i, j) : s \in S, i \in A(s), j \in B(s)\}$ to be the set of admissible triplets. For each $(s, i, j) \in K$, we denote $q(. \mid s, i, j)$ to be the transition law of the game. Given $(s, i, j) \in K$ and $s^{'} \in S$, let $\tau_{ij}^{ss^{'}}$ be the transition time random variable which denotes the time for a transition to a state $s^{'}$ from a state s by a pair of actions $(i, j) \in A(s) \times B(s)$. Let $P_{ij}^{ss^{'}} = Prob(\tau_{ij}^{ss^{'}} \leq t)$ for each $(s, i, j) \in K, s^{'} \in S$ be a probability distribution function on $[0, \infty)$ and it is called the conditional transition time distribution function. Finally r is the real valued functions on K, which represents the immediate (expected) rewards for the player-I (and $-r$ is the immediate reward for player-II).*

Let us consider player I as the maximiser and player II as the minimiser in the zero-sum two person SMG. The semi-Markov game over infinite time is played as follows. At the 1st decision epoch, the game starts at $s \in S$ and the players I and II simultaneously and independently choose actions $i \in A(s)$ and $j \in B(s)$ respectively. Consequently player I and II get immediate rewards $r(s, i, j)$ and $-r(s, i, j)$ respectively and the game moves to the state $s^{'}$ with probability $q(s^{'} \mid s, i, j)$. The sojourn time to move from state s to the state $s^{'}$ is determined by the distribution function $P_{ij}^{ss^{'}}(.)$. After reaching the state $s^{'}$ on the next decision epoch, the game is repeated over infinite time with the state s replaced by $s^{'}$.

By a strategy (behavioural) π_1 of the player I, we mean a sequence $\{(\pi_1)_n(. \mid hist_n)\}_{n=1}^{\infty}$, where $(\pi_1)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n$=$(s_1, a_1, b_1, s_2, a_2, b_2 \cdots, s_{n-1}, a_{n-1}, b_{n-1}, s_n)$ for $n \geq 1$, $hist_1 = (s_1)$ and $(s_k, a_k, j_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi_1)_n(. \mid hist_n)$ on $A(s_n)$. Behavioural strategy π_2 for player II can be defined analogously. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π_1 and Π_2 to be the sets of strategies (behavioural) of the players I and II respectively.

A strategy $f = \{f_n\}_{n=1}^{\infty}$ for the player I is called semi-Markov if for each n, f_n depends on s_1, s_n and the decision epoch number n. Similarly we can define a semi-Markov strategy $g = \{g_n\}_{n=1}^{\infty}$ for the player II.

A stationary strategy is a strategy that depends only on the current state and not on the decision epoch number. A stationary strategy for player I is defined as z tuple $f = (f(1), f(2), \cdots, f(z))$, where each $f(s)$ is the probability distribution on $A(s)$ given by $f(s) = (f(s,1), f(s,2), \cdots, f(s, m_s))$, where $f(s,i)$ denotes the probability of choosing action i in the state s. By similar manner, one can define a stationary strategy g for player II as $g = (g(1), g(2), \cdots, g(z))$ where each $g(s)$ is the probability distribution on $B(s)$. Let us denote F_1^s and F_2^s to be the set of stationary strategies for player I and II respectively.

A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F_1^{sp} and F_2^{sp} to be the set of pure stationary strategies of the players I and II respectively.

A semi-stationary strategy is a semi-Markov strategy which is independent of the decision epoch n, i.e., for a initial state s_1 and present state s_n, if a semi-Markov strategy $g(s_1, s_n, n)$ turns out to be independent of n, then we call it a semi-stationary strategy. Let ξ_1^s and ξ_2^s denote the set of semi-stationary strategies for the players I and II respectively and ξ_1^{sp} and ξ_2^{sp} denote the set of pure semi-stationary strategies for the players I and II respectively.

For a general semi-Markov game model, we need some regularity condition i.e., infinite number of transitions do not occur in finite time interval. Also, Luque-Vasquez [27] and Lal-Sinha [22] studied some continuity and boundedness assumptions. If K is endowed with discrete topology in a finite semi-Markov game, then all the previous assumptions are satisfied trivially. Thus, we have the following assumptions:

1. There exists a finite number M such that $\mid r(s,i,j) \mid \leq M$, for all $(s,i,j) \in K$.

2. For all $s \in S$, $A(s)$ and $B(s)$ are compact subsets of the action spaces A and B respectively.

3. For all $s, s^{'} \in S$ and t, the reward function $r(s,i,j), P_{ss^{'}}^{ij}(t)$ and $q(s^{'} \mid s,i,j)$ are jointly continuous on $A(s) \times B(s)$.

4. There exists $\epsilon > 0$ and $T \in \mathbb{N}$ such that:

$$\sum_{s^{'} \in S} q(s^{'} \mid s,i,j) \sum_{t=1}^{T} P_{ss^{'}}^{ij}(t) \leq 1 - \epsilon \ \forall (s,i,j) \in K.$$

The above inequality states that there exists a probability of at least ϵ that the transition time is greater than T. From Shapley [42] we can conclude that if all the transition times are identical, then the SMG reduces to a Markov (Stochastic) game.

1.4.1 β-Discounted Semi Markov Games

Given a pair of behavioural strategies $(\pi_1, \pi_2) \in (\Pi_1 \times \Pi_2)$, the β-discounted pay-off for player-I from player-II in a semi-Markov game is

$$v_\beta(s, \pi_1, \pi_2) = \mathbb{E}_{\pi_1,\pi_2}[\sum_{m=1}^{\infty} \beta^{(\tau_1+\tau_2+\cdots+\tau_m)} r(X_m, A_m, B_m) \mid X_1 = s] \tag{1.15}$$

Where $\tau_1 = 0$ and τ_k is the time between the $(k-1)$th and kth decision epoch $(k \geq 2)$. Note that τ_k is a random variable which depends on $(X_{k-1}, A_{k-1}, B_{k-1}$ and $X_k)$ for each k.

For a pair of stationary strategies $(f_1, f_2) \in F_1^s \times F_2^s$, we can write the discounted pay-off for player-I as:

$$\begin{aligned} v_\beta(s, f_1, f_2) = r(s, f_1, f_2) + \sum_{s' \in S} \sum_{a \in A(s)} \sum_{b \in B(s)} [q(s^{'} \mid s, a, b) v_\beta(s^{'}, f_1, f_2) \\ \times \sum_{t=1}^{T} \beta^t P_{ss^{'}}^{ab}(t)] f_1(s, a) f_2(s, b), \text{for all} s \in S. \end{aligned} \tag{1.16}$$

Where $r(s, f_1, f_2) = \sum_{a \in A(s)} \sum_{b \in B(s)} r(s, a, b) f_1(s, a) f_2(s, b)$. Let us denote $v_\beta(f_1, f_2) = [v_\beta(s, f_1, f_2)]_{z \times 1}$, $r(f_1, f_2) = [r(s, f_1, f_2)]_{z \times 1}$ and

$$M_\beta(f_1, f_2) = [\textstyle\sum_{a \in A(s)} \sum_{b \in B(s)} q(s^{'} \mid s, a, b) \sum_{t=1}^{T} \beta^t P_{ss^{'}}^{ab}(t) f_1(s, a) f_2(s, b)]_{s, s^{'}=1}^{z}$$

Then (1.16) can be expressed as a matrix equation:

$$v_\beta(f_1, f_2) = r(f_1, f_2) + M_\beta(f_1, f_2) v_\beta(f_1, f_2) \tag{1.17}$$

For a pair of stationary strategies, the total discounted pay-off for the player-I can be calculated by the following result:

Theorem 1.4.1. *Let $(f_1, f_2) \in F_1^s \times F_2^s$ be a pair of stationary strategies of the players-I and II respectively. The the value vector $v_\beta(f_1, f_2)$ is the unique solution of:*

$$X = r(f_1, f_2) + M_\beta(f_1, f_2) X, \tag{1.18}$$

such that

$$v_\beta(f_1, f_2) = [I - M_\beta(f_1, f_2)]^{-1} r(f_1, f_2). \tag{1.19}$$

Proof. The sum of the s-th row of the matrix $M_\beta(f_1, f_2)$ can be written as:

$$\sum_{a\in A(s)}\sum_{b\in B(s)} q(s^{'} \mid s,a,b)\sum_{t=1}^{T}\beta^t P_{ss^{'}}^{ab}(t) f_1(s,a) f_2(s,b) = E_{f_1 f_2}[\beta^{\tau_{ss^{'}}^{ab}}] < 1$$

Thus we can conclude that the matrix $M_\beta(f_1, f_2)$ is a sub stochastic matrix whose sum of all row-elements is strictly less than 1. Now following similar technique as MDP model, we conclude that $(I - M_\beta(f_1, f_2))$ is non-singular and

$$(I - M_\beta(f_1, f_2)) = \textstyle\sum_{p=0}^{\infty} M_\beta^p(f_1, f_2)$$

Thus we can conclude that $X = [I - M_\beta(f_1, f_2)]^{-1} r(f_1, f_2)$ is the unique solution of (1.18). □

Now we define value vector and the optimal strategy pair of the players in a β-discounted zero-sum two person semi-Markov game.

Definition 1.11. *A zero-sum two person β-discounted semi-Markov game has a value $v_\beta(s)$ for the initial state $s \in S$ if*

$$\inf\nolimits_{\pi_2\in\Pi_2}\sup\nolimits_{\pi_1\in\Pi_1} v_\beta(s, \pi_1, \pi_2) = v_\beta(s) = \sup\nolimits_{\pi_1\in\Pi_1}\inf\nolimits_{\pi_2\in\Pi_2} v_\beta(s, \pi_1, \pi_2) \textit{ for all } s \in S.$$

A pair of strategies (π_1^*, π_2^*) is called an optimal strategy pair of the players if

$$v_\beta(s, \pi_1^*, \pi_2) \geq v_\beta(s, \pi_1^*, \pi_2^*) \geq v_\beta(s, \pi_1, \pi_2^*) \textit{ for all } s \in S \textit{ and } (\pi_1, \pi_2) \in \Pi_1 \times \Pi_2.$$

Luque-Vasquez[27] and Lal-Sinha [22] showed that under some boundedness and continuity assumptions, a discounted zero-sum two person semi-Markov game has a value and a pair of pure stationary optimal strategies for countable state space, compact action space and continuous sojourn time. The Shapley equation for such games has been obtained by them also. Thus we have the following result for a discounted zero-sum two person semi-Markov game:

Theorem 1.4.2. *The value vector v_β of zero-sum two person semi-Markov game satisfies the Shapley equation*

$$v_\beta(s) = val[r(s,a,b) + \sum_{s^{'}\in S} q(s^{'} \mid s,a,b) v_\beta(s^{'})\sum_{t=1}^{T}\beta^t P_{ss^{'}}^{ab}(t)] \forall s \in S \tag{1.20}$$

Furthermore (f_1^*, f_2^*) *is an optimal pure stationary strategy pair if and only if* $(f_1^*(s), f_2^*(s))$ *is a pair of pure optimal stationary strategy pair of the matrix game in the right hand side of the above equation* (1.20) *for all* $s \in S$.

1.4.2 Zero-Sum Two-Person Semi-Markov Games Under Limiting Ratio Average (Undiscounted) Pay-off

Let $(X_1, A_1, B_1, X_2, A_2, B_2 \cdots)$ be a co-ordinate sequence in $S \times (A \times B \times S)^\infty$. Given behavioural strategy pair $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi_1 \pi_2}(. \mid X_1 = s)$ (hence an expectation $\mathbb{E}_{\pi_1 \pi_2}(. \mid X_1 = s)$) on the product σ- field of $S \times (A \times B \times S)^\infty$ by Kolmogorov's extension theorem. For a pair of strategies $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$ for the players I and II respectively, the limiting ratio average (undiscounted) pay-off for player I, starting from a state $s \in S$ is defined by:

$$\phi(s, \pi_1, \pi_2) = \liminf_{n \to \infty} \frac{\mathbb{E}_{\pi_1 \pi_2} \sum_{m=1}^{n} [r(X_m, A_m, B_m) | X_1 = s]}{\mathbb{E}_{\pi_1 \pi_2} \sum_{m=1}^{n} [\bar{\tau}(X_m, A_m, B_m) | X_1 = s]}.$$

Here $\bar{\tau}(s, i, j) = \sum_{s' \in S} q(s^{'} \mid s, i, j) \sum_{t=1}^{T} t P_{ij}^{ss^{'}}(t)$ is the expected sojourn time in the state s for a pair of actions $(i, j) \in A(s) \times B(s)$. For a finite semi-Markov game, the transition times are finite. Furthermore, they are positive and can never be zero again by the fact that the transition times are $t = 1, 2, \cdots, T$. Thus we have the following observation:

Observation: There exists an $\epsilon > 0$ and a finite number $M (> \epsilon)$ such that:

$$\epsilon \leq \bar{\tau}(s, i, j) \leq M \text{ for all } (s, i, j) \in K.$$

Definition 1.12. *For each pair of stationary strategies* $(f, g) \in F_1^s \times F_2^s$ *we define the transition probability matrix as* $Q(f, g) = [q(s^{'} \mid s, f, g)]_{z \times z}$, *where* $q(s^{'} \mid s, f, g) = \sum_{i \in A(s)} \sum_{j \in B(s)} q(s^{'} \mid s, i, j) f(s, i) g(s, j)$ *is the probability that starting from the state s, next state is* $s^{'}$ *when the players choose strategies* f *and* g *respectively (For a stationary strategy* f, $f(s, i)$ *denotes the probability of choosing action* i *in the state* s*).*

For any pair of stationary strategies $(f, g) \in F_1^s \times F_2^s$ of player I and II, we write the undiscounted pay-off for player I as:

$$\phi(s, f, g) = \lim_{n \to \infty} \frac{\sum_{m=1}^{n} r^m(s, f, g)}{\sum_{m=1}^{n} \bar{\tau}^m(s, f, g)} \text{ for all } s \in S.$$

Where $r^m(s, f, g)$ and $\bar{\tau}^m(s, f, g)$ are respectively the expected reward and expected sojourn time for player I at the m th decision epoch, when player I chooses f and player

II chooses g respectively and the initial state is s. We define $r(f,g) = [r(s,f,g)]_{z\times 1}$, $\bar{\tau}(f,g) = [\bar{\tau}(s,f,g)]_{z\times 1}$ and $\phi(f,g) = [\phi(s,f,g)]_{z\times 1}$ as expected reward, expected sojourn time and undiscounted pay-off vector for a pair of stationary strategy $(f,g) \in F_1^s \times F_2^s$. Now

$$\begin{aligned} r^m(s,f,g) \quad &= \textstyle\sum_{s'\in S} \mathbb{P}_{fg}(X_m = s' \mid X_1 = s) r(s',f,g) \\ &= \textstyle\sum_{s'\in S} r(s',f,g) q^{m-1}(s' \mid s,f,g) \\ &= [Q^{m-1}(f,g) r(f,g)](s) \end{aligned}$$

and

$$\begin{aligned} \bar{\tau}^m(s,f,g) \quad &= \textstyle\sum_{s'\in S} \mathbb{P}_{fg}(X_m = s' \mid X_1 = s) \bar{\tau}(s',f,g) \\ &= \textstyle\sum_{s'\in S} \bar{\tau}(s',f,g) q^{m-1}(s' \mid s,f,g) \\ &= [Q^{m-1}(f,g) \bar{\tau}(f,g)](s) \end{aligned}$$

Since $Q(f,g)$ is a Markov matrix, we have by Kemeny et al., [21]

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{m=1}^{n} Q^m(f,g) \text{ exists and equals to } Q^*(f,g).$$

It is obvious that

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{m=1}^{n} r^m(f,g) = [Q^*(f,g) r(f,g)](s)$$

and

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{m=1}^{n} \bar{\tau}^m(f,g) = [Q^*(f,g) \bar{\tau}(f,g)](s).$$

Thus we have for any pair of stationary strategies $(f_1, f_2) \in F_1^s \times F_2^s$,

$$\phi(s,f,g) = \tfrac{[Q^*(f,g) r(f,g)](s)}{[Q^*(f,g) \bar{\tau}(f,g)](s)} \text{ for all } s \in S$$

where $Q^*(f,g)$ is the Cesaro limiting matrix of $Q(f,g)$.

Definition 1.13. *A zero-sum two person undiscounted semi-Markov game is said to have a value vector $\phi = [\phi(s)]_{z\times 1}$ if*

$$\sup_{\pi_1\in\Pi_1} \inf_{\pi_2\in\Pi_2} \phi(s,\pi_1,\pi_2) = \phi(s) = \inf_{\pi_2\in\Pi_2} \sup_{\pi_1\in\Pi_1} \phi(s,\pi_1,\pi_2) \textit{ for all } s \in S.$$

A pair of strategies $(\pi_1^*, \pi_2^*) \in \Pi_1, \times\Pi_2$ is said to be an optimal strategy pair for the players if $\phi(s,\pi_1^*,\pi_2) \geq \phi(s) \geq \phi(s,\pi_1,\pi_2^*)$ for all $s \in S$ and all $(\pi_1,\pi_2) \in \Pi_1 \times \Pi_2$. Throughout this paper, we use the notion of undiscounted pay-off as limiting ratio average pay-off.

For the undiscounted case, the following result by Lal-Sinha [22] is very significant.

Theorem 1.4.3. *If there exists a bounded function $h^* \in \mathcal{B}(S)$ and a constant $g^* \in \mathbb{R}$ satisfying the equation:*

$$h^*(s) = val\{r(s,i,j) + \textstyle\sum_{s' \in S} q(s^{'} \mid s,i,j)h^*(s^{'}) - g^*\bar{\tau}(s,i,j)\} \text{ for all } s \in S$$

where $\mathcal{B}(S)$ is the set of all bounded measurable functions on S, then the undiscounted semi-Markov game has a value g^* in every state and both players have stationary optimal strategies.

1.5 Semi-Markov Decision Processes

Semi-Markov decision processes(SMDPs) were introduced by Jewell [19] and Howard [16] and later generalised by Ross [40], Lippman [25], Federgruen and Tijms et al.[9]. It is a generalisation of Markov decision processes(MDPs) in a larger class of dynamical models. SMDPs are used in modelling stochastic control problems arising in MDPs where the sojourn(transition) time is a random variable depending on the present state, the next state and the action chosen in the present state. A finite SMDP is similar to an MDP, where the state changing occurs according to Markov property. These models are useful in the field of inventory management, production engineering, reliability testing and queuing theory.

Definition 1.14. *A finite semi-Markov decision process is defined by the pen-tuple $< S, A = \{A(s) : s \in S\}, q, P, r >$, where the state space $S = \{1, 2, \cdots, z\}$ is a non-empty finite set and A is the finite action set for the decision maker. A subset $A(s) \subseteq A$ is defined as the set of admissible actions for the state s of the decision maker. The set of admissible state-action pairs is defined by $K = \{(s,a) : a \in A(s), s \in S\} \subseteq S \times A$. For each $x \in K$, $q(. \mid x)$ is the transition law. Given $(s,a) \in K$ and $s^{'} \in S$, let $\tau^a_{ss^{'}}$ be the sojourn(transition)time random variable(discrete) denoting the time from state s to state $s^{'}$ by an action $a \in A(s)$. For each $(s,a) \in K$, $s^{'} \in S$, $P^a_{ss^{'}}(t) = Prob(\tau^a_{ss^{'}} \leq t)$ is a probability distribution function on $(0,\infty)$ given $K \times S$ and it is called the conditional transition time distribution function. r is a real valued function on K (i.e., r: $K \to \mathbb{R}$ is the reward function) which is called the immediate reward function.*

The process, i.e., SMDP proceeds over infinite time as follows:
At the first decision epoch, the process starts at a state $s \in S$. The decision maker looks at s and chooses an action $a \in A(s)$. Then he/she gets an immediate reward

$r(s,a)$. At next decision epoch the system moves to a new state $s^{'} \in S$ with probability $q(s^{'} \mid s,a)$. $P^{a}_{ss^{'}}(.)$, which denotes the transition time distribution function, determines the sojourn time from the state s to state $s^{'}$, when the decision maker chooses action a. After the transition to the state $s^{'}$ on the next decision epoch, the above described process with s replaced by $s^{'}$, is repeated all over again, and thus SMDP proceeds over infinite time period. Clearly, for identical transition times, an SMDP reduces to an MDP.

Definition 1.15. *By a strategy (behavioural) π of the decision maker, we mean a sequence $\{(\pi_n)(. \mid hist_n)\}_{n=1}^{\infty}$, where (π_n) specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n$=$(s_1,a_1,s_2,a_2,\cdots,s_{n-1},a_{n-1},s_n)$ for $n \geq 1$, $hist_1 = (s_1)$ and $(s_k,a_k) \in K$ are respectively the state and actions of the decision maker at the k-th decision epoch) a probability distribution $(\pi)_n(. \mid hist_n)$ on $A(s_n)$. We denote Π to be the sets of strategies (behavioural) of the decision maker.*
A strategy $f = \{f_n\}_{n=1}^{\infty}$ for the decision maker is called semi-Markov if for each n, f_n depends on s_1, s_n and the decision epoch number n.
A strategy $f = \{f_n\}_{n=1}^{\infty}$ is called stationary if $\exists$ a map $\xi : S \to \mathbb{P}(A)$, such that (i) $f_n = \xi$ $\forall n \in \mathbb{N}$, i.e., $f_n(. \mid s_1, s_2, \cdots, s_n = s) = \xi(s)$ and (ii) $\xi(s)$ has a finite 'support' in $A(s)$, i.e., $\xi(s) \in \mathbb{P}(A(s))$ for each $s \in S$. Let us denote F^s as the set of stationary strategies of the decision maker.
A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F^{sp} to be the set of pure stationary strategies of the decision maker.
A strategy $g = \{g_n\}_{n=1}^{\infty}$ is called semi-stationary strategy if there exists a map $\xi : S \times S \to \mathbb{P}(A)$ such that (i) $g_n(. \mid s_1 = s, \cdots, s_n = s^{'}) = \xi(s,s^{'})$ for all $s_1, s_2, \cdots, s_n \in S$, $n \in \mathbb{N}$ and (ii) $\xi(s,s^{'})$ has a support in $A(s^{'})$ for each pair $(s,s^{'}) \in S \times S$. So, it can also be viewed as a semi-Markov strategy which is independent of the decision epoch n, i.e., for a initial state s_1 and present state s_2, if a semi-Markov strategy $g(s_1,s_2,n)$ turns out to be independent of n, then we call it a semi-stationary strategy. Let ξ^s denote the set of semi-stationary strategies and ξ^{sp} denote the set of pure semi-stationary strategies of the decision maker.

We consider a sequence of co-ordinates $(X_1,A_1,X_2,A_2,\cdots)$ in $S \times (A \times S)^{\infty}$. For a specified probability distribution η on $(S,2^S)$ and a strategy $\pi \in \Pi$, we can conclude from Kolmogorov's extension theorem that there exists an unique probability measure on the Borel-σ field of $S \times (A \times S)^{\infty}$, such that the marginal distribution of X_1 is η and for each $n \in \mathbb{N}$ and almost surely $A_n \in A(X_n)$. For a degenerate probability distribution

η at $s \in S$, we denote $\mathbb{P}_\pi(. \mid X_1 = s)$ to be the probability measure determined by η and π and the corresponding expectation operator is $\mathbb{E}_\pi(. \mid X_1 = s)$. For every strategy $\pi \in \Pi$ we have a sequence of random rewards $r(X_m, A_m)$ which denotes the immediate reward on the m-th decision epoch ($m = 1, 2, 3, \cdots$), where X_m and A_m are respectively the state and action chosen on that decision epoch. The expectation of $r(X_m, A_m)$ is well defined and will be denoted by:

$$\mathbb{E}_\pi[r(X_m, A_m) \mid X_1 = s].$$

1.5.1 β-Discounted Semi-Markov Decision Processes

We consider SMDP model where rewards are discounted by a discount factor $\beta \in [0, 1)$. It means that a reward 'r' at time 't' is equivalent to a reward $r\beta^t$ at $t = 0$, i.e., the initial time. For each $\beta \in [0, 1)$ and a fixed initial state $s \in S$, the total discounted pay-off for a strategy $\pi \in \Pi$ is defined by:

$$v_\beta(s, \pi) = \mathbb{E}_\pi[\sum_{m=1}^{\infty} \beta^{(\tau_1+\tau_2+\cdots+\tau_m)} r(X_m, A_m) \mid X_1 = s] \tag{1.21}$$

where $\tau_1 = 0$ and $\tau_k (k \geq 1)$ is the time between $(k-1)$th and k-th decision epochs. τ_k is a random variable depending on X_{k-1}, A_{k-1} and X_k, where $X_k \in S$ and $A_k \in A(X_k)$ are respectively the state and action at the k-th decision epoch.

Definition 1.16. *A strategy π^* is called and optimal strategy for the discounted SMDP if*

$$v_\beta(s, \pi^*) \geq v_\beta(s, \pi) \textit{ for all } s \in S, \pi \in \Pi.$$

Let us denote $v_\beta(s)$= $sup_{\pi\in\Pi} v_\beta(s, \pi) \forall s \in S$. Now for a stationary strategy $f \in F^s$, initial state $s \in S$ and reward function r, we have from (1.14)

$$\begin{aligned}
v_\beta(s, f) &= \mathbb{E}_f[\sum_{m=1}^{\infty} \beta^{(\tau_1+\tau_2+\cdots+\tau_m)} r(X_m, A_m) \mid X_1 = s] \\
&= \mathbb{E}_f[r(X_1, A_1)] + \mathbb{E}[\mathbb{E}_f\{\sum_{m=1}^{\infty} \beta^{(\tau_2+\cdots+\tau_m)} r(X_m, Am) \mid hist_2, \tau_2\}] \\
&= r(s, f) + \mathbb{E}[\beta^{\tau_2} \mathbb{E}_f\{\sum_{m=3}^{\infty} \beta^{(\tau_3+\cdots+\tau_m)} r(X_m, A_m) \mid hist_2, \tau_2\}] \\
&= r(s, f) + \mathbb{E}[\beta^{\tau_2} v_\beta(f, X_2)] \ \forall s \in S \text{ being an initial state.}
\end{aligned} \tag{1.22}$$

Which implies

$$v_\beta(s,f) = r(s,f) + \sum_{a\in A(s)} \left[\sum_{s'\in S} q(s^{'} \mid s,a) v_\beta(s^{'},f) \int_0^\infty \beta^t dP^a_{ss'}(t) \right] f(s,a) \tag{1.23}$$

where $f(s,a)$ is the probability of choosing action 'a' when system is in state 's'. Now putting equation (1.23) in matrix form we get

$$v_\beta(f) = r(f) + P_\beta(f) v_\beta(f), \tag{1.24}$$

where $v_\beta(f) = \left[v_\beta(s,f)\right]_{z\times 1}$ is the value vector, $r(f) = [r(s,f)]_{z\times 1}$ and

$$P_\beta(f) = \left[\textstyle\sum_{a\in A(s)} q(s^{'} \mid s,a) \int_0^\infty \beta^t dP^a_{ss'}(t) f(s,a)\right]^z_{s,s'=1}.$$

$r(s,f)$ can be calculated as $\sum_{a\in A(s)} r(s,a)f(s,a)$. Here $P_\beta(f)$ is a sub-stochastic matrix whose all rows have a sum strictly less than 1. Using this fact and by induction, we can further write (1.24) as

$$v_\beta(f) = (I - P_\beta(f))^{-1} r(f). \tag{1.25}$$

Theorem 1.5.1. *(Howard [16], Ross [40]) The value vector of a discounted SMDP with discount factor β is the unique solution of the optimality equation*

$$v_\beta(s) = max_{a\in A(s)}\{r(s,a) + \sum_{s'\in S} q(s^{'} \mid s,a) v_\beta(s^{'}) \int_0^\infty \beta^t dP^a_{ss'}(t)\},\ s\in S. \tag{1.26}$$

Further a strategy $f^* \in F^s$ is an optimal stationary strategy if and only if $f^*(s)$ is an optimal action of (1.26).

A β-discounted SMDP can be solved via linear programming problem (Puterman [38]).

1.5.2 Limiting Ratio Average Pay-Off Criterion in SMDP

For a behavioural strategy $\pi \in \Pi$, the limiting ratio average pay-off for the decision maker can be defined as:

$$\phi_1(s,\pi) = \liminf_{n\to\infty} \frac{\mathbb{E}_\pi \sum_{m=1}^n [r(X_m,A_m)|X_1=s]}{\mathbb{E}_\pi \sum_{m=1}^n [\bar{\tau}(X_m,A_m)|X_1=s]} \text{ for all } s\in S.$$

and

$$\phi_2(s,\pi)= \limsup_{n\to\infty} \frac{\mathbb{E}_\pi \sum_{m=1}^{n}[r(X_m,A_m)|X_1=s]}{\mathbb{E}_\pi \sum_{m=1}^{n}[\bar{\tau}(X_m,A_m)|X_1=s]} \text{ for all } s \in S.$$

Where $\bar{\tau}(s,a) = \sum_{s^{'}\in S} q(s^{'} \mid s,a) \int_0^\infty t dP_{ss^{'}}(t \mid a)$ is the expected sojourn time in the state s when decision maker chooses the action $a \in A(s)$. We assume that $\bar{\tau}(s,a)$ is bounded away from zero for each $(s,a) \in K$.

Definition 1.17. *A strategy π^* is called an optimal strategy under limiting ratio average pay-off if*

$$\phi_1(s,\pi^*) \geq \phi_1(s,\pi) \ (\phi_2(s,\pi^*) \geq \phi_2(s,\pi))$$

Let $\phi_1(s) = \sup_{s\in S}(\phi_1(s,\pi))$ and $\phi_2(s) = \sup_{s\in S}(\phi_2(s,\pi))$. Then $\phi_1(s)$ (resp. $\phi_2(s)$) is called the limiting ratio average value of the undiscounted SMDP for the initial state $s \in S$. For a stationary strategy $f \in F^s$ and $s \in S$, let $r(s,f) = \sum_{a\in A(s)} r(s,a).f(s,a)$, $\tau(s,f) = \sum_{a\in A(s)} \tau(s,a)f(s,a)$ and $q(s^{'} \mid s,f) = \sum_{a\in A(s)} q(s^{'} \mid s,a)f(s,a)$. Then the reward vector, sojourn time vector and transition probability matrix can be defined respectively as: $r(f) = [r(s,f)]_{z\times 1}$, $\tau(f) = [\tau(s,f)]_{z\times 1}$ and $Q(f) = [q(s^{'} \mid s,f)]_{z\times z}$. Let $Q^m(f) = [q^m(s^{'} \mid s,f)]_{z\times z}$ where $q^m(s^{'} \mid s,f)$ is the m-step transition probability from state s to $s^{'}$ under the stationary strategy f.

Lemma 1.5.2. *(Doob, theorem* 2.1*, page-* 175*) [6] Let $Q = [q(s^{'} \mid s)]_{z\times z}$ be a transition probability matrix. Then $\exists$ a stochastic matrix $Q^* = [q(s^{'} \mid s)]_{z\times z}$, which is called the Cesaro-limiting matrix of Q, is defined as:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} q^m(s^{'} \mid s) = q^*(s^{'} \mid s) \ \ s,s^{'} \in S$$

Now by Sinha et.al [43] we have the following result to compute the undiscounted (limiting ratio average) pay-off of the decision maker.

Proposition 1.5.3. *Let $f \in F^s$ be a stationary strategy, then*

$$\phi_1(s,f) = \phi_2(s,f) = \frac{[Q^*(f)r(f)](s)}{[Q^*(f)\bar{\tau}(f)](s)} \text{ for all } s \in S.$$

Where $Q^*(f)$ is the Cesaro limiting matrix of $Q(f)$.

Previously Xiaobo et.al [20] showed that optimal pure stationary strategies may not exist in the undiscounted (limiting ratio average) SMDP model. The following example illustrates the fact:

Example 1.1. *State*-1: $\begin{array}{|r|}\hline 13\\ (1,0,0,0,0)\\ 4\\ \hline 9\\ (0,0,0,1,0)\\ 2\\ \hline\end{array}$ *State*-2: $\begin{array}{|r|}\hline 4\\ (0,1,0,0,0)\\ 2\\ \hline 3\\ (0,1,0,0,0)\\ 1.6\\ \hline\end{array}$ *State*-3: $\begin{array}{|r|}\hline 7\\ (\frac{1}{3},\frac{2}{3},0,0,0)\\ 2\\ \hline 3\\ (0,0,0,0,1)\\ 1.5\\ \hline\end{array}$

State-4: $\begin{array}{|r|}\hline 15\\ (1,0,0,0,0)\\ 5\\ \hline\end{array}$ *State*-5: $\begin{array}{|r|}\hline 5\\ (0,0,1,0,0)\\ 3\\ \hline\end{array}$

Where a cell $\begin{array}{|r|}\hline r\\ (q_1,q_2,q_3,q_4,q_5)\\ \bar{\tau}\\ \hline\end{array}$ represents that r is the immediate rewards of the playes, q_1, q_2, q_3, q_4, q_5 represents that the next states are 1, 2, 3, 4 and 5 respectively and $\bar{\tau}$ is the expected sojourn time if this cell is chosen at present. There are 8 pure stationary strategies in this SMDP model, which are given by $f_1=(1,1,1,1,1)$, $f_2=(1,2,1,1,1)$, $f_3=(1,1,2,1,1)$, $f_4=(1,2,2,1,1)$, $f_5=(2,1,1,1,1)$, $f_6=(2,2,1,1,1)$, $f_7=(2,1,2,1,1)$ and $f_8=(2,2,2,1,1)$. One can verify that:

$$Q^*(f_1)=Q^*(f_2)\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\\frac{1}{3}&\frac{2}{3}&0&0&0\\1&0&0&0&0\\\frac{1}{3}&\frac{2}{3}&0&0&0\end{bmatrix},\ Q^*(f_3)=Q^*(f_4)=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&\frac{1}{2}&0&\frac{1}{2}\\1&0&0&0&0\\0&0&\frac{1}{2}&0&\frac{1}{2}\end{bmatrix},$$

$$Q^*(f_5)=Q^*(f_6)=\begin{bmatrix}\frac{1}{2}&0&0&\frac{1}{2}&0\\0&1&0&0&0\\\frac{1}{6}&\frac{2}{3}&0&\frac{1}{6}&0\\\frac{1}{2}&0&0&\frac{1}{2}&0\\\frac{1}{6}&\frac{2}{3}&0&\frac{1}{6}&0\end{bmatrix},\ Q^*(f_7)=Q^*(f_8)=\begin{bmatrix}\frac{1}{2}&0&0&\frac{1}{2}&0\\0&1&0&0&0\\0&0&\frac{1}{2}&0&\frac{1}{2}\\1&0&0&0&0\\0&0&\frac{1}{2}&0&\frac{1}{2}\end{bmatrix}$$

One can verify that $\phi_1(f_1)=(3.25,2,2.625,3.25,2.625)$, $\phi_1(f_2)=(3.25,1.875,2.639,3.25,2.639)$, $\phi_1(f_3)=(3.25,2,1.778,3.25,1.778)$, $\phi_1(f_4)=(3.25,1.875,1.778,3.25,1.778)$, $\phi_1(f_5)=(3.429,2,2.667,3.429,2.667)$, $\phi_1(f_6)=(3.429,1.875,2.687,3.429,2.687)$, $\phi_1(f_7)=(3.429,2,1.778,3.429,1.778)$ and $\phi_1(f_8)=(3.429,1.875,1.778,3.429,1.778)$. The value vector of this SMDP is $(3.429,2,2.687,3.429,2.687)$ but no pure stationary strategies exist in this example since $\phi_1(3)=2.687$ is attained by f_6 only, whereas $\phi_1(2)=2>1.875=\phi_1(2,f_6)$.

Now, we state a very important result from Sinha et al.[43] which states the existence of a pure semi-stationary strategy instead of stationary strategy in an undiscounted(limiting ratio average) SMDP.

Theorem 1.5.4. *(i)There exists pure semi-stationary strategies $g^{'}$ and $g^{''} \in \xi^{sp}$ such that:*

$$\phi_1(s, g^{'}) \geq \phi_1(s, \pi) \text{ and } \phi_2(s, g^{''}) \leq \phi_2(s, \pi) \text{ for all } \pi \in \Pi \text{ and all } s \in S.$$

(ii)There exists pure semi-stationary strategies g^{*} and $g^{**} \in \xi^{sp}$ such that:

$$\phi_1(s, g^{*}) \leq \phi_1(s, \pi) \text{ and } \phi_2(s, g^{''}) \geq \phi_2(s, \pi) \text{ for all } \pi \in \Pi \text{ and all } s \in S.$$

Note: The first statement of theorem 1.5.4 asserts the existence of pure optimal semi-stationary strategies for an SMDP under maximisation (with liminf objective) and minimisation (with limsup objective) problems respectively. The second statement asserts the existence of such optimal policies when the optimisation problems are reversed.

1.6 Results obtained in this thesis

The dissertation is divided into four chapters including introduction and literature review. In chapter 2 we deal with discounted/undiscounted (limiting ratio average) SMDP with vector pay-offs. In chapter 3 we consider undiscounted (limiting ratio average) perfect information semi-Markov /stochastic games and prove the existence of optimal pure semi-stationary/stationary strategy pair of the players. Chapter 4 presents the existence of a near-optimal pure semi-stationary strategy of the decision maker in an undiscounted (limiting ratio average) SMDP with countably infinite action space. The summary of the results obtained in each chapter are given below:

In the 2nd chapter semi-Markov decision processes with vector rewards are studied under discounted as well as undiscounted (limiting ratio average) pay-off criterion. The notion of optimality is replaced by Pareto-optimality here. We show that a pure stationary/ semi-stationary Pareto-optimal strategy exists for the discounted and undiscounted (limiting ratio average) SMDP. We also propose an algorithm for discounted SMDP with vector rewards to compute all the stationary Pareto-optimal strategies of the decision maker. For the undiscounted case, we use an existing algorithm by Mondal(2020)[33] to compute a pure semi-stationary Pareto-optimal strategy of the decision maker.

In chapter 3 we study zero-sum two-person perfect information semi-Markov games as well as stochastic games (PISMGs/ PISGs) under limiting ratio average (limiting average) payoff criteria and prove that these games have a value and both the maximiser and the minimiser have optimal pure semi-stationary (stationary) strategies. We arrive at the result by first fixing an arbitrary initial state and forming the matrix of undiscounted payoffs corresponding to each pair of pure stationary strategies of the two players and proving that this matrix has a pure saddle point (as optimality in the pure stationary/semi-stationary strategy space is equivalent to optimality in the behavioural strategy space, so it is necessary to only consider pure stationary/ semi-stationary strategies to prove the optimality of strategies of the players). In the case of PISG, we further use the results by Derman [5] to prove the existence of optimal pure stationary strategy pair of the players. A crude but finite step algorithm is given to compute such an optimal pure stationary/ semi-stationary strategy pair of the players in a PISG/PISMG.

Chapter 4 deals with undiscounted semi-Markov decision processes with countably infinite action spaces, where the state space is finite and the action spaces of the decision maker are (possibly) countably infinite. Here we do not put any restrictions on the rewards. We prove that in such model, the value of the decision process exists and the decision maker has a near-optimal pure semi-stationary strategies. We have shown this result by exploiting the results of Sinha et. al [44] as the decision process can be regarded as a one player game. We also propose an efficient algorithm to compute the value and a near-optimal pure semi-stationary strategy of the decision maker.

Chapter 2

Semi-Markov Decision Processes With Vector Pay-Offs

2.1 Introduction

Semi-Markov decision processes(SMDPs) were introduced by Jewell [19] and Howard [16] and later generalised by Ross [40], Lippman [25], Federgruen and Tijms et al.[9]. It is a generalisation of Markov decision processes(MDPs) in a larger class of dynamical models. SMDPs are used in modelling stochastic control problems arising in MDPs where the sojourn(transition) time is a random variable depending on the present state, the next state and the action chosen in the present state. A finite SMDP is similar to an MDP, where the state changing occurs according to Markov property. Pareto-optimality is generally studied in multi criterion optimisation theory. For details see [49]. Owen [37] showed the usefulness of Pareto-optimality in the field of co-operative game theory. Chatterjee et al.[3] showed that in a Markov decision process(discounted) with vector pay-offs, pure-stationary Pareto-optimal strategies exist. We generalise this result to both discounted and undiscounted (limiting ratio average) semi-Markov decision processes.

The results of this chapter are organised as follows. Section 2.1.1 contains preliminaries and basic definitions about SMDPs. Section 2.1.2 deals with discounted semi-Markov decision processes and definition of Pareto-optimal strategies of a discounted SMDP. In section 2.1.3, we show the existence of a pure stationary Pareto-optimal strategy in a discounted SMDP. Section 2.1.4 contains an algorithm to compute a stationary Pareto-optimal strategy of the decision maker. Section 2.1.5 contains an numerical example establishing our algorithm. Further in section 2.1.6 we concentrate on undiscounted (limiting ratio average pay-off) SMDP. In section 2.1.7 we show the existence of a pure semi-stationary Pareto-optimal strategy in an undiscounted SMDP. In section 2.1.8 we modify an algorithm proposed by Mondal [33] to calculate a pure semi-stationary Pareto-optimal strategy of the decision maker in an undiscounted SMDP. The chapter concludes with a numerical example based on the algorithm described in section 2.1.8.

2.1.1 Preliminaries

Definition 2.1. *A finite semi-Markov decision process is defined by the collection of five objects* $< S, A = \{A(s) : s \in S\}, q, P, r >$ *where the state space* $S = \{1, 2, \cdots, z\} (\neq \phi)$ *is a non-empty finite set and* A *is the finite action set for the decision maker. A subset* $A(s) \subseteq A$ *is defined as the set of admissible actions for the state* s *of the decision maker. The set of admissible state-action pairs is defined by* $\kappa_A = \{(s, a) : a \in A(s), s \in S\}$ $\subseteq S \times A$*. For each* $x \in \kappa_A$*,* $q(. \mid x)$ *is the transition law. Given* $(s, a) \in \kappa_A$ *and* $s^{'} \in S$*, let* $\tau^a_{ss^{'}}$ *be the sojourn(transition)time random variable(discrete) denoting the time from state* s *to state* $s^{'}$ *by an action* $a \in A(s)$*. For each* $(s, a) \in k_A$*,* $s^{'} \in S$*,* $P^a_{ss^{'}}(t) = Prob(\tau^a_{ss^{'}} \leq t)$ *is a probability distribution function on* $(0, \infty)$ *given* $\kappa_A \times S$ *and it is called the conditional transition time distribution function. Without loss of generality we write* P *instead of* $P^a_{ss^{'}}(t)$ *in some cases.* r *is a real valued function on* κ_A *(i.e.,* r*:* $\kappa_A \rightarrow \mathbb{R}$ *is the reward function) which is called the immediate reward function.*

At the first decision epoch, the process starts at a state $s_1 \in S$. The decision maker looks at s_1 and chooses an action $a_1 \in A(s_1)$. Then he / she gets an immediate reward $r(s_1, a_1)$. At next decision epoch the system moves to a new state $s_2 \in S$ with probability $q(s_2 \mid s_1, a_1)$. $P^{a_1}_{s_1 s_2}(.)$, which denotes the transition time distribution function, determines the sojourn time from the state s_1 to state s_2, when the decision maker chooses action a_1. After the transition to the state s_2 on the next decision epoch, the above described process with s_1 replaced by s_2, is repeated all over again, and thus SMDP proceeds over infinite time period. Clearly, for identical transition times, an SMDP reduces to an MDP.

Definition 2.2. *By a strategy (behavioural)* π *of the decision maker, we mean a sequence* $\{(\pi)_n(. \mid hist_n)\}_{n=1}^{\infty}$*, where* $(\pi)_n$ *specifies which action is to be chosen on the* n*-th decision epoch by associating with each history* $hist_n$ *of the system up to* n*th decision epoch (where* $hist_n$=$(s_1, a_1, s_2, a_2, \cdots, s_{n-1}, a_{n-1}, s_n)$ *for* $n \geq 2$*,* $hist_1 = (s_1)$ *and* s_k*,* a_k *are respectively the state and action of the decision maker at the* k*-th decision epoch) a probability distribution* $(\pi)_n(. \mid hist_n)$ *on* $A(s_n)$*. Generally by any unspecified strategy, we mean behavioural strategy here.*

A strategy $f=\{f_n\}_{n=1}^{\infty}$ is called stationary if $\exists$ a map $\xi : S \to \mathbb{P}(A)$, such that (i) $f_n = \xi \ \forall n \in \mathbb{N}$, i.e., $f_n(. \mid s_1, \cdots, s_n = s) = \xi(s)$ and (ii) $\xi(s)$ has a finite 'support' in $A(s)$, i.e., $\xi(s) \in \mathbb{P}(A(s))$ for each $s \in S$.

A strategy is called a policy, when for any $\pi = \{(\pi)_n\}_{n=1}^{\infty}$, each $(\pi)_n$ is degenerate probability distribution.

Let us denote Π and F^s as the respective classes of all behavioural and stationary strategies of the decision maker. We consider a sequence of co-ordinates $(X_1, A_1, X_2, A_2, \cdots)$ in $S \times (A \times S)^{\infty}$. For a specified probability distribution η on $(S, 2^S)$ and a strategy $\pi \in \Pi$, we can conclude from Kolmogorov's extension theorem that there exists an unique probability measure on the Borel-σ field of $S \times (A \times S)^{\infty}$, such that the marginal distribution of X_1 is η and for each $n \in \mathbb{N}$ and almost surely $A_n \in A(X_n)$. For a degenerate probability distribution η at $s \in S$, we denote $P_{\pi}(. \mid X_1 = s)$ to be the probability measure determined by η and π and the corresponding expectation operator is $\mathbb{E}_{\pi}(. \mid X_1 = s)$. For every strategy $\pi \in \Pi$ we have a sequence of random rewards $r(X_m, A_m)$ which denotes the immediate reward on the m-th decision epoch ($m = 1, 2, 3, \cdots$), where X_m and A_m are respectively the state and action chosen on that decision epoch. The expectation of $r(X_m, A_m)$ is well defined and will be denoted by

$$\mathbb{E}_{\pi}[r(X_m, A_m) \mid X_1 = s]. \tag{2.1}$$

2.1.2 Discounted Semi-Markov Decision Processes With Vector Pay-Offs

In this section we consider SMDP model where rewards are discounted by a discount factor $\beta \in [0,1)$. It means that a reward r at time t is equivalent to a reward $r\beta^t$ at $t = 0$, i.e., the initial time.

Definition 2.3. *For each $\beta \in [0,1)$ and a fixed initial state $s \in S$, the total discounted pay-off for a strategy $\pi \in \Pi$ is defined by*

$$V_{\beta}^{\pi}(r, s) = \mathbb{E}_{\pi}[\sum_{m=1}^{\infty} \beta^{(\tau_1 + \tau_2 + \cdots + \tau_m)} r(X_m, A_m) \mid X_1 = s] \tag{2.2}$$

where $\tau_1 = 0$ and $\tau_i (i \geq 2)$ is the time between $(i-1)$th and i-th decision epochs. τ_i is a random variable depending on X_{i-1}, A_{i-1} and X_i, where $X_i \in S$ and $A_i \in A(X_i)$ are respectively the state and action at the ith decision epoch(each X_i and A_i is a random variable $\forall i \geq 1$). For an SMDP with a fixed initial state s, strategy π and l *different reward functions $r_1,r_2,\cdots,r_l$, the discounted value profile(or pay-off profile) w.r.t. $\bar{r}=(r_1,r_2,\cdots,r_l)$ is defined by*

$$V_\beta^\pi(\bar{r},s)=(V_\beta^\pi(r_1,s),\cdots,V_\beta^\pi(r_l,s)). \tag{2.3}$$

We denote $\mathbb{P}(B)$ to be the set of probability distributions on the set B and $\mathbb{R}^n$ as the set of n tuples of real numbers, where $n\in\mathbb{N}$. For $x=(x_1,x_2,\cdots,x_n)$, $y=(y_1,y_2,\cdots,y_n)\in\mathbb{R}^n$, we assume that comparison(equality) between these two vectors is comparison(equality) between their respective co-ordinates, i.e., for any $x,y\in\mathbb{R}^n$, $x\geq y$ holds if $x_i\geq y_i$, $\forall i=1,2,\cdots,n$.

Definition 2.4. *For a discounted SMDP $\Gamma=<S,A,q,P,\bar{r}>$ with reward vector $\bar{r}=(r_1,r_2,\cdots,r_l)$ a strategy $\pi^*\in\Pi$ is called Pareto-optimal such that for all $s\in S$, if there is no strategy $\pi^{'}\in\Pi$ such that both $V_\beta^{\pi^*}(\bar{r},s)\leq V_\beta^{\pi^{'}}(\bar{r},s)$ and $V_\beta^{\pi^*}(\bar{r},s)\neq V_\beta^{\pi^{'}}(\bar{r},s)$ hold, i.e., $\nexists$ any strategy $\pi^{'}$ such that $\forall\ 1\leq j\leq l$, we have $V_\beta^{\pi^*}(r_j,s)\leq V_\beta^{\pi^{'}}(r_j,s)$ and $\exists 1\leq j\leq l$, we have $V_\beta^{\pi^*}(r_j,s)<V_\beta^{\pi^{'}}(r_j,s)$.*

For a Pareto-optimal strategy π^*, the corresponding value profile or discounted value profile $V_\beta^{\pi^*}(\bar{r},s)$ is referred to as a Pareto-optimal point.

Definition 2.5. *Given a β-discounted SMDP $\Gamma=<S,A,q,P,\bar{r}>$ with vector reward function $\bar{r}=(r_1,r_2,\cdots,r_l)$, a strategy $\pi^s\in\Pi$ is called sufficient for Pareto-optimality if for every discount factor $\beta\in[0,1)$, $s\in S$ and a Pareto-optimal strategy π^* such that*

$$V_\beta^{\pi^*}(\bar{r},s)\leq V_\beta^{\pi^s}(\bar{r},s) \tag{2.4}$$

where, $\bar{r}=(r_1,r_2,\cdots,r_l)$.

2.1.3 Main Results

Theorem 2.1.1. *There exist pure stationary Pareto-optimal strategies for discounted SMDPs, with vector pay-offs.*

Proof. Suppose $r_1,r_2,\cdots,r_l$ are the given reward functions of the discounted SMDP with reward vector $\bar{r}$ and discount factor $\beta\in[0,1)$. Now we construct a reward function r_{sum} as follows

$$r_{\text{sum}}=r_1+r_2+\cdots+r_l.$$

i.e., for all $s \in S$, we have $r_{\text{sum}}(s,a) = r_1(s,a) + r_2(s,a) + \cdots + r_l(s,a)$, where $a \in A(s)$. Let f^* be a pure stationary strategy, which is optimal for the reward function r_{sum}. Our objective is to show that f^* is Pareto-optimal, which proves the corollary. Let us assume by contradiction that f^* is not Pareto-optimal. Then there exists a strategy $\pi' \in \Pi$ such that for all $s \in S$, $V_\beta^{f^*}(\bar{r},s) \leq V_\beta^{\pi'}(\bar{r},s)$ and for some $1 \leq j \leq l$, $V_\beta^{f^*}(r_j,s) < V_\beta^{\pi'}(r_j,s)$, where $\bar{r} = (r_1, r_2, \cdots, r_l)$. Thus we have

$$\begin{aligned} V_\beta^{f^*}(r_{sum},s) &= \sum_{j=1}^{l} V_\beta^{f^*}(r_j,s) \\ &< \sum_{j=1}^{l} V_\beta^{\pi'}(r_j,s) \\ &= V_\beta^{\pi'}(r_{sum},s). \end{aligned} \tag{2.5}$$

This leads to the contradiction that f^* is optimal for r_{sum}. Thus f^* is a pure stationary Pareto-optimal strategy. □

Note that, pure strategies are not sufficient strategies for Pareto-optimality. This holds for pure stationary strategies also. The following example illustrates this fact.

Example 2.1. *Consider an SMDP with reward vector $\bar{r} = (r_1, r_2)$, discount factor $\beta \in [0,1)$ with state space $S = \{s_1, s_2\}$. $A(s_1) = \{a_1^1, a_1^2\}, A(s_2) = \{a_2^1\}$ are the action sets chosen on state s_1 and s_2 respectively. The reward vectors, sojourn times and transition probabilities are given as*

	state s_1
a_1^1	$(1,0)$ $(1\{2(1)\},0)$
a_1^2	$(0,1)$ $(0,1\{3(1)\})$

	state s_2
a_2^1	$(0,0)$ $(0,1\{1(1/2),2(1/2)\})$

where the rectangle $\boxed{\begin{array}{c}(r_1,r_2)\\ (q_1\{c_1(t_1),c_2(t_2)\},q_2\{g_1(t_1'),g_2(t_2')\})\end{array}}$ *represents that the reward vector is $\bar{r} = (r_1, r_2)$, q_1, q_2 are transition probabilities which denote that next states are s_1, s_2 respectively. $c_1, c_2; g_1, g_2$ are the discrete transition times and $t_1, t_2; t_1', t_2'$ are the transition time probability distributions for the states s_1, s_2 respectively, if this cell is chosen at present.*

Let us fix the initial state s_1 once and for all. Consider f to be a stationary strategy, defined by $f = \{(\frac{1}{2}, \frac{1}{2}), (1)\}$. This means the decision maker chooses action a_1^1 and a_1^2 each with probability $\frac{1}{2}$ in the state s_1. In state s_2 the decision maker chooses action a_2^1 with probability 1.

Now from equation (1.24) *discussed in section* 1.5.1, *the sub-stochastic matrix $P_\beta(f) = \begin{bmatrix} \frac{\beta^2}{2} & \frac{\beta^3}{2} \\ 0 & \frac{\beta(1+\beta)}{2} \end{bmatrix}$. Here $\bar{r}(f) = [\bar{r}(s_1, f), \bar{r}(s_2, f)]$ where $\bar{r}(s_i, f) = [r_1(s_i, f)$, $r_2(s_i, f)]$, $(i = 1, 2)$. The entries of $\bar{r}(s_1, f)$ can be calculated as $r_1(s_1, f) = \frac{1}{2}.1 + \frac{1}{2}.0$ and $r_2(s_1, f) = \frac{1}{2}.0 + \frac{1}{2}.1$. Similarly we can calculate $\bar{r}(s_2, f)$. Thus from* (1.25) *we get the value profile of the discounted SMDP $V_\beta^f(s_1, \bar{r}) = ((1 - \frac{\beta(1+\beta)}{2})(\frac{1}{2}, \frac{1}{2}))$. Now clearly f is a Pareto-optimal strategy and we can not get a better value profile than $V_\beta^f(s_1, \bar{r})$ by any other pure stationary strategy. This example shows that pure stationary strategies are not sufficient for Pareto-optimality.*

Let us consider $1_{(X_m = s^{'})}$ to be the reward function, which denotes that the reward is 1 when the system is in the state $s^{'}$ on the m-th decision epoch and 0 otherwise For a strategy $\pi \in \Pi$, a fixed initial state $s \in S$ once and for all, and a specified discount factor $\beta \in [0, 1)$, we define the counter of the state $s^{'} \in S$ as

$$Count_\beta^{s^{'}, \pi}(s) = \mathbb{E}_\pi \left[\sum_{m=1}^{\infty} \beta^{(\tau_1 + \tau_2 + \cdots + \tau_m)} 1_{(X_m = s^{'})} \mid X_1 = s \right]. \tag{2.6}$$

Theorem 2.1.2. *Suppose a stationary strategy $f \in F^s$ is given, the initial state s and discount factor $\beta \in [0, 1)$ is fixed in the above discussed SMDP model. Consider the vector $\bar{x} = (x_1, x_2, \cdots, x_z)$ of z variables, where $S = \{1, 2, \cdots, z\}$ is the finite state space, then the set of z equations*

$$x_s = \delta(s, s^{'}) + \sum_{a \in A(s)} [\sum_{s^{''} \in S} q(s^{''} \mid s, a) x_{s^{''}} \sum_{t=1}^{T} \beta^t P(\tau_{ss^{''}}^a = t)].f(s, a) \tag{2.7}$$

has a unique solution $Count_\beta^{s^{'}, f}(s)$, $\forall s, s^{'} \in S$.

Proof. We claim that, $Count_\beta^{s^{'}, f}(s) = \delta(s, s^{'}) + \sum_{a \in A(s)} [\sum_{s^{''} \in S} q(s^{''} \mid s, a) Count_\beta^{s^{'}, f}(s^{''})]$ $\sum_{t=1}^{T} \beta^t P(\tau_{ss^{''}}^a = t)].f(s, a)$. The proof of the above claim proves the theorem. Now, we have

$$\begin{aligned}
Count_{\beta}^{s',f}(s) &= \mathbb{E}_f[\sum_{m=1}^{\infty} \beta^{(\tau_1+\tau_2+\cdots+\tau_m)} 1_{(X_m=s')} \mid X_1 = s] \\
&= \delta(s,s') + \mathbb{E}[\mathbb{E}_f[\{\sum_{m=2}^{\infty} \beta^{(\tau_2+\cdots+\tau_m)} 1_{(X_m=s')}\} \mid hist_2, \tau_2]]
\end{aligned} \tag{2.8}$$

(where, we write $\mathbb{E}$ instead of $\mathbb{E}_f(. \mid X_1 = s)$) Thus from (2.8) we have

$$\begin{aligned}
Count_{\beta}^{s',f}(s) &= \delta(s,s') + \mathbb{E}[\beta^{\tau_2}.Count_{\beta}^{s',f}(X_2)] \\
&= \delta(s,s') + \sum_{a\in A(s)} \left[\sum_{s''\in S} q(s'' \mid s,a) Count_{\beta}^{s',f}(s'') \sum_{t=1}^{T} \beta^t P_{ss''}^a(t)\right].f(s,a)
\end{aligned} \tag{2.9}$$

The uniqueness follows from the uniqueness of values under stationary strategies[10]. Thus our theorem is proved. □

Now, we define Pareto-curve and the basic structure of a multi-objective linear programming problem.

Definition 2.6. *Let $\Gamma =< S, A, q, P, \bar{r} >$ (where $\bar{r} = (r_1, r_2, \cdots, r_l)$ is the vector reward function) be a discounted SMDP with discount factor β ($0 \le \beta < 1$) with vector reward. Let $s \in S$ be an arbitrary but fixed initial state in the SMDP model Γ. The Pareto curve $Par^{\beta}(\Gamma, s, \bar{r})$ of the SMDP $\Gamma =< S, A, q, P, \bar{r} >$ can be defined as the set of all of value profiles with l co-ordinates such that for each $p \in Par^{\beta}(\Gamma, s, \bar{r})$ there is a Pareto-optimal strategy π^* such that $V_{\beta}^{\pi^*}(\bar{r}, s) = p$.*

Definition 2.7. *Consider following optimisation problem*

$$\begin{gathered}
Maximise\ z_1(\bar{x}) = \bar{c_1}\bar{x} \\
Maximise\ z_2(\bar{x}) = \bar{c_2}\bar{x} \\
. \\
. \\
. \\
Maximise\ z_l(\bar{x}) = \bar{c_l}\bar{x}
\end{gathered}$$

Subject to $A.\bar{x} \ge \bar{b}$, $\bar{x} \ge 0$. A is a matrix of order $m \times l$ ($m \in \mathbb{N}$) and $\bar{b}$ is a vector of order $m \times 1$. $\bar{x} \ge 0$ is a vector of $l(\in \mathbb{N})$ variables. Here $\bar{c_i}$ ($i = 1, 2, \cdots, l$)s are vectors of co-efficients.

The set of equations above is the basic set up of multi-objective linear programming problem. We often use the abbreviation MOLP, instead of multi-objective linear programming problem to avoid complexity. In the above problem, l objective functions are to be optimised (here, maximised). In a multi-objective optimisation, unlike a single objective optimisation, there is no typical optimal solution, i.e., which is optimal for all objective functions.

2.1.4 An Algorithm To Compute Stationary Pareto-optimal Strategies Of A Discounted SMDP

Puterman [38] previously showed that an SMDP with discount factor $0 < \beta < 1$ can be solved by using a linear programming. We now propose a multi objective linear programming algorithm to solve a β-discounted SMDP, with finite state space $S = \{1, 2, \cdots, z\}$ and a vector reward $\bar{r} = (r_1, r_2, \cdots, r_l)$ for $l \in \mathbb{N}$. Consider the following linear programming problem:

MOLP 2.1

$$\max \textstyle\sum_{s=1}^{z} \sum_{a\in A(s)} r_i(s,a) x_{sa} \text{ for } i \in 1, 2, \cdots, l$$

subject to the constraints:
(i)$\sum_{a\in A(t)} x_{ta} = \delta(s,t) + \sum_{s^{'}\in S} \sum_{a^{'}\in A(s^{'})} [q(t \mid s^{'}, a^{'}) x_{s^{'}a^{'}} \sum_{y=1}^{T} \beta^y P(\tau_{s^{'}t}^{a^{'}} = y)]$, $t, s \in S$.
(ii)$x_{ta} \geq 0$ for $t \in S, a \in A(t)$.
Let us consider x_{ta} for $t \in S$ and $a \in A(t)$ to be any solution of the above multi-objective linear programming problem. Let $x_t = \sum_{a\in A(t)} x_{ta}$. This solution derives a stationary strategy which chooses action a at state t with probability $\frac{x_{ta}}{x_t}$. The LP with the i-th objective maximises the discounted reward for the i-th reward function r_i over the set of stationary strategies. Furthermore, from the solution of the above multi-objective linear programming we can compute the pay-off profile. We illustrate the above MOLP (multi-objective linear programming) algorithm in the following example.

2.1.5 Numerical Example

Example 2.2. *Consider the following β- discounted SMDP $\Gamma =< S, A, q, \bar{r}, P >$ with finite state space $S = \{1, 2, 3\}$, $A(2) = A(3) = \{1, 2\}$ and $A(1) = \{1\}$. Here, the vector reward function is defined by $\bar{r} = (r_1, r_2)$. The model is given as follows:*

State-1: $\boxed{\begin{array}{r}(1,1)\\(1\{1(1)\},0,0)\end{array}}$ *State*-2: $\boxed{\begin{array}{r}(2,3)\\(0,1\{1(\frac{1}{2}),2(\frac{1}{2})\},0)\\ \hline (4,2)\\(0,1\{1(\frac{1}{2}),2(\frac{1}{2})\},0)\end{array}}$

State-3: $\boxed{\begin{array}{r}(3,1)\\(\frac{1}{3}\{2(\frac{1}{2}),1(\frac{1}{2})\},\frac{2}{3}\{1(1)\},0)\\ \hline (3,5)\\(\frac{1}{9}\{1(\frac{1}{3}),2(\frac{1}{3}),3(\frac{1}{3})\},\frac{2}{9}\{1(1)\},\frac{2}{3}\{2(1)\})\end{array}}$

where the rectangle $\boxed{\begin{array}{c}(r_1,r_2)\\(q_1\{c_1(t_1)\},q_2\{g_1(t_1^{'})\},q_3\{d_1(t_1^{''})\})\end{array}}$ *represents that the reward vector is* $\bar{r}=(r_1,r_2)$*,* q_1,q_2,q_3 *are transition probabilities which denote that next states are* 1, 2 *and* 3 *respectively.* $c_1;g_1$ *and* d_1 *are the discrete transition times and* $t_1;t_1^{'}$ *and* $t_1^{''}$ *are the transition time probability distributions for the states* $1,2$ *and* 3 *respectively, if this cell is chosen at present. If we fix the initial state to be* 1*, then we can write the MOLP with respect to the variables* $(x_{11},x_{21},x_{22},x_{31},x_{32})$ *as follows:*

$$f_1=\max(x_{11}+2x_{21}+4x_{22}+3x_{31}+3x_{32})$$
$$\textit{and}$$
$$f_2=\max(x_{11}+3x_{21}+2x_{22}+x_{31}+5x_{32})$$

with respect to the constraints:

$$(i)x_{11}=1+x_{11}(\beta^2)+\frac{1}{3}x_{31}\frac{(\beta^2+\beta)}{2}+\frac{1}{9}x_{32}\frac{\beta+\beta^2+\beta^3}{3} \tag{2.10}$$
$$(ii)x_{21}+x_{22}=x_{21}\frac{(\beta+\beta^2)}{2}+x_{22}\frac{(\beta+\beta^2)}{2}+\frac{2\beta}{3}x_{31}+\frac{2\beta}{9}x_{32}. \tag{2.11}$$
$$(iii)x_{31}+x_{32}=\frac{2\beta^2}{3}x_{32} \tag{2.12}$$
$$\tag{2.13}$$

Now, if we fix the value of $\beta=0.5$*, then the Pareto-optimal solutions of the above MOLP is:*

x_{11}	x_{21}	x_{22}	x_{31}	x_{32}
0.3681	0.0327	0.0000	0.0248	0.0000
0.3682	0.0325	0.0000	0.0248	0.0000
0.3682	0.0326	0.0000	0.0248	0.0000

The values of the objective functions corresponding to each Pareto-optimal solution are given below:

f_1	f_2
0.5079	0.49126
0.50839	0.49119
0.50852	0.49112

Now, if we plot the values of the objective function corresponding to each Pareto-optimal solution, then the Pareto-curve will look like this:

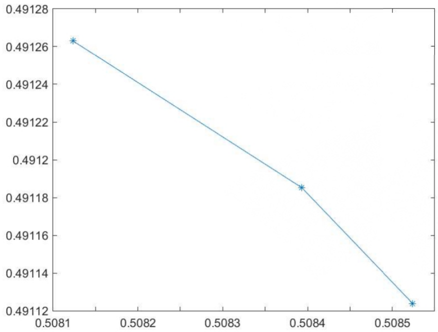

Fig. 2.1 Pareto Curve

Thus, from the above set of solutions, we can calculate a stationary Pareto-optimal strategy for the decision maker. Taking the values of x_{ta} $(t \in S, a \in A(t))$ *from* 2*nd row of the above table, the optimal stationary strategy is* $f^* = \{1,(1,0),(1,0)\}$ *and the pay-off profile corresponding to this solution is:* $(0.50839, 0.49119)$.

2.1.6 Undiscounted Semi-Markov Decision Processes With Vector Pay-Offs

A finite semi-Markov decision process (SMDP) with finite state and action spaces is defined by a collection of objects $< S, A = A(s) : s \in S, q, P, r >$, where $S = \{1,2,\cdots,z\}$ is the finite state space, $A(s) = \{1,2,\cdots,m_s\}$ is the finite admissible action space in the state $s \in S$. The transition probability is denoted by $q(s^{'} \mid s,a)$ for each $s,s^{'} \in S$, i.e., $q(s^{'} \mid s,a) \geq 0$ and $\sum_{s^{'} \in S} q(s^{'} \mid s,a) = 1$. $P^a_{ss^{'}}(.)$ is the conditional time distribution function on $[0,\infty)$ and $r(s,a)$ is the immediate (expected) reward. The process starts at s state $s \in S$ and the decision maker chooses an action $a \in A(s)$. Consequently he/she receives an immediate reward $r(s,a)$ and the system moves to the next state $s^{'} \in S$ with probability $q(s^{'} \mid s,a)$ and following transition time distribution $P^a_{ss^{'}}(.)$. Once the transition to the state $s^{'}$ is completed, on the next decision epoch, the entire process, with s replaced by $s^{'}$ is repeated over and over again. Thus the SMDP proceeds over infinite time. The definitions of behavioural, Markov, stationary, semi-Markov and semi-stationary strategies are defined similarly as in Chapter 1. Let $(X_1, A_1, X_2, A_2, \cdots)$ be a co-ordinate sequence in $S \times (A \times S)^{\infty}$. Given a behavioural strategy $\pi \in \Pi$ and an initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi}(. \mid X_1 = s)$ (hence an expectation $\mathbb{E}_{\pi}(. \mid X_1 = s)$) on the product σ-field of $S \times (A \times S)^{\infty}$ by Kolmogorov's extension theorem.

Definition 2.8. *For a behavioural strategy* $\pi \in \Pi$ *and a single reward function* r*, the limiting ratio average pay-off is defined as:*

$$\phi(r,s,\pi) = \liminf_{n\to\infty} \tfrac{\mathbb{E}_{\pi}[\sum_{m=1}^{n} r(X_m,A_m)|X_1=s]}{\mathbb{E}_{\pi}[\sum_{m=1}^{n} \bar{\tau}(X_m,A_m)|X_1=s)]} \text{ for all } s \in S.$$

The expected sojourn time here is defined by:

$$\bar{\tau}(s,a) = \textstyle\sum_{s^{'} \in S} q(s^{'} \mid s,a) \int_0^{\infty} t dP_{ss^{'}}(t \mid a).$$

We assume that $\bar{\tau}(s,a)$ is bounded away from zero for all $(s,a) \in K$, where $K = \{(s,a) \mid s \in S, a \in A(s)\}$. In this section we consider an undiscounted semi-Markov decision process where instead of a single reward function, a vector reward function is considered. Thus the notion of optimality is changed to Pareto-optimality here. For an undiscounted SMDP with a fixed initial state s, strategy $\pi \in \Pi$ and $l(l \in \mathbb{N})$ different reward functions $r_1, r_2, \cdots, r_l$, the undiscounted value profile(or pay-off profile) w.r.t. $\bar{r} = (r_1, r_2, \cdots, r_l)$ is defined by

$$\phi(\bar{r}, s, \pi) = (\phi(r_1, s, \pi), \cdots, \phi(r_l, s, \pi)). \tag{2.14}$$

Definition 2.9. *For an undiscounted SMDP Γ with reward vector $\bar{r} = (r_1, r_2, \cdots, r_l)$, a strategy $\pi^* \in \Pi$ is called Pareto-optimal if there is no strategy $\pi^{'} \in \Pi$ such that both $\phi(\bar{r}, s, \pi^*) \leq \phi(\bar{r}, s, \pi^{'})$ and $\phi(\bar{r}, s, \pi^*) \neq \phi(\bar{r}, s, \pi^{'})$ hold, i.e., $\nexists$ any strategy $\pi^{'}$ such that $\forall\ 1 \leq j \leq l$, we have $\phi(r_j, s, \pi^*) \leq \phi(r_j, s, \pi^{'})$ and $\exists 1 \leq j \leq l$, we have $\phi(r_j, s, \pi^*) < \phi(r_j, s, \pi^{'})$.*

For a Pareto-optimal strategy π^*, the corresponding value profile or undiscounted value profile $\phi(\bar{r}, s, \pi^*)$ is referred to as a Pareto-optimal point.

The next theorem investigates the existence of a pure semi-stationary Pareto-optimal strategy for an undiscounted SMDP with vector rewards.

2.1.7 Main Result

Theorem 2.1.3. *There exist pure semi-stationary Pareto-optimal strategies for undiscounted SMDPs, with vector pay-offs.*

Proof. Suppose $r_1, r_2, \cdots, r_l$ are the given reward functions of the undiscounted SMDP with reward vector $\bar{r}$. Now we construct a reward function r_{sum} as follows

$$r_{sum} = r_1 + r_2 + \cdots + r_l.$$

Let us fix an initial state $s \in S$. Now by Sinha et al.[43], we know that for a semi-Markov decision process with limiting ratio average pay-off (undiscounted SMDP), there exists an optimal pure semi-stationary strategy $f^* = (f_1^*, f_2^*, \cdots, f_s^*, \cdots, f_z^*)$, where f_s^* is a pure stationary strategy of the decision maker in the undiscounted SMDP, when the initial state is s. Suppose, f_s^* be a pure stationary strategy, which is optimal for the reward function r_{sum} for the initial state $s \in S$. Our objective is to show that f_s^* is Pareto-optimal, for the initial state $s \in S$ in the undiscounted SMDP which proves the theorem. Let us assume by contradiction that f_s^* is not Pareto-optimal. Then there exists a strategy $\pi \in \Pi$ such that $\phi(\bar{r}, s, f_s^*) \leq \phi(\bar{r}, s, \pi)$ and for some $1 \leq j \leq l$, $\phi(r_j, s, f_s^*) < \phi(r_j, s, \pi)$, where $\bar{r} = (r_1, r_2, \cdots, r_l)$. Thus we have

$$\begin{aligned}\phi(r_{sum},s,f_s^*) &= \sum_{j=1}^{l}\phi(r_j,s,f_s^*)\\ &< \sum_{j=1}^{l}\phi(r_j,s,\pi)\\ &= \phi(r_{sum},s,\pi). \qquad (2.15)\end{aligned}$$

This leads to the contradiction that f_s^* is optimal for r_{sum} for the initial state $s \in S$. Thus f_s^* is a pure stationary Pareto-optimal strategy of the undiscounted SMDP for the initial state $s \in S$. Consequently, $f^* = \{f_1^*, f_2^*, \cdots, f_s^*, \cdots, f_z^*\}$ is the pure-semi stationary Pareto strategy for the undiscounted SMDP with vector reward $\bar{r} = (r_1, r_2, \cdots, r_l)$. □

2.1.8 Algorithm To Compute A Pure Semi-Stationary Pareto-Optimal Strategy For An Undiscounted SMDP

Here we use an existing algorithm by Mondal (2020) [33] to compute the value and an optimal pure semi-stationary strategy of the decision maker. We extend the reward function to a vector reward function here to compute a pure semi-stationary Pareto-optimal strategy of the decision maker.

Suppose Γ be an undiscounted SMDP with vector reward function $\bar{r} = (r_1, r_2, \cdots, r_l)$. Let s_0 be a fixed but arbitrary initial state of Γ. We consider the following linear programming problem in the variables $v(s_0)$, $g = (g_s : s \in S)$ and $h = (h_s : s \in S)$ as:

$$LP : \min v(s_0)$$

subject to

$$g_s \geq \sum_{s' \in S} q(s' \mid s,a) g_{s'} \ \forall s \in S, a \in A(s). \qquad (2.16)$$

$$g_s + h_s \geq r(s,a) - v(s_0)\bar{\tau}(s,a) + \sum_{s' \in S} q(s' \mid s,a) h_{s'} \ \forall s \in S, a \in A(s). \qquad (2.17)$$

$$g_{s_0} \leq 0. \qquad (2.18)$$

The variables $v(s_0), (g_s : s \in S, s_0 \neq s)$ and $(h_s : s \in S)$ are unrestricted in sign. The dual linear programming problem of this primal for the variables $x = (x_{sa} : s \in S, a \in A(s))$ and $y = (y_{sa} : s \in S, a \in A(s))$ and t is given by

$$DLP : \max R_s, \text{ where } R_s = \sum_{s\in S}\sum_{a\in A(s)} r(s,a)x_{sa}$$

subject to

$$\sum_{s\in S}\sum_{a\in A(s)}\{\delta(s,s^{'}) - q(s^{'} \mid s,a)\}x_{sa} = 0 \ \forall s^{'} \in S. \tag{2.19}$$

$$\sum_{a\in A(s')} x_{s'a} + \sum_{s\in S}\sum_{a\in A(s)}\{\delta(s,s^{'}) - q(s^{'} \mid s,a)\}y_{sa} = 0 \ \forall s^{'} \in S - \{s_0\}. \tag{2.20}$$

$$\sum_{a\in A(s_0)} x_{s_0 a} + \sum_{s\in S}\sum_{a\in A(s)}\{\delta(s,s_0) - q(s_0 \mid s,a)\}y_{sa} - t = 0. \tag{2.21}$$

$$\sum_{s\in S}\sum_{a\in A(s)} \bar{\tau}(s,a)x_{sa} = 1. \tag{2.22}$$

$$x_{sa}, y_{sa} \geq 0 \ \forall s \in S, a \in A(s), t \geq 0. \tag{2.23}$$

where $\delta(s,s^{'})$ is the Kronecker delta function. For a feasible solution (x,y,t) of the DLP, we define the following sets associated with the feasible solution:

$$\begin{aligned}
&S_x = \{s \in S : \textstyle\sum_{a\in A(s)} x_{sa} > 0\}\\
&S_y = \{s \in S : \textstyle\sum_{a\in A(s)} x_{sa} = 0 \text{ and } \sum_{a\in A(s)} y_{sa} > 0\}\\
&S_{xy} = \{s \in S : \textstyle\sum_{a\in A(s)} x_{sa} = 0 \text{ and } \sum_{a\in A(s)} y_{sa} = 0\}.
\end{aligned}$$

Thus $S = S_x \cup S_y \cup S_{xy}$, where S_x, S_y and S_{xy} are pairwise disjoint sets. A pure stationary strategy corresponding to the feasible solution (x,y,t) of the DLP is defined by $f^{ps_0}_{xyt}$, where s_0 is the fixed but arbitrary initial state $f^{ps_0}_{xyt}(s) = a_s, s \in S$ such that:

$$a_s = \begin{cases} a & \text{if } s \in S_x \text{ and } x_{sa} > 0 \\ a^{'} & \text{if } s \in S_y \text{ and } y_{sa^{'}} > 0 \\ \text{arbitrary} & \text{if } s \in S_{xy} \end{cases}$$

From [33], we have the following theorem.

Theorem 2.1.4. *Let (x^*,y^*,t^*) be an optimal solution of the DLP. Then $f^{ps_0}_{x^*y^*t^*}$is a pure stationary optimal strategy of the SMDP for the initial state s_0.*

2.1.9 Numerical Example

Example 2.3. *Consider an undiscounted SMDP* Γ *with five states* $S=\{1,2,3,4,5\}$, $A(1)=\{1,2\}=A(2)=A(3)=A(4)$. $A(5)=\{1\}$. *Rewards, transition probabilities and expected sojourn times for the decision maker are given below:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline (11,10)\\ (\frac{1}{2},\frac{1}{2},0,0,0)\\ 1.1\\ \hline (1,1)\\ (\frac{1}{3},\frac{2}{3},0,0,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-2:}\ \begin{array}{|r|}\hline (2,3)\\ (\frac{1}{2},\frac{1}{2},0,0,0)\\ 1\\ \hline (4,2)\\ (\frac{2}{3},\frac{1}{3},0,0,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-3:}\ \begin{array}{|r|}\hline (2,3)\\ (0,0,1,0,0)\\ 0.9\\ \hline (4,5)\\ (0,0,1,0,0)\\ 1\\ \hline\end{array}$$

$$\textit{State-4:}\ \begin{array}{|r|}\hline (4,2)\\ (\frac{1}{2},0,0,0,\frac{1}{2})\\ 2\\ \hline (2,3)\\ (0,0,0,1,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-5:}\ \begin{array}{|r|}\hline (3,8)\\ (\frac{1}{3},0,\frac{2}{3},0,0)\\ 2\\ \hline\end{array}$$

Now we get the following undiscounted SMDP $\hat{\Gamma}$ *by converting the reward vectors to a single reward in each cell from this SMDP:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline 21\\ (\frac{1}{2},\frac{1}{2},0,0,0)\\ 1.1\\ \hline 2\\ (\frac{1}{3},\frac{2}{3},0,0,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-2:}\ \begin{array}{|r|}\hline 5\\ (\frac{1}{2},\frac{1}{2},0,0,0)\\ 1\\ \hline 6\\ (\frac{2}{3},\frac{1}{3},0,0,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-3:}\ \begin{array}{|r|}\hline 5\\ (0,0,1,0,0)\\ 0.9\\ \hline 9\\ (0,0,1,0,0)\\ 1\\ \hline\end{array}$$

$$\textit{State-4:}\ \begin{array}{|r|}\hline 6\\ (\frac{1}{2},0,0,0,\frac{1}{2})\\ 2\\ \hline 5\\ (0,0,0,1,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-5:}\ \begin{array}{|r|}\hline 11\\ (\frac{1}{3},0,\frac{2}{3},0,0)\\ 2\\ \hline\end{array}$$

Next we implement our LP algorithm to solve this SMDP and obtain pure semi-stationary strategy of the decision maker. For a fixed initial state s_0*, the DLP in the variables* $x = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32}, x_{41}, x_{42}, x_{51})$*,* $y = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32}, y_{41}, y_{42}, y_{51})$ *and* t *can be written as*

$$\max R_{s_0} = 21x_{11} + 2x_{12} + 5x_{21} + 6x_{22} + 5x_{31} + 9x_{32} + 6x_{41} + 5x_{42} + 11x_{51}$$

subject to

$$3x_{11} + 4x_{12} - 3x_{21} - 4x_{22} - 3x_{41} - 2x_{51} = 0 \tag{2.24}$$

$$-3x_{11} - 4x_{12} + 3x_{21} + 4x_{22} = 0 \tag{2.25}$$

$$x_{51} = 0 \tag{2.26}$$

$$x_{41} = 0 \tag{2.27}$$

$$2x_{51} - x_{41} = 0 \tag{2.28}$$

$$6x_{11} + 6x_{12} + 3y_{11} + 4y_{12} - 3y_{21} - 4y_{22} - 3y_{41} - 2y_{51} - 6\delta(s_0, 1)t = 0 \tag{2.29}$$

$$6x_{21} + 6x_{22} - 3y_{11} - 4y_{12} + 3y_{21} + 4y_{22} - 6\delta(s_0, 2)t = 0 \tag{2.30}$$

$$3x_{31} + 3x_{32} - 2y_{51} - 3\delta(s_0, 3)t = 0 \tag{2.31}$$

$$x_{41} + x_{42} + y_{41} - \delta(s_0, 4)t = 0 \tag{2.32}$$

$$2x_{51} + 2y_{51} - y_{41} - 2\delta(s_0, 5)t = 0 \tag{2.33}$$

$$x_{11} + 0.9x_{12} + x_{21} + 1.1x_{22} + x_{31} + 2x_{32} + 2x_{41} + 1.1x_{42} + 2x_{51} = 1 \tag{2.34}$$

$$x, y, t \geq 0. \tag{2.35}$$

The solution of the above linear prgramming problem by dual-simplex method for different initial states s_0 *are given by:*
(i) For $s_0 = 1$*:* $\max R_1 = 3.8806$*,* $x = (0, 0.4478, 0.5970, 0, 0, 0, 0, 0, 0)$*,*
$y = (0, 0.8955, 0, 0, 0, 0, 0, 0, 0)$*,* $t = 1.0448$*.*
(ii) For $s_0 = 2$*:* $\max R_2 = 3.8806$*,* $x = (0, 0.4478, 0.5970, 0, 0, 0, 0, 0, 0)$*,*
$y = (0, 0.8955, 0, 0, 0, 0, 0, 0, 0)$*,* $t = 1.04$*.*
(iii) For $s_0 = 3$*:* $\max R_3 = 4.5$*,* $x = (0, 0, 0, 0, 0, 0.5, 0, 0, 0)$*,*
$y = (0, 0, 0, 0, 0, 0, 0, 0, 0)$*,* $t = 0.5$*.*
(iv) For $s_0 = 4$*:* $\max R_4 = 4.4525$*,* $x = (0, 0.2190, 0.2920, 0, 0.5109, 0, 0, 0, 0)$*,*
$y = (0, 0.4380, 0, 0, 0, 0, 0, 1.0219, 0)$*,* $t = 1.0219$*.*
(v) For $s_0 = 5$*:* $\max R_5 = 4.379492$*,* $x = (0, 0.0864553, 0.115274, 0, 0, 0.403458, 0, 0, 0)$*,*
$y = (0, 0.172911, 0, 0, 0, 0, 0, 0, 0.605187)$*,* $t = 0.605187$*.*

Thus, we find that the SMDP has a pure semi-stationary Pareto-optimal strategy, which is given by $\hat{f}^* = (f_1^*, f_2^*, f_3^*, f_4^*, f_5^*)$*, where* $f_1^* = (2,1,1,1,1)$ *and* $f_2^* = (2,1,2,1,1)$ *and* $f_3^* = (2,1,1,2,1)$*,* $f_4^* = (2,1,1,1,1) = f_5^*$*.*

2.2 Concluding Remarks

In this chapter, we deal with vector reward function instead of a single reward function in both discounted and undiscounted (limiting ratio average) SMDP. Thus the concept of Pareto optimality comes into play. We prove the existence of pure stationary/ semi-stationary Pareto-optimal strategies in discounted and undiscounted SMDP models respectively. Also, we can find a stationary/ semi-stationary Pareto-optimal strategy of the decision maker in the discounted as well as undiscounted SMDP models by the algorithms discussed in the sections 2.1.4 and 2.1.8.

Chapter 3

Zero-Sum Two Person Undiscounted Perfect Information Stochastic and Semi-Markov Games

3.1 Zero-Sum Two Person Undiscounted Perfect Information Stochastic Game

3.1.1 Introduction

Stochastic games are generalisations of Markov decision processes (MDPs) to the case of two or more players. Shapley (1953) [42] introduced 'Stochastic games', which is also known as Markov games these days. If two players play a matrix game repeatedly over the infinite time horizon and the limiting average payoff is considered, then the value of this infinitely repeated game coincides with the value of the one shot game (by Folk Theorem [11]). Shapley [42] introduced the idea of not playing the same matrix game everyday (i.e., in every stage of the game), but playing one among finitely many matrix games, with a motion among them governed by the present game and the actions chosen there in such a manner that the game is certain to stop in finite time. Then the payoffs of the players can be formulated as the ratio of two bilinear forms. Neumann [36] established the minimax theorem for such games and Loomis [26] gave an elementary proof of this theorem. The case of non-terminating limiting average stochastic games was studied by Gillette [12] and Hoffman and Karp [13] . By undiscounted pay-off we mean limiting average pay-off in this paper. Gillette [12] and Liggett and Lippman [24] previously proved the existence of a pure stationary optimal strategy pair of the players in an undiscounted perfect information stochastic game. We propose an alternate proof of the result by Liggett and Lippman [24] which is simple and straightforward. By forming the matrix of undiscounted payoffs corresponding to each pair of pure stationary strategies (for each initial state) of the two players we prove that this matrix has a pure saddle point, which is essentially a pure semi-stationary strategy pair of the players. Then we prove the existence of optimal pure stationary strategy pair of the players by using the results by Derman [5]. We consider the policy-improvement algorithm to compute an optimal pure stationary strategy pair of the players. This is a best response algorithm, in which each player looks for his own undiscounted optimal strategy. As the set of pure stationary strategies is finite, so this is a finite step algorithm and it terminates in finite time (Raghavan and Syed (2002) [39]) . This section of the chapter 3 is organised as follows. Section 3.1.2 contains definitions and properties of an undiscounted two person zero-sum stochastic game considered under limiting average pay-off. Section 3.1.5 contains main result of this paper. In section 3.1.6 we propose a policy improvement algorithm to compute an optimal stationary strategy pair of the players in a zero-sum two person perfect information undiscounted stochastic game. Section 3.1.7 contains some numerical examples illustrating our theorem and proposed algorithm.

3.1.2 Preliminaries

A zero-sum two person finite stochastic game is described by a collection of five objects $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, r >$, where $S = \{1, 2, \cdots, z\}$ is the finite non-empty state space and $A(s) = \{1, 2, \cdots, m_s\}, B(s) = \{1, 2, \cdots, n_s\}$ are respectively the non-empty sets of admissible actions of the players I and II respectively in the state s. Let us denote $K = \{(s, i, j) : s \in S, i \in A(s), j \in B(s)\}$ to be the set of admissible triplets. For each $(s, i, j) \in K$, we denote $q(. \mid s, i, j)$ to be the transition law of the game. Finally r is the real valued functions on K, which represents the immediate (expected) reward for the player-I (whereas -r is the reward for the player-II). Let us consider player-I as the maximiser and player-II as the minimiser in the zero-sum two person stochastic game.

The Stochastic game over infinite time is played as follows. At the 0th decision epoch, the game starts at $s_0 \in S$ and the players I and II simultaneously and independently choose actions $i_0 \in A(s_0)$ and $j_0 \in B(s_0)$ respectively. Consequently player-I and II get immediate rewards $r(s_0, i_0, j_0)$ and $-r(s_0, i_0, j_0)$ respectively and the game moves to the state s_1 with probability $q(s_1 \mid s_0, i_0, j_0)$. After reaching the state s_1 on the next decision epoch, the game is repeated over infinite time with the state s_0 replaced by s_1. Shapley extended the idea of defining SGs where $\sum_{s_1 \in S} q(s_1 \mid s_0, i_0, j_0) < 1$ for all $(s_0, i_0, j_0) \in K$ and the play terminates with probability $1 - \sum_{s_1 \in S} q(s_1 \mid s_0, i_0, j_0) < 1$. Such games are called 'stopping SGs'. The 'non-stopping SGs' are those where $\sum_{s_1 \in S} q(s_1 \mid s_0, i_0, j_0) = 1$ for all $(s_0, i_0, j_0) \in K$, i.e., the play never terminates.

By a strategy (behavioural) π_1 of the player-I, we mean a sequence $\{(\pi_1)_n(. \mid hist_n)\}_{n=1}^{\infty}$, where $(\pi_1)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n$=$(s_0, a_0, b_0, s_1, a_1, b_1 \cdots, s_{n-1}, a_{n-1}, b_{n-1}, s_n)$ for $n \geq 2$, $hist_1 = (s_0)$ and $(s_k, a_k, j_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi_1)_n(. \mid hist_n)$ on $A(s_n)$. Behavioural strategy π_2 for player -II can be defined analogously. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π_1 and Π_2 to be the sets of strategy (behavioural) spaces of the players I and II respectively. A strategy $f^{'} = \{f_n^{'}\}_{n=1}^{\infty}$ for the player-I is called semi-Markov if for each n, $f_n^{'}$ depends on s_0, s_n and the decision epoch number n. Similarly we can define a semi-Markov strategy $g^{'} = \{g_n^{'}\}_{n=1}^{\infty}$ for the player-II.
A strategy $\pi_1 = \{\pi_{1n}\}_{n=1}^{\infty}$ is called a stationary strategy if $\exists$ a map $f : S \to \mathbb{P}(A) = \{\mathbb{P}(A(s)) : s \in S\}$, where $\mathbb{P}(A(s))$ is the set of probability distribution on $A(s)$ such that $\pi_{1n} = f$ for all n and $f(s) \in \mathbb{P}(A(s))$. A stationary strategy for player-I is defined as z tuple $f = (f(1), f(2), \cdots, f(z))$, where each $f(s)$ is the probability distribution on $A(s)$ given by $f(s) = (f(s,1), f(s,2), \cdots, f(s, m_s))$. $f(s,i)$ denotes the probability of choosing action i in the state s by player-I. By similar manner, one can define a stationary strategy g for player-II as $g = (g(1), g(2), \cdots, g(z))$ where each $g(s)$ is the probability distribution on $B(s)$. Let us denote F_1^s and F_2^s to be the set of stationary strategies for player-I and II respectively. A semi-stationary strategy is a semi-Markov strategy which is independent of the decision epoch n, i.e., for a initial state s_0 and present state s_1, if a semi-Markov strategy $f^{'}(s_0, s_1, n)$ turns out to be independent of n, then we call it a semi-stationary strategy. Let us denote ξ_1^s and ξ_2^s to be the set of semi-stationary strategies for player-I and II respectively.
A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F_1^{sp} and F_2^{sp} to be the set of pure stationary strategies of the players I and II respectively. Also ξ_1^{sp} and ξ_2^{sp} are denoted as the set of pure semi-stationary strategies for the player-I and II respectively.

Definition 3.1. *A zero-sum two person SG* $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, r >$ *is called a perfect information stochastic game (PISG) if the following properties hold*
(i)$S = S_1 \cup S_2, S_1 \cap S_2 = \phi$.
(ii)$\mid B(s) \mid= 1$, *for all* $s \in S_1$, *i.e., on* S_1 *player-II is a dummy (i.e., for states* $\{1, 2, \cdots, \mid S_1 \mid\}$ *player-II has only* 1 *action).*
(iii)$\mid A(s) \mid= 1$, *for all* $s \in S_2$, *i.e., on* S_2 *player-I is a dummy (i.e., for states* $\{\mid S_1 \mid +1, \mid S_1 \mid +2, \cdots, \mid S_1 \mid + \mid S_2 \mid\}$ *player-I has only* 1 *action).*

3.1.3 Undiscounted Zero-Sum Two Person Stochastic Games

Let $(X_0, A_0, B_0, X_1, A_1, B_1 \cdots)$ be a co-ordinate sequence in $S \times (A \times B \times S)^{\infty}$. Given behavioural strategy pair $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi_1 \pi_2}(. \mid X_0 = s)$ (hence an expectation $\mathbb{E}_{\pi_1 \pi_2}(. \mid X_0 = s)$) on the product σ- field of $S \times (A \times B \times S)^{\infty}$ by Kolmogorov's extension theorem. For a pair of strategies $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$ for the players I and II respectively, the limiting average (undiscounted) pay-off for player-I from player-II, starting from a state $s \in S$ is defined by:

$$\phi(s, \pi_1, \pi_2) = \liminf_{n \to \infty} \frac{1}{n+1} \mathbb{E}_{\pi_1 \pi_2} \sum_{m=0}^{n} [r(X_m, A_m, B_m) \mid X_0 = s] \tag{3.1}$$

Alternatively, for any pair of stationary strategies $(f_1, f_2) \in F_1^s \times F_2^s$ of player-I and II, we write the undiscounted pay-off for player-I from player-II as:

$$\phi(s, f_1, f_2) = \lim_{n \to \infty} \frac{1}{n+1} \sum_{m=0}^{n} r^m(s, f_1, f_2) \tag{3.2}$$

(as limit exists for a pair of stationary strategies (f_1, f_2)) for all $s \in S$. Where $r^m(s, f_1, f_2)$ is the expected reward for player-I at the m-th decision epoch, when player-I chooses f_1 and player-II chooses f_2 respectively and the initial state is s.

Definition 3.2. *For a pair of strategies* $(f_1, f_2) \in F_1^s \times F_2^s$, *we define the transition probability matrix by:*

$$Q(f_1,f_2) = [q(s^{'} \mid s, f_1(s), f_2(s))]_{s,s^{'}=1}^{z},$$

where $q(s^{'} \mid s, f_1(s), f_2(s)) = \sum_{i\in A(s)} \sum_{j\in B(s)} q(s^{'} \mid s,i,j) f_1(s,i) f_2(s,j)$ is the probability is that the system jumps to the state $s^{'}$ from given state s when the players play the stationary strategies f_1 and f_2 respectively.

Lemma 3.1.1. *(Kemeney and Snell, 1976, [21]) Let Q be any $z\times z$ Markov matrix, then the sequence $\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} Q^m(f_1,f_2)$ converges as $n\to\infty$ to a Markov matrix Q^* (the Cesaro limiting matrix) such that $QQ^* = Q^*Q = Q^*Q^* = Q^*$.*
For each $(f_1,f_2)\in F_1^s \times F_2^s$, we define $r(f_1,f_2) = [r(s,f_1,f_2)]_{z\times 1}$ as the expected reward, where for each $s\in S$,

$$r(s,f_1,f_2) = \sum_{i\in A(s)} \sum_{j\in B(s)} r(s,i,j) f_1(s,i) f_2(s,j).$$

Now we have the following result:
***Proposition 1** For each player of pure stationary strategies $(f_1,f_2)\in F_1^{sp}\times F_2^{sp}$,*

$$\phi(s,f_1,f_2) = [Q^*(f_1,f_2) r(f_1,f_2)](s) \forall s\in S.$$

Where $Q^*(f_1,f_2)$ is the Cesaro limiting matrix of $Q(f_1,f_2)$.

Definition 3.3. *A zero-sum two person undiscounted stochastic game is said to have a value vector $\phi = [\phi(s)]_{z\times 1}$ if*

$$\sup_{\pi_1\in\Pi_1} \inf_{\pi_2\in\Pi_2} \phi(s,\pi_1,\pi_2) = \phi(s) = \inf_{\pi_2\in\Pi_2} \sup_{\pi_1\in\Pi_1} \phi(s,\pi_1,\pi_2) \text{ for all } s\in S.$$

A pair of strategies $(\pi_1^*,\pi_2^*)\in \Pi_1,\times\Pi_2$ is said to be an optimal strategy pair for the players if

$$\phi(s,\pi_1^*,\pi_2) \geq \phi(s) \geq \phi(s,\pi_1,\pi_2^*) \text{ for all } s\in S \text{ and all } (\pi_1,\pi_2)\in \Pi_1\times\Pi_2.$$

3.1.4 Undiscounted Markov Decision Processes

A finite (state and action spaces) Markov decision process (**MDP**) is defined by a collection of four objects $\hat{\Gamma} =< S, \hat{A} = \{A(s) : s\in S\}, \hat{q}, \hat{r} >$, where $S = \{1,2,\cdots,z\}$ is the finite state space, $\hat{A}(s) = \{1,2,\cdots,d\}$ is the finite set of admissible actions in the state s. $\hat{q}(s^{'} \mid s,a)$ is the transition probability (i.e., $\hat{q}(s^{'} \mid s,a) \geq 0$ and $\sum_{s^{'}\in S} \hat{q}(s^{'} \mid s,a) = 1$) that the next state is $s^{'}$, where s is the initial state and the decision maker chooses action a in the state s. The decision process proceeds over infinite time just as stochastic game, where instead of two players we consider a single decision maker. The definition of strategy spaces for the decision maker is same as in the case of stochastic games. Let us denote Π, F^s, F^{sp} as the set of behavioural, stationary, pure-stationary strategies respectively of the decision maker. Let $(X_0, A_0, X_1, A_1, \cdots)$ be a coordinate sequence in $S \times (\hat{A} \times S)^\infty$. Given a behavioural strategy $\pi \in \Pi$, initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_\pi(. \mid X_0 = s)$ (hence an expectation $\mathbb{E}_\pi(. \mid X_0 = s)$) on the product σ- field of $S \times (\hat{A} \times S)^\infty$ by Kolmogorov's extension theorem.

For a behavioural strategy $\pi \in \Pi$, the expected limiting average pay-off is defined by

$$\hat{\phi}(s, \pi) = \liminf_{n \to \infty} \frac{1}{n+1} \sum_{m=0}^{n} \mathbb{E}_\pi[\hat{r}(X_m, A_m) \mid X_0 = s]. \tag{3.3}$$

for all $s \in S$.

3.1.5 Main Result

Theorem 3.1.2. *Any zero-sum two person undiscounted perfect information Stochastic game has a solution in pure stationary strategies.*

Note: Liggett and Lippmann (1969) [24] gave the first correct proof of the above theorem. But their proof is indirect as well as involved. The proof given below is direct and straightforward.

Proof. Let $\Gamma =< S = S_1 \cup S_2, A = \{A(s) : s \in S_1\}, B = \{B(s) : s \in S_2\}, q, r >$ be a zero-sum two person perfect information Stochastic game under limiting average pay-off, where $S = \{1, 2, \cdots, z\}$ is the finite state space. We assume that in $\mid S_1 \mid$ number of states, player-II is a dummy and for states $\{\mid S_1 \mid +1, \cdots, \mid S_1 \mid + \mid S_2 \mid\}$ player-I is a dummy. Without loss of generality, we assume that each player has d number of pure actions in each state where they are non-dummy. Thus, player-I has $\mid S_1 \mid .d$ number of pure actions available in each state $s \in S$, where he/she is non-dummy and player-II has $\mid S_2 \mid .d$ number of pure actions where he/she is non-dummy in the PISG Γ. Let us the consider the pay-off matrix $A(s)$ for the initial state $s \in S$:

$$A(s)_{|S_1|.d \times |S_2|.d} = \begin{bmatrix} \phi(s, f_1, g_1) & \phi(s, f_1, g_2) & \cdots & \phi(s, f_1, g_{|S_2|.d}) \\ \phi(s, f_2, g_1) & \phi(s, f_2, g_2) & \cdots & \phi(s, f_2, g_{|S_2|.d}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(s, f_{|S_1|.d}, g_1) & \phi(s, f_{|S_1|.d}, g_2) & \cdots & \phi(s, f_{|S_1|.d}, g_{|S_2|.d}) \end{bmatrix}$$

Where $(f_1, f_2, \cdots, f_{|S_1|.d})$ and $(g_1, g_2, \cdots, g_{|S_2|.d})$ are the pure stationary strategies chosen by player-I and II respectively. In order to prove the existence of a pure semi-stationary strategy, we have to prove that this matrix has a pure saddle point for each initial state $s \in S$. Now by Shapley ([8], theorem 2.1 page-6) if A is the matrix of a two-person zero-sum game and if every 2×2 submatrix of A has a saddle point, then A has a saddle point. So, we concentrate only on a generic 2×2 submatrix and prove by contradiction that it has a saddle point. We consider the 2×2 submatrix:

$$\begin{bmatrix} \phi(s,f_i,g_j) & \phi(s,f_i,g_{j'}) \\ \phi(s,f_{i'},g_j) & \phi(s,f_{i'},g_{j'}) \end{bmatrix}$$

Where $i^{'},i \in \{1,2,\cdots,\mid S_1 \mid .d\},(i \neq i^{'})$ and $j,j^{'} \in \{1,2,\cdots,\mid S_2 \mid .d\},(j \neq j^{'})$. Now, by suitably renumbering the strategies, we can write the above sub-matrix as:

$$\begin{bmatrix} \phi(s,f_1,g_1) & \phi(s,f_1,g_2) \\ \phi(s,f_2,g_1) & \phi(s,f_2,g_2) \end{bmatrix}$$

Using the definition of $\phi(s,f_1,f_2)$ in section 3.1.3, we get that

$$\phi(s,f_i,g_j) = \sum_{s^{'} \in S} q^*(s^{'} \mid s,f_i,g_j) r(s^{'},f_i,g_j)$$

$$\Longrightarrow \phi(s,f_i,g_j) = \sum_{t=1}^{S_1}[q^*(t \mid s,f_{i.})r(t,f_{i.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s,g_{.j})r(v,g_{.j})] \qquad (*)$$

where

$$f_i(s,.) = \begin{cases} f_{i.}(s,.) & s \in S_1 \\ 1 & s \in S_2 \end{cases}$$

and

$$g_j(s,.) = \begin{cases} 1 & s \in S_1 \\ g_{.j}(s,.) & s \in S_2. \end{cases}$$

We consider the following two cases when $A(s)$ can not have a pure saddle point.

Case-1: $\phi(s,f_1,g_1)$ is row-minimum and column-minimum, $\phi(s,f_1,g_2)$ is row-maximum and column-maximum, $\phi(s,f_2,g_1)$ is row-maximum and column-maximum and $\phi(s,f_2,g_2)$ is row-minimum and column-minimum. These four conditions can be written as:

1. $\phi(s, f_1, g_1) < \phi(s, f_1, g_2)$.

2. $\phi(s, f_1, g_1) < \phi(s, f_2, g_1)$.

3. $\phi(s, f_2, g_2) < \phi(s, f_2, g_1)$.

4. $\phi(s, f_2, g_2) < \phi(s, f_1, g_2)$.

Thus using the equation $(*)$ we get the following inequalities:

$$\begin{aligned}
&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] && (3.4)\\
<&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})] \\
&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] && (3.5)\\
<&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] \\
&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})] && (3.6)\\
<&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] \\
&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})] && (3.7)\\
<&\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]
\end{aligned}$$

Hence, (3.4) yields

$$\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s, g_{.2})r(v, g_{.2}) - q^*(v \mid s, g_{.1})r(v, g_{.1}) > 0 \tag{3.8}$$

(3.6) yields

$$\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s, g_{.1})r(v, g_{.1}) - q^*(v \mid s, g_{.2})r(v, g_{.2}) > 0 \tag{3.9}$$

From (3.8) and (3.9) we clearly get a contradiction. Now we consider the next case:

Case-2: $\phi(s, f_1, g_1)$ is row-maximum and column-maximum, $\phi(s, f_1, g_2)$ is row-minimum and column-minimum, $\phi(s, f_2, g_1)$ is row-minimum and column-minimum and $\phi(s, f_2, g_2)$ is row-maximum and column-maximum. These four conditions can be written as:

1. $\phi(s, f_1, g_1) > \phi(s, f_1, g_2)$.
2. $\phi(s, f_1, g_1) > \phi(s, f_2, g_1)$.
3. $\phi(s, f_2, g_2) > \phi(s, f_2, g_1)$.
4. $\phi(s, f_2, g_2) > \phi(s, f_1, g_2)$.

We can re-write them as follows:

$$\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] \tag{3.10}$$
$$> \sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]$$

$$\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})] \tag{3.11}$$
$$> \sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]$$

$$\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})] \tag{3.12}$$
$$> \sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]$$

$$\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})] \tag{3.13}$$
$$> \sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})].$$

Hence, (3.10) yields

$$\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s, g_{.1}) r(v, g_{.1}) - q^*(v \mid s, g_{.2}) r(v, g_{.2}) > 0 \tag{3.14}$$

and (3.12) yields

$$\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s, g_{.2}) r(v, g_{.2}) - q^*(v \mid s, g_{.1}) r(v, g_{.1}) > 0 \tag{3.15}$$

From (3.14) and (3.15) we clearly get a contradiction. Thus, every 2×2 submatrix has a pure saddle point and by Shapley ([8], theorem 2.1, page-6) we conclude that the matrix $A(s)$ has a pure saddle point, namely (f_s, g_s) for some $s \in S$. Now $f^* = (f_1, f_2, \cdots, f_t, \cdots, f_z)$ and $g^* = (g_1, g_2, \cdots, g_t, \cdots, g_z)$ where f_t and g_t are the pure stationary strategies for the initial state t chosen by player-I and II respectively. As we know, optimality among pure stationary strategy space is equivalent to optimality in behavioural strategy space also, it is sufficient to consider only pure stationary strategy space to compute the optimal pure stationary strategies of the players. Thus, it only remains to prove the following lemma to prove the existence of pure stationary strategy pair which is optimal for the players:

Lemma 3.1.3. *Let us fix an initial state $t \in S$ in the PISG Γ. Suppose $(f_t, g_t) \in F_1^{sp} \times F_2^{sp}$ be an optimal pure stationary strategy pair of the players in $A(t)$ satisfying:*

$$\phi(t, f_t, g_t) \leq \phi(t, f_t, g) \forall g \in F_2^{sp} \textit{and for some initial state } \ t \in S. \tag{3.16}$$

Let us denote D_t to be the t-th row of the bi-matrix identifying the strategy pair (f_t, g_t), i.e., $D_t = ((f_t(t,1), g_t(t,1)) \cdots, (f_t(t,d), g_t(t,d))$, where d is the total number of pure actions in state t for both the players. Let $f^* = (f_1(1,1), f_2(2,2), \cdots, f_z(z,d))$ and $g^* = (g_1(1,1), g_2(2,2), \cdots, g_z(z,d))$. Then $(f^*, g^*) \in F_1^{sp} \times F_2^{sp}$ is a pure stationary strategy pair of the players identified by the bi-matrix D^* having D_t as its t-th row. We can write the bi-matrix D^* as:

$$D^*_{z \times d} = \begin{bmatrix} (f_1(1,1), g_1(1,1)) & (f_1(1,2), g_1(1,2)) & \cdots & (f_1(1,d), g_1(1,d)) \\ (f_2(2,1), g_2(2,1))) & (f_2(2,2), g_2(2,2)) & \cdots & (f_2(2,d), g_2(2,d)) \\ \vdots & \vdots & \ddots & \vdots \\ (f_z(z,1), g_z(z,1)) & (f_z(z,2), g_z(z,2)) & \cdots & (f_z(z,d), g_z(z,d)) \end{bmatrix}$$

and the pair (f^*, g^*) *satisfies:*

$$\phi(t, f^*, g^*) \leq \phi(t, f^*, g) \forall g \in F_2^{sp}, \forall t \in S.$$

Proof. For an initial state $t \in S(= \{1, 2, \cdots, z\})$ and a pair of behavioural strategy $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$ of the players, we consider the $z.d^2$ component vector:

$$\xi_n^{\pi_1\pi_2} = \{x_{n111}^t, x_{n112}^t, \cdots, x_{ns^{'}ab}^t, \cdots, x_{nzd^2}^t\}$$

where $x_{ns^{'}ab}^t = \frac{1}{n}\sum_{m=1}^n P_{\pi_1\pi_2}(X_m = s^{'}, A_m = a, B_m = b \mid X_0 = t)$.
Let $\xi^{\pi_1\pi_2}(t) = \lim_{n\to\infty} \xi_n^{\pi_1\pi_2}(t)$, whenever the limit exists and $\lim_{n\to\infty} x_{ns^{'}ab}^t = x_{s^{'}ab}^t$. Denote $\Theta(\xi_n^{\pi_1\pi_2}(t)) = \sum_{s^{'}\in S}\sum_{a\in A(s^{'})}\sum_{b\in B(s^{'})} x_{ns^{'}ab}^t.r(s^{'}, a, b)$. Then

$$\begin{aligned}\phi(t, \pi_1, \pi_2) = \liminf_{n\to\infty} \sum_{s^{'}\in S}\sum_{a\in A(s^{'})}\sum_{b\in B(s^{'})} x_{ns^{'}ab}^t.r(s^{'}, a, b) &= \liminf_{n\to\infty} [\xi_n^{\pi_1\pi_2}(t)].\bar{r} \\ &= \liminf_{n\to\infty} \Theta(\xi_n^{\pi_1\pi_2})(t)\end{aligned}$$

Where $\bar{r}$ is the reward vector of order zd^2. When (π_1, π_2) is a pure stationary strategy pair (f, g), the above limit exists and we can write

$$\phi(t, f, g) = \liminf_{n\to\infty} \Theta(\xi_n^{fg})(t) = \lim_{n\to\infty} \Theta(\xi_n^{fg})(t) \tag{3.17}$$

Define $\Theta(\xi^{fg}(t)) = \lim_{n\to\infty} \Theta(\xi_n^{fg}(t))$, considering that the limit exists for all pure stationary strategy pair $(f, g) \in F_1^{sp} \times F_2^{sp}$.

Let $p(s^{'} \mid t, f^*, g^*)$ be the transition probability from the state t to $s^{'}$ defined for the strategy pair (f^*, g^*). As this is a stochastic game, we can apply the Markov property that for any two states $x, y \in S$ and $m, n \in \mathbb{N}$,

$$\begin{aligned} p^{n+m}(x, y) &= P(X_{n+m} = y \mid X_0 = x) \\ &= \sum_{z\in S} P(X_n = z \mid X_0 = x) P(X_{n+m} = y \mid X_0 = x, X_n = z) \\ &= \sum_{z\in S} p^n(x, z) P(X_{n+m} = y \mid X_0 = x, X_n = z) \\ &= \sum_{z\in S} p^n(x, z) p^m(z, y) \end{aligned} \tag{3.18}$$

where $p^m(z,y)$ is the m-th step transition probability from the state z to y. Now using the above property and using the definition of $\xi^{f_t g_t}(t)$ we have

$$\begin{aligned}\Theta(\xi^{f_t g_t}(t)) &= \Theta(\sum_{s^{'}\in S} p(s^{'} \mid t,f^*,g^*)\xi^{f_t g_t}(s^{'})) \\ &= \sum_{s^{'}\in S} p(s^{'} \mid t,f^*,g^*)\Theta(\xi^{f_t g_t}(s^{'}))[\text{as } \Theta \text{ is a continuous function}] \end{aligned} \tag{3.19}$$

Now, as (f_t,g_t) is an optimal pure stationary strategy pair for the players when the initial state is t, we can write (3.19) as

$$\Theta(\xi^{f_t g_t}(t)) = \sum_{s^{'}\in S} p(s^{'} \mid t,f^*,g^*)\Theta(\xi^{f_{s'} g_{s'}}(s^{'})) \tag{3.20}$$

Now iterating (3.20) l times we get

$$\Theta(\xi^{f_t g_t}(t)) = \sum_{s^{'}\in S} p^l(s^{'} \mid t,f^*,g^*)\Theta(\xi^{f_{s'} g_{s'}}(s^{'})) \tag{3.21}$$

If we expand the right hand side of the above expression, the right hand side becomes:

$$p^l(1 \mid t,f^*,g^*)[\sum_{s\in S}\sum_{a\in A(s)}\sum_{b\in B(s)} r(s,a,b)x^1_{sab}]+\cdots+p^l(z \mid t,f^*,g^*)[\sum_{s\in S}\sum_{a\in A(s)}\sum_{b\in B(s)} r(s,a,b)x^z_{sab}] \tag{3.22}$$

Thus we can write (3.22) as:

$$\begin{aligned}\Theta(\xi^{f_t g_t}(t)) &= \sum_{s\in S}\sum_{a\in A(s)}\sum_{b\in B(s)} r(s,a,b).x^t_{sab} \\ &= \Theta(\xi^{f^* g^*}(t)) \end{aligned} \tag{3.23}$$

Thus from (3.16), (3.17) and (3.23) we get that

$$\phi(t,f^*,g^*) \le \phi(t,f^*,g) \forall t \in S \text{ and } \forall g \in F_2^{sp}.$$

□

By similar manner we can show that $\phi(t,f^*,g^*) \ge \phi(t,f,g^*) \forall t \in S$ and $\forall f \in F_1^{sp}$. Thus the pair (f^*,g^*) is the optimal pure stationary strategy pair of the players in the PISG Γ. □

3.1.6 Algorithm To Solve An Undiscounted Zero-Sum Two Person Perfect Information Stochastic Game

Let Γ be a zero-sum two person perfect information stochastic game. We consider the following policy-improvement algorithm to compute optimal stationary strategies of the players. This is a best response algorithm, in which each player looks for his own undiscounted optimal strategy. The algorithm is stated below:
Step 1: Choose a random pure stationary strategy for player-II g_k.
Step 2: Find the undiscounted optimal pure stationary strategy f_k for player-I in the MDP $\Gamma(g_k)$.
Step 3: if g_k is undiscounted optimal pure stationary strategy for player-II in $\Gamma(f_k)$, set $(f^*, g^*) = (f_k, g_k)$ and stop.
Step 4: else find the best response undiscounted optimal pure stationary strategy g_{k+1} for player-II in the MDP $\Gamma(f_k)$, set $k = k+1$ and go to step 2. As the set of pure stationary strategy pair of the players is a finite set, the algorithm stops in finite time.
Note: The process of finding an undiscounted optimal strategy for an undiscounted MDP $\hat{\Gamma} =< S, \hat{A} = \{A(s) : s \in S\}, \hat{q}, \hat{r} >$ was proposed by Hordijk et al.(1985) [14]. It consists of a linear programming problem with several parameters as given below:

$$\max \textstyle\sum_{s=1}^{z} \sum_{a\in A(s)} r(s,a) w_{sa}$$

subject to:

$$\begin{gathered}\textstyle\sum_{s=1}^{z} \sum_{a\in A(s)} (\delta(s,s^{'}) - q(s^{'} \mid s,a)) w_{sa} = 0,\ s^{'} \in S \\ \textstyle\sum_{a\in A(s)} w_{sa} + \sum_{s=1}^{z} \sum_{a\in A(s)} (\delta(s,s^{'}) - q(s^{'} \mid s,a)) y_{sa} = \beta_s,\ s^{'} \in S \\ w_{sa} \geq 0\ \forall s \in S \text{ and } a \in A(s).\end{gathered}$$

where $\beta_s > 0$ are given numbers for each $s \in S$, such that $\sum_{s\in S} \beta_s = 1$. The undiscounted optimal pure stationary strategy is computed as:

$$f^*(s) = \frac{w^*_{sa}}{\sum_{a\in A(s) w^*_{sa}}}$$

where w^*_{sa} is the optimal solution of the above LP. By Hordijk et al.[14], this pure stationary strategy is average optimal as well. We elaborate the above algorithm by following examples:

3.1.7 Numerical Examples

Example 3.1. *Consider a PISG* Γ *with three states* $S = \{1,2,3\}$, $A(1) = \{1,2\} = A(2)$, $A(3) = \{1\}$, $B(1) = \{1\} = B(2)$ *and* $B(3) = \{1,2,3\}$. *In this example player-I is a dummy player here for the state* 3 *and player-II is dummy for states* 1 *and* 2. *Rewards and transition probabilities for the players are given below*

$$\textit{State-1: }\begin{array}{|r|}\hline 5\\ (\frac{1}{2},\frac{1}{2},0)\\ \hline 7\\ (0,1,0)\\ \hline\end{array}\quad \textit{State-2: }\begin{array}{|r|}\hline 1\\ (\frac{1}{3},0,\frac{2}{3})\\ \hline 0.5\\ (0,0,1)\\ \hline\end{array}\quad \textit{State-3: }\begin{array}{|c|c|c|}\hline 3 & 4 & 2\\ (0,\frac{1}{2},\frac{1}{2}) & (1,0,0) & (\frac{1}{2},\frac{1}{4},\frac{1}{4})\\ \hline\end{array}$$

where a cell $\begin{array}{|r|}\hline r\\ (q_1,q_2,q_3)\\ \hline\end{array}$ *represents that* r *is the immediate reward function and* (q_1,q_2,q_3) *are the transition probabilities that the next states are* 1, 2 *and* 3 *respectively if this cell is chosen at present state. The pure strategies for player-I are:* $f_0 = \{(1,0),(1,0),1\}$, $f_1 = \{(1,0),(0,1),1\}$, $f_2 = \{(0,1),(1,0)\}$,$f_3 = \{(0,1),(0,1)\}$. *The pure strategies of player-II are:* $g_0 = \{1,1,(1,0,0)\}$, $g_1 = \{1,1,(0,1,0)\}$, $g_2 = \{1,1,(0,0,1)\}$. *Firstly we fix the strategy* g_1 *of the player-II in* Γ. *Thus we get a reduced MDP* $\Gamma(g_1)$ *given below:*

$$\textit{State-1: }\begin{array}{|r|}\hline 5\\ (\frac{1}{2},\frac{1}{2},0)\\ \hline 7\\ (0,1,0)\\ \hline\end{array}\quad \textit{State-2: }\begin{array}{|r|}\hline 1\\ (\frac{1}{3},0,\frac{2}{3})\\ \hline 0.5\\ (0,0,1)\\ \hline\end{array}\quad \textit{State-3: }\begin{array}{|r|}\hline 4\\ (1,0,0)\\ \hline\end{array}$$

Now we formulate the following linear programming problem in the variables $x = (x_{11},x_{12},x_{21},x_{22},x_{31})$ *and* $y = (y_{11},y_{12},y_{21},y_{22},y_{31})$ *to obtain player-I's undiscounted optimal strategy:*

$$\max R = 5x_{11} + 7x_{12} + x_{21} + 0.5x_{22} + 4x_{31}$$

subject to

$$3x_{11}+6x_{12}-2x_{21}-6x_{31}=0 \tag{3.24}$$

$$-3x_{11}-6x_{12}+6x_{21}+6x_{22}=0 \tag{3.25}$$

$$-4x_{21}-6x_{22}+6x_{31}=0 \tag{3.26}$$

$$6x_{11}+6x_{12}+3y_{11}+6y_{12}-2y_{21}-6y_{31}=6\beta_1 \tag{3.27}$$

$$2x_{21}+2x_{22}-y_{11}-2y_{12}+2y_{21}+2y_{22}=2\beta_2 \tag{3.28}$$

$$12x_{31}-8y_{21}-12y_{22}+12y_{31}=12\beta_3 \tag{3.29}$$

$$x,y \geq 0. \tag{3.30}$$

We fix $\beta_1=\beta_2=\beta_3=\frac{1}{3}$*. The solution of the above linear programming problem by dual-simplex method is given below:*
$\max R=4$*,* $x=(0,0.375,0.375,0,0.25)$*,* $y=(0,0.111,0,0.111,0)$*.*
Now by the method to compute optimal pure stationary strategy described in section 3.1.6*, we get that* $f^0=\{(0,1),(1,0),1\}$ *is the optimal pure stationary strategy for player-I in* $\Gamma(g_1)$*. Now we fix this strategy for player-I. Thus we get a resultant MDP as follows:*

$$\textit{State-}1\textit{:}\ \begin{array}{|r|}\hline 7 \\ (0,1,0) \\ \hline \end{array} \quad \textit{State-}2\textit{:}\ \begin{array}{|r|}\hline 1 \\ (\frac{1}{3},0,\frac{2}{3}) \\ \hline \end{array} \quad \textit{State-}3\textit{:}\ \begin{array}{|r|}\hline 3 \\ (0,\frac{1}{2},\frac{1}{2}) \\ \hline 4 \\ (1,0,0) \\ \hline 2 \\ (\frac{1}{2},\frac{1}{4},\frac{1}{4}) \\ \hline \end{array}$$

We formulate the linear programming problem of the above MDP for the variables $x=(x_{11},x_{21},x_{31},x_{32},x_{33})$ *and* $y=(y_{11},y_{21},y_{31},y_{32},y_{33})$ *as follows:*

$$\min R' = 7x_{11}+x_{21}+3x_{31}+4x_{32}+2x_{33}$$

subject to

$$6x_{11}-2x_{21}-6x_{32}-3x_{33}=0 \tag{3.31}$$

$$-4x_{11}+4x_{21}-2x_{31}-x_{33}=0 \tag{3.32}$$

$$-8x_{21}+6x_{31}+12x_{32}+9x_{33}=0 \tag{3.33}$$

$$6x_{11}+3y_{11}-2y_{21}-6y_{32}-3y_{33}=6\beta_1 \tag{3.34}$$

$$4x_{21}-4y_{11}+4y_{21}-2y_{31}-y_{33}=4\beta_2 \tag{3.35}$$

$$12x_{31}+12x_{32}+12x_{33}-8y_{21}+6y_{31}+12y_{32}+9y_{33}=12\beta_3 \tag{3.36}$$

$$x,y \geq 0. \tag{3.37}$$

The solution of the above LP by dual-simplex method is given below: $\min R^{'}=2.4667,\ x=(0.2333,0.3,0.17866,0,0),\ y=(0.333,0.1667,0,0,0)$. *So by the same method described in section* 3.1.6*, we compute the optimal pure stationary strategy for player-II as:* $g^0=\{1,1,(1,0,0)\}$. *As we observe* g_1 *is not optimal in* $\Gamma(f_0)$*, we now fix the strategy* g^0 *for player-II in the PISG* Γ. *Thus, we get a reduced MDP* $\Gamma(g^0)$ *which is given below:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline 5 \\ (\frac{1}{2},\frac{1}{2},0) \\ \hline 7 \\ (0,1,0) \\ \hline \end{array} \quad \textit{State-2:}\ \begin{array}{|r|}\hline 1 \\ (\frac{1}{3},0,\frac{2}{3}) \\ \hline 0.5 \\ (0,0,1) \\ \hline \end{array} \quad \textit{State-3:}\ \begin{array}{|r|}\hline 3 \\ (0,\frac{1}{2},\frac{1}{2}) \\ \hline \end{array}$$

Now we formulate the following linear programming problem in the variables $x=(x_{11},x_{12},x_{21},x_{22},x_{31})$ *and* $y=(y_{11},y_{12},y_{21},y_{22},y_{31})$ *to obtain player-I's undiscounted optimal strategy:*

$$\max R=5x_{11}+7x_{12}+x_{21}+0.5x_{22}+3x_{31}$$

subject to

$$3x_{11}+6x_{12}-2x_{21}=0 \tag{3.38}$$

$$-3x_{11}-6x_{12}+6x_{21}+6x_{22}-3x_{31}=0 \tag{3.39}$$

$$-8x_{21}-12x_{22}+6x_{31}=0 \tag{3.40}$$

$$6x_{11}+6x_{12}+3y_{11}+6y_{12}-2y_{21}=6\beta_1 \tag{3.41}$$

$$2x_{21}+2x_{22}-y_{11}-2y_{12}+2y_{21}+2y_{22}-y_{31}=2\beta_2 \tag{3.42}$$

$$12x_{31}-8y_{21}-12y_{22}+6y_{31}=12\beta_3 \tag{3.43}$$

$$x,y\geq 0. \tag{3.44}$$

We fix $\beta_1=\beta_2=\beta_3=\frac{1}{3}$*. The solution of the above linear programming problem by dual-simplex method is given below:*
$\max R=2.778,\ x=(0.222,0,0.333,0,0.444),\ y=(0,0.111,0,0.111,0)$.
Now by the method to compute optimal pure stationary strategy described in section 3.1.6*, we get that* $f^1=\{(1,0),(1,0),1\}$ *is the optimal pure stationary strategy for player-I in* $\Gamma(g^0)$*. Now we fix the strategy for player-I in the PISG* Γ*. The resultant MDP is given below:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline 5\\ (\frac{1}{2},\frac{1}{2},0)\\ \hline\end{array}\quad \textit{State-2:}\ \begin{array}{|r|}\hline 1\\ (\frac{1}{3},0,\frac{2}{3})\\ \hline\end{array}\quad \textit{State-3:}\ \begin{array}{|r|}\hline 3\\ (0,\frac{1}{2},\frac{1}{2})\\ \hline 4\\ (1,0,0)\\ \hline 2\\ (\frac{1}{2},\frac{1}{4},\frac{1}{4})\\ \hline\end{array}$$

We formulate the linear programming problem of the above MDP for the variables $x=(x_{11},x_{21},x_{31},x_{32},x_{33})$ *and* $y=(y_{11},y_{21},y_{31},y_{32},y_{33})$ *as follows:*

$$\min R^{'}=5x_{11}+x_{21}+3x_{31}+4x_{32}+2x_{33}$$

subject to

$$3x_{11}-2x_{21}-6x_{32}-3x_{33}=0 \tag{3.45}$$

$$-2x_{11}+4x_{21}-2x_{31}-x_{33}=0 \tag{3.46}$$

$$-8x_{21}+6x_{31}+12x_{32}+9x_{33}=0 \tag{3.47}$$

$$6x_{11}+3y_{11}-2y_{21}-6y_{32}-3y_{33}=6\beta_1 \tag{3.48}$$

$$4x_{21}-2y_{11}+4y_{21}-2y_{31}-y_{33}=4\beta_2 \tag{3.49}$$

$$12x_{31}+12x_{32}+12x_{33}-8y_{21}+6y_{31}+12y_{32}+9y_{33}=12\beta_3 \tag{3.50}$$

$$x,y\geq 0. \tag{3.51}$$

The solution of the above LP by dual-simplex method is given below: $\min R^{'}=2.778$, $x=(0.222,0.333,0.444,0,0)$, $y=(0.333,0.1667,0,0,0)$. *So by the same method described in section* 3.1.6, *we compute the optimal pure stationary strategy for player-II as:* $g^1=\{1,1,(1,0,0)\}$.

Thus the algorithm stops in this step and we get the optimal pure (limiting average) stationary strategy $(F^*,G^*)=(f_0,g_0)$.

Example 3.2. *Consider a PISG* Γ *with four states* $S=\{1,2,3,4\}$, $A(1)=\{1,2\}=A(2)$, $A(3)=\{1\}$, $B(1)=\{1\}=B(2)$ *and* $B(3)=\{1,2\}=B(4)$. *In this example player-I is a dummy player here for the state* 3 *and* 4 *and player-II is a dummy player for states* 1 *and* 2. *Rewards and transition probabilities for the players are given below:*

State-1:

2 $(\frac{1}{2},\frac{1}{2},0,0)$
3 $(0,1,0,0)$

State-2:

1 $(\frac{1}{3},0,\frac{2}{3},0)$
0.5 $(0,0,1,0)$

State-3:

5 $(0,0,\frac{1}{2},\frac{1}{2})$	0 $(0,0,1,0)$

State 4:

11 $(\frac{1}{2},0,\frac{1}{2},0)$	12 $(1,0,0,0)$

where a cell $\boxed{\begin{array}{r} r \\ (q_1,q_2,q_3,q_4) \end{array}}$ *represents that* r *is the immediate reward function and* (q_1,q_2,q_3,q_4) *are the transition probabilities that the next states are* 1, 2, 3 *and* 4 *respectively if this cell is chosen at present state. The pure strategies for player-I are:* $f_0=\{(1,0),(1,0),1,1\}$, $f_1=\{(1,0),(0,1),1,1\}$, $f_2=\{(0,1),(1,0),1,1\}$,$f_3=$

$\{(0,1),(0,1),1,1\}$. *The pure strategies of player-II are:* $g_0 = \{1,1,(1,0),(1,0)\}$, $g_1 = \{1,1,(1,0),(0,1)\}$, $g_2 = \{1,1,(0,1),(1,0)\}$ *and* $g_3 = \{1,1,(0,1),(0,1)\}$. *Firstly set* $k=0$ *and we fix the strategy* g_0 *of the player-II in* Γ. *Thus we get a reduced MDP* $\Gamma(g_0)$ *given below:*

State-1:

2 $(\frac{1}{2},\frac{1}{2},0,0)$
3 $(0,1,0,0)$

State-2:

1 $(\frac{1}{3},0,\frac{2}{3},0)$
0.5 $(0,0,1,0)$

State-3:

5 $(0,0,\frac{1}{2},\frac{1}{2})$

State 4:

11 $(\frac{1}{2},0,\frac{1}{2},0)$

Now we formulate the following linear programming problem in the variables $x = (x_{11},x_{12},x_{21},x_{22},x_{31},x_{32},x_{41},x_{42})$ *and* $y = (y_{11},y_{12},y_{21},y_{22},y_{31},y_{32},y_{41},y_{42})$ *to obtain player-I's undiscounted optimal strategy:*

$$\max R = 2x_{11} + 3x_{12} + x_{21} + 0.5x_{22} + 5x_{31} + 11x_{41}$$

subject to

$$3x_{11} + 6x_{12} - 2x_{21} - 3x_{41} = 0 \tag{3.52}$$

$$-3x_{11} - 6x_{12} + 6x_{21} + 6x_{22} = 0 \tag{3.53}$$

$$-4x_{21} - 6x_{22} + 3x_{31} - 3x_{41} = 0 \tag{3.54}$$

$$-3x_{31} + 6x_{41} = 0 \tag{3.55}$$

$$6x_{11} + 6x_{12} + 3y_{11} + 6y_{12} - 2y_{21} - 3y_{41} = 6\beta_1 \tag{3.56}$$

$$2x_{21} + 2x_{22} - 3y_{11} - 6y_{12} + 6y_{21} + 6y_{22} = 2\beta_2 \tag{3.57}$$

$$6x_{31} - 4y_{21} - 6y_{22} + 3y_{31} - 3y_{41} = 12\beta_3 \tag{3.58}$$

$$2x_{41} - x_{31} + 2x_{41} = 2\beta_4 \tag{3.59}$$

$$x, y \geq 0. \tag{3.60}$$

We fix $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \frac{1}{4}$. *The solution of the above linear programming problem by dual-simplex method is given below:*

$\max R = 5.6875$, $x = (0, 0.1250, 0, 0.1250, 0.5000, 0.2500)$, $y = (0, 0.1250, 0, 0.2500, 0, 0)$. *Now by the method to compute optimal pure stationary strategy described in section* 3.1.6*, we get that* $f^0 = \{(0,1),(0,1),1,1\}$ *is the optimal pure stationary strategy for player-I in* $\Gamma(g_0)$. *Now we fix this strategy for player-I. Thus we get a resultant MDP as follows:*

State-1:

3 $(0,1,0,0)$

State-2:

0.5 $(0,0,1,0)$

State-3:

5 $(0,0,\frac{1}{2},\frac{1}{2})$	0 $(0,0,1,0)$

State 4:

11 $(\frac{1}{2},0,\frac{1}{2},0)$	12 $(1,0,0,0)$

Now we formulate the following linear programming problem in the variables $x = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32}, x_{41}, x_{42})$ *and* $y = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32}, y_{41}, y_{42})$ *to obtain player-I's undiscounted optimal strategy:*

$$\min R = 3x_{11} + 0.5x_{21} + 5x_{31} + 11x_{41} + 12x_{42}$$

subject to

$$2x_{11} - x_{41} - 2x_{42} = 0 \tag{3.61}$$

$$-x_{11} + x_{21} = 0 \tag{3.62}$$

$$-2x_{21} + x_{31} - x_{41} = 0 \tag{3.63}$$

$$-x_{31} + 2x_{41} + 2x_{42} = 0 \tag{3.64}$$

$$2x_{11} + 2y_{11} - y_{41} - 2y_{42} = 2\beta_1 \tag{3.65}$$

$$x_{21} - y_{11} + y_{21} = \beta_2 \tag{3.66}$$

$$2x_{31} + 2x_{32} - 2y_{21} + y_{31} - y_{41} = 2\beta_3 \tag{3.67}$$

$$x_{41} + x_{42} - y_{31} + 2y_{41} + y_{42} = 2\beta_4 \tag{3.68}$$

$$x, y \geq 0. \tag{3.69}$$

The solution of the above LP by dual-simplex method is given below:
$\min R' = 5.6875$, $x = (1, 0.1250, 0.5, 0, 0.011, 0)$, $y = (0.1250, 0.2500, 0, 0, 0, 0, 0)$. *So by the same method described in section* 3.1.6*, we compute the optimal pure stationary strategy for player-II as:* $g_0 = \{1, 1, (1,0), (1,0))\}$. *Thus the algorithm stops in this step and we get the optimal pure (limiting average) stationary strategy* $(F^*, G^*) = (f_2, g_0)$.

3.2 Undiscounted Perfect Information Semi-Markov Games

3.2.1 Introduction

A semi-Markov game (SMG) is a generalisation of a Stochastic (Markov) game (Shapley(1953) [42]). Such games have already been studied in the literature (e.g. Lal-Sinha(1992) [22], Luque-Vasquez(1999) [28], Mondal(2015) [34]). Single player SMGs are called semi-Markov decision processes (SMDPs) which were introduced by Jewell(1963) [18] and Howard(1971) [17]. A perfect information semi-Markov game (PISMG) is a natural extension of perfect information stochastic games (PISGs) (Raghavan et al.(1997) [45]), where at each state all but one player is a dummy (i.e., he has only one available action in that state). Note that for such a game, perfect information is a state property. In this paper, we prove that such games (PISMGs) have a value and both players have pure semi-stationary optimal strategies under undiscounted (limiting ratio average) pay-offs. We prove this by showing the existence of a pure saddle point in the pay-off matrix of the game for each initial state. This section of chapter 3 is organised as follows. Section 3.2.2 contains definitions and properties of an undiscounted two person zero-sum semi-Markov game considered under limiting ratio average pay-off. Section 3.2.5 contains main result of this paper. In Section 3.2.6 we state the algorithm to compute a Cesaro limiting matrix of a transition matrix, proposed by Lazari et al. [23]. In section 3.2.7 we give a policy improvement algorithm to solve an undiscounted zero-sum two person perfect information semi-Markov game. section 3.2.8 contains a numerical example illustrating our theorem. Section 3.2.9 is reserved for the conclusion.

3.2.2 Preliminaries

A zero-sum two-person finite SMG is described by a collection of objects $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, P, r >$, where $S = \{1, 2, \cdots, z\}$ is the finite non-empty state space and $A(s) = \{1, 2, \cdots, m_s\}, B(s) = \{1, 2, \cdots, n_s\}$ are respectively the non-empty sets of admissible actions of the players I and II respectively in the state s. Let us denote $K = \{(s, i, j) : s \in S, i \in A(s), j \in B(s)\}$ to be the set of admissible triplets. For each $(s, i, j) \in K$, we denote $q(. \mid s, i, j)$ to be the transition law of the game. Given $(s, i, j) \in K$ and $s^{'} \in S$, let $\tau_{ss^{'}}^{ij}$ be the transition time random variable which denotes the time for a transition to a state $s^{'}$ from a state s by a pair of actions $(i, j) \in A(s) \times B(s)$. Let $P_{ss^{'}}^{ij} = Prob(\tau_{ss^{'}}^{ij} \leq t)$ for each $(s, i, j) \in K, s^{'} \in S$ be a probability distribution function on $[0,\infty)$ and it is called the conditional transition time distribution function. Finally r is the real valued functions on K, which represents the immediate (expected) rewards for the player-I (and $-r$ is the immediate reward for player-II). Let us consider player-I as the maximiser and player-II as the minimiser in the zero-sum two person SMG. The semi-Markov game over infinite time is played as follows. At the 1st decision epoch, the game starts at $s_1 \in S$ and the players I and II simultaneously and independently choose actions $i_1 \in A(s_1)$ and $j_1 \in B(s_1)$ respectively. Consequently player I and II get immediate rewards $r(s_1,i_1,j_1)$ and $-r(s_1,i_1,j_1)$ respectively and the game moves to the state s_2 with probability $q(s_2 \mid s_1,i_1,j_1)$. The sojourn time to move from state s_1 to the state s_2 is determined by the distribution function $P^{s_1 s_2}_{i_1 j_1}(.)$. After reaching the state s_2 on the next decision epoch, the game is repeated over infinite time with the state s_1 replaced by s_2.

By a strategy (behavioural) π_1 of the player I, we mean a sequence $\{(\pi_1)_n(. \mid hist_n)\}_{n=1}^{\infty}$, where $(\pi_1)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n{=}(s_1,a_1,b_1,s_2,a_2,b_2\cdots,s_{n-1},a_{n-1},b_{n-1},s_n)$ for $n \geq 2$, $hist_1 = (s_1)$ and $(s_k,a_k,j_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi_1)_n(. \mid hist_n)$ on $A(s_n)$. Behavioural strategy π_2 for player II can be defined analogously. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π_1 and Π_2 to be the sets of strategies (behavioural) of the players I and II respectively.

A strategy $f = \{f_n\}_{n=1}^{\infty}$ for the player I is called semi-Markov if for each n, f_n depends on s_1, s_n and the decision epoch number n. Similarly we can define a semi-Markov strategy $g = \{g_n\}_{n=1}^{\infty}$ for the player II.

A stationary strategy is a strategy that depends only on the current state. A stationary strategy for player I is defined as z tuple $f = (f(1), f(2), \cdots, f(z))$, where each $f(s)$ is the probability distribution on $A(s)$ given by $f(s) = (f(s,1), f(s,2), \cdots, f(s,m_s))$. $f(s,i)$ denotes the probability of choosing action i in the state s. By similar manner, one can define a stationary strategy g for player II as $g = (g(1), g(2), \cdots, g(z))$ where each $g(s)$ is the probability distribution on $B(s)$. Let us denote F_1^s and F_2^s to be the set of stationary strategies for player I and II respectively.

A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F_1^{sp} and F_2^{sp} to be the set of pure stationary strategies of the players I and II respectively.

A semi-stationary strategy is a semi-Markov strategy which is independent of the decision epoch n, i.e., for a initial state s_1 and present state s_n and decision epoch n, if a semi-Markov strategy $g(s_1, s_n, n)$ turns out to be independent of n, then we call it a semi-stationary strategy. Let ξ_1^s and ξ_2^s denote the set of semi-stationary strategies for the players I and II respectively and ξ_1^{sp} and ξ_2^{sp} denote the set of pure semi-stationary strategies for the players I and II respectively.

3.2.3 Zero-Sum Two-Person Semi-Markov Games Under Limiting Ratio Average (Undiscounted) Pay-off

Let $(X_1, A_1, B_1, X_2, A_2, B_2 \cdots)$ be a co-ordinate sequence in $S \times (A \times B \times S)^\infty$. Given behavioural strategy pair $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi_1\pi_2}(. \mid X_1 = s)$ (hence an expectation $\mathbb{E}_{\pi_1\pi_2}(. \mid X_1 = s)$) on the product σ- field of $S \times (A \times B \times S)^\infty$ by Kolmogorov's extension theorem. For a pair of strategies $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$ for the players I and II respectively, the limiting ratio average (undiscounted) pay-off for player I, starting from a state $s \in S$ is defined by:

$$\phi(s, \pi_1, \pi_2) = \liminf_{n\to\infty} \frac{\mathbb{E}_{\pi_1\pi_2} \sum_{m=1}^{n} [r(X_m, A_m, B_m) | X_1 = s]}{\mathbb{E}_{\pi_1\pi_2} \sum_{m=1}^{n} [\bar{\tau}(X_m, A_m, B_m) | X_1 = s]}.$$

Here $\bar{\tau}(s, i, j) = \sum_{s' \in S} q(s^{'} \mid s, i, j) \int_0^\infty t dP_{ij}^{ss^{'}}(t)$ is the expected sojourn time in the state s for a pair of actions $(i, j) \in A(s) \times B(s)$. We assume that $\bar{\tau}(s, i, j)$ is bounded away from zero for all $(s, i, j) \in K$.

Definition 3.4. *For each pair of stationary strategies $(f, g) \in F_1^s \times F_2^s$ we define the transition probability matrix as $Q(f, g) = [q(s^{'} \mid s, f, g)]_{z \times z}$, where $q(s^{'} \mid s, f, g) = \sum_{i \in A(s)} \sum_{j \in B(s)} q(s^{'} \mid s, i, j) f(s, i) g(s, j)$ is the probability that starting from the state s, next state is $s^{'}$ when the players choose strategies f and g respectively (For a stationary strategy f, $f(s, i)$ denotes the probability of choosing action i in the state s).*

For any pair of stationary strategies $(f, g) \in F_1^s \times F_2^s$ of player I and II, we write the undiscounted pay-off for player I as:

$$\phi(s, f, g) = \lim_{n\to\infty} \frac{\sum_{m=1}^{n} r^m(s, f, g)}{\sum_{m=1}^{n} \bar{\tau}^m(s, f, g)} \text{ for all } s \in S.$$

Where $r^m(s, f, g)$ and $\bar{\tau}^m(s, f, g)$ are respectively the expected reward and expected sojourn time for player-I at the m-th decision epoch, when player-I chooses f and player-II chooses g respectively and the initial state is s.

We define $r(f,g)=[r(s,f,g)]_{z\times 1}$, $\bar{\tau}(f,g)=[\bar{\tau}(s,f,g)]_{z\times 1}$ and $\phi(f,g)=[\phi(s,f,g)]_{z\times 1}$ as expected reward, expected sojourn time and undiscounted pay-off vector for a pair of stationary strategy $(f,g)\in F_1^s\times F_2^s$. Now

$$\begin{aligned} r^m(s,f,g) \quad &= \textstyle\sum_{s'\in S}\mathbb{P}_{fg}(X_m=s^{'}\mid X_1=s)r(s^{'},f,g)\\ &= \textstyle\sum_{s'\in S} r(s^{'},f,g)q^{m-1}(s^{'}\mid s,f,g)\\ &= [Q^{m-1}(f,g)r(f,g)](s)\end{aligned}$$

and

$$\begin{aligned} \bar{\tau}^m(s,f,g) \quad &= \textstyle\sum_{s'\in S}\mathbb{P}_{fg}(X_m=s^{'}\mid X_1=s)\bar{\tau}(s^{'},f,g)\\ &= \textstyle\sum_{s'\in S} \bar{\tau}(s^{'},f,g)q^{m-1}(s^{'}\mid s,f,g)\\ &= [Q^{m-1}(f,g)\bar{\tau}(f,g)](s)\end{aligned}$$

Since $Q(f,g)$ is a Markov matrix, we have by Kemeny et al., [21]

$$\lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{m=1}^{n}Q^m(f,g) \text{ exists and equals to } Q^*(f,g).$$

It is obvious that

$$\lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{m=1}^{n}r^m(f,g)=[Q^*(f,g)r(f,g)](s)$$

and

$$\lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{m=1}^{n}\bar{\tau}^m(f,g)=[Q^*(f,g)\bar{\tau}(f,g)](s).$$

Thus we have for any pair of stationary strategies $(f_1,f_2)\in F_1\times F_2$,

$$\phi(s,f,g)=\tfrac{[Q^*(f,g)r(f,g)](s)}{[Q^*(f,g)\bar{\tau}(f,g)](s)} \text{ for all } s\in S$$

where $Q^*(f,g)$ is the Cesaro limiting matrix of $Q(f,g)$.

Definition 3.5. *A zero-sum two person undiscounted semi-Markov game is said to have a value vector $\phi=[\phi(s)]_{z\times 1}$ if*

$$\sup_{\pi_1\in\Pi_1}\inf_{\pi_2\in\Pi_2}\phi(s,\pi_1,\pi_2)=\phi(s)=\inf_{\pi_2\in\Pi_2}\sup_{\pi_1\in\Pi_1}\phi(s,\pi_1,\pi_2) \textit{ for all } s\in S.$$

A pair of strategies $(\pi_1^*,\pi_2^*)\in\Pi_1,\times\Pi_2$ is said to be an optimal strategy pair for the players if

$$\phi(s,\pi_1^*,\pi_2)\geq\phi(s)\geq\phi(s,\pi_1,\pi_2^*) \textit{ for all } s\in S \textit{ and all } (\pi_1,\pi_2)\in\Pi_1\times\Pi_2.$$

Note: Throughout this paper, we use the notion of undiscounted pay-off as limiting ratio average pay-off.

3.2.4 Zero-Sum Two Person Perfect Information Semi-Markov Games

A zero-sum two person SMG is denoted by $\Gamma =< S, \{A(s) : s \in S\}, \{B(s) : s \in S\}, q, P, r >$, where $S = \{1, 2, \cdots, z\}$ is the finite non-empty state space and $A(s) = \{1, 2, \cdots, m_s\}$, $B(s) = \{1, 2, \cdots, n_s\}$ are respectively the non-empty sets of admissible actions of the players I and II respectively in the state s. Let us denote $K = \{(s, i, j) : s \in S, i \in A(s), j \in B(s)\}$ to be the set of admissible triplets. For each $(s, i, j) \in K$, we denote $q(. \mid s, i, j)$ to be the transition law of the game whereas $P_{ss'}(. \mid i, j)$ is a distribution function on $[0, \infty)$ given $K \times S$, which is called the conditional transition (sojourn) time distribution. Finally r and $-r$ are real valued functions on K, which represent the immediate (expected) rewards for the players I and II respectively. Let us consider player-I as the maximiser and player-II as the minimiser in the zero-sum two person perfect information SMG. Γ is called a perfect information semi-Markov game (PISMG) if the following properties hold
(i)$S = S_1 \cup S_2, S_1 \cap S_2 = \phi$.
(ii)$\mid B(s) \mid= 1$, for all $s \in S_1$, i.e., on S_1 player-II is a dummy (i.e., for states $\{1, 2, \cdots,$ $\mid S_1 \mid\}$ player-II has only 1 action).
(iii)$\mid A(s) \mid= 1$, for all $s \in S_2$, i.e., on S_2 player-I is a dummy (i.e., for states $\{\mid S_1 \mid +1,$ $\mid S_1 \mid +2, \cdots, \mid S_1 \mid + \mid S_2 \mid\}$ player-I has only 1 action).

3.2.5 Main Result

Theorem 3.2.1. *Any zero-sum two person undiscounted perfect information semi-Markov game has a solution in pure semi-stationary strategies under limiting ratio average pay-offs.*

Proof. Let $\Gamma =< S = S_1 \cup S_2, A = \{A(s) : s \in S_1\}, B = \{B(s) : s \in S_2\}, q, P, r >$ be a zero-sum two person perfect information semi-Markov game under limiting ratio average pay-off, where $S = \{1, 2, , \cdots z\}$ is the finite state space. Let us fix an initial state $s \in S$. We assume that in $\mid S_1 \mid$ number of states (i.e., states $\{1, 2, \cdots, S_1\}$), player-II is a dummy and from states $\{\mid S_1 \mid +1, \cdots, \mid S_1 \mid + \mid S_2 \mid\}$ player-I is a dummy. We assume that in this perfect information game, player-I has $d_1, d_2, \cdots, d_{S_1}$ number of pure actions in the states where he is non-dummy and similarly player-II has $t_{S_1+1}, t_{S_1+2}, \cdots, t_{S_1+S_2}$ number of pure actions available in the states where he is non-dummy. Let $D_1 = \Pi_{i=1}^{S_1} d_i$ and $D_2 = \Pi_{i=S_1+1}^{S_1+S_2} t_i$. Let us the consider the pay-off matrix

$$A_{D_1 \times D_2} = \begin{bmatrix} \phi(s, f_1, g_1) & \phi(s, f_1, g_2) & \cdots & \phi(s, f_1, g_{D_2}) \\ \phi(s, f_2, g_1) & \phi(s, f_2, g_2) & \cdots & \phi(s, f_2, g_{D_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(s, f_{D_1}, g_1) & \phi(s, f_{D_1}, g_2) & \cdots & \phi(s, f_{D_1}, g_{D_2}) \end{bmatrix}$$

Where $(f_1, f_2, \cdots, f_{D_1})$ and $(g_1, g_2, \cdots, g_{D_2})$ are the pure stationary strategies chosen by player-I and II respectively. In order to prove the existence of a pure semi-stationary strategy, we have to prove that this matrix has a pure saddle point for each initial state $s \in S$. Now by theorem 2.1 ("Some topics in two-person games", in the Advances in Game Theory.(AM-52), Volume 52, 1964, page-6) proposed by Shapley [7], we know that, if A is the pay-off matrix of a two-person zero-sum game and if every 2×2 sub-matrix of A has a saddle point, then A has a saddle point. So, we concentrate only on a 2×2 matrix and observe if it has a saddle point or not. We consider the 2×2 sub-matrix:

$$\begin{bmatrix} \phi(s, f_i, g_j) & \phi(s, f_i, g_{j'}) \\ \phi(s, f_{i'}, g_j) & \phi(s, f_i^{'}, g_{j'}) \end{bmatrix}$$

Where $i^{'}, i \in \{d_1, d_2 \cdots, d_{S_1}\}, (i \neq i^{'})$ and $j, j^{'} \in \{t_{S_1+1}, t_{S_1+2}, \cdots, t_{S_1+S_2}\}, (j \neq j^{'})$. Now, by suitably renumbering the strategies, we can write the above sub-matrix as:

$$A^{'}_{2 \times 2} = \begin{bmatrix} \phi(s, f_1, g_1) & \phi(s, f_1, g_2) \\ \phi(s, f_2, g_1) & \phi(s, f_2, g_2) \end{bmatrix}$$

Now we know

$$\phi(s, f_{i.}, g_{.j}) = \frac{\sum_{t=1}^{S_1}[q^*(t|s,f_{i.})r(t,f_{i.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v|s,g_{.j})r(v,g_{.j})]}{\sum_{t=1}^{S_1}[q^*(t|s,f_{i.})\bar{\tau}(t,f_{i.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v|s,g_{.j})\bar{\tau}(v,g_{.j})]}$$

We replace $\phi(s, f_{i.}, g_{.j})$ by the expression above in the matrix A. Let us rename the elements of the 2×2 sub-matrix as we consider the following two cases when A can not have a pure saddle point.

Case-1: $\phi(s,f_1,g_1)$ is row-minimum and column-minimum, $\phi(s,f_1,g_2)$ is row-maximum and column-maximum, $\phi(s,f_2,g_1)$ is row-maximum and column-maximum and $\phi(s,f_2,g_2)$ is row-minimum and column-minimum. These four conditions can be written as:

1. $\phi(s,f_1,g_1) < \phi(s,f_1,g_2)$.

2. $\phi(s,f_1,g_1) < \phi(s,f_2,g_1)$.

3. $\phi(s,f_2,g_2) < \phi(s,f_2,g_1)$.

4. $\phi(s,f_2,g_2) < \phi(s,f_1,g_2)$.

So, the above four inequalities can be written elaborately as:

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})r(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})r(v,g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})\bar{\tau}(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})\bar{\tau}(v,g_{.1})]}\\ &<\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})r(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})r(v,g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})\bar{\tau}(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})\bar{\tau}(v,g_{.2})]}\end{aligned} \tag{3.70}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})r(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})r(v,g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})\bar{\tau}(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})\bar{\tau}(v,g_{.1})]}\\ &<\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})r(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})r(v,g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})\bar{\tau}(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})\bar{\tau}(v,g_{.1})]}\end{aligned} \tag{3.71}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})r(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})r(v,g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})\bar{\tau}(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})\bar{\tau}(v,g_{.2})]}\\ &<\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})r(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})r(v,g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})\bar{\tau}(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.1})\bar{\tau}(v,g_{.1})]}\end{aligned} \tag{3.72}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})r(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})r(v,g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{2.})\bar{\tau}(t,f_{2.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})\bar{\tau}(v,g_{.2})]}\\ &<\frac{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})r(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})r(v,g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t\mid s,f_{1.})\bar{\tau}(t,f_{1.})]+\sum_{v=S_1+1}^{S_1+S_2}[q^*(v\mid s,g_{.2})\bar{\tau}(v,g_{.2})]}\end{aligned} \tag{3.73}$$

We rename the strategies $f_{1.}, f_{2.}, g_{.1}$ and $g_{.2}$ as 1., 2., .1 and .2 respectively to avoid notational complexity. Hence, (3.70) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,1.)q^*(v \mid s,.2)[\bar{\tau}(t,1.)r(v,.2) - r(t,1.)\bar{\tau}(v,.2)] \\
&\qquad + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,1.)q^*(v \mid s,.1)[r(t,1.)\bar{\tau}(v,.1) \\
&- r(v,.1)\bar{\tau}(t,1.)] \sum_{v=S_1+1}^{S_1+S_2}\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s,.1)[\bar{\tau}(v,.1)r(v,.2) - r(v,.1)\bar{\tau}(v,.2)] > 0
\end{aligned} \tag{3.74}$$

(3.72) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,2.)q^*(v \mid s,.1)[\bar{\tau}(t,2.)r(v,.1) - r(t,2.)\bar{\tau}(v,.1)] \\
&\qquad + \sum_{v=S_1+1}^{S_1+S_2}\sum_{v=S_1+1}^{S_1+S_2} q^*(v \mid s,.2)q^*(v \mid s,.1)[\bar{\tau}(v,.2)r(v,.1) \\
&- r(v,.2)\bar{\tau}(v,.1)] + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,2.)q^*(v \mid, s.2)[r(t,2.)\bar{\tau}(v,.2) - \bar{\tau}(t,2.)r(v,.2)] > 0
\end{aligned} \tag{3.75}$$

(3.71) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{t=1}^{S_1} q^*(t \mid s,1.)q^*(t \mid s,.2)[\bar{\tau}(t,1.)r(t,2.) - r(t,1.)\bar{\tau}(t,2.)] \\
&\qquad + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,1.)q^*(v \mid s,.1)[r(v,1.)\bar{\tau}(t,.1) \\
&- r(t,1.)\bar{\tau}(v,.1)] + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,2.)q^*(v \mid s,.1)[\bar{\tau}(v,.1)r(t,2.) - r(v,.1)\bar{\tau}(t,2.)] > 0
\end{aligned} \tag{3.76}$$

(3.73) yields

$$\sum_{t=1}^{S_1}\sum_{t=1}^{S_1} q^*(t \mid s,1.)q^*(t \mid s,.2)[r(t,1.)\bar{\tau}(t,2.) - r(t,2.)\bar{\tau}(t,1.)] + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,2.)q^*(v \mid s,.2)[r(v,.2)\bar{\tau}(t,2.) - r(t,2.)\bar{\tau}(v,.2)] + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t \mid s,1.)q^*(v \mid s,.2)[r(t,1.)\bar{\tau}(v,.2) - r(v,.2)\bar{\tau}(t,1.)] > 0 \tag{3.77}$$

Using the fact that, $0 < q^*(s^{'} \mid s,a) < 1$, (where $s,s^{'} \in \{1,2,\cdots,z\}$, a is the action chosen by either player-I or II) and adding (3.74) and (3.75), we get

$$\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (\bar{\tau}(t,1.)r(v,.2) - r(t,1.)\bar{\tau}(v,.2)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(t,1.)\bar{\tau}(v,.1) - \bar{\tau}(t,1.)r(v,.1)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(v,.1)\bar{\tau}(t,2.) - r(t,2.)\bar{\tau}(v,.1)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(t,2.)\bar{\tau}(v,.2) - \bar{\tau}(t,2.)r(v,.2)) > 0 \tag{3.78}$$

Similarly adding (3.76) and (3.77), we get

$$\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (\bar{\tau}(v,.2)r(t,1.) - r(v,.2)\bar{\tau}(t,1.)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(v,.1)\bar{\tau}(t,1.) - \bar{\tau}(v,.1)r(t,1.)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(t,2.)\bar{\tau}(v,.1) - r(v,.1)\bar{\tau}(t,2.)) + \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} (r(v,.2)\bar{\tau}(t,2.) - \bar{\tau}(v,.2)r(t,2.)) > 0 \tag{3.79}$$

From (3.78) and (3.79) we clearly get a contradiction. Now we consider the next case:

Case-2: $\phi(s, f_1, g_1)$ is row-maximum and column-maximum, $\phi(s, f_1, g_2)$ is row -minimum and column-minimum, $\phi(s, f_2, g_1)$ is row-minimum and column-minimum and $\phi(s, f_2, g_2)$ is row-maximum and column-maximum. These four conditions can be written as:

1. $\phi(s, f_1, g_1) > \phi(s, f_1, g_2)$.

2. $\phi(s, f_1, g_1) > \phi(s, f_2, g_1)$.

3. $\phi(s, f_2, g_2) > \phi(s, f_2, g_1)$.

4. $\phi(s, f_2, g_2) > \phi(s, f_1, g_2)$.

We can re-write them as follows:

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})\bar{\tau}(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})\bar{\tau}(v, g_{.1})]} \\ &> \frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})\bar{\tau}(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})\bar{\tau}(v, g_{.2})]}.\end{aligned} \tag{3.80}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})\bar{\tau}(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})\bar{\tau}(v, g_{.1})]} \\ &> \frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})\bar{\tau}(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})\bar{\tau}(v, g_{.1})]}.\end{aligned} \tag{3.81}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})\bar{\tau}(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})\bar{\tau}(v, g_{.2})]} \\ &> \frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})r(v, g_{.1})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})\bar{\tau}(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.1})\bar{\tau}(v, g_{.1})]}.\end{aligned} \tag{3.82}$$

$$\begin{aligned}&\frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})r(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{2.})\bar{\tau}(t, f_{2.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})\bar{\tau}(v, g_{.2})]} \\ &> \frac{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})r(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})r(v, g_{.2})]}{\sum_{t=1}^{S_1}[q^*(t \mid s, f_{1.})\bar{\tau}(t, f_{1.})] + \sum_{v=S_1+1}^{S_1+S_2}[q^*(v \mid s, g_{.2})\bar{\tau}(v, g_{.2})]}.\end{aligned} \tag{3.83}$$

Like the previous case we also rename the strategies $f_{1.}, f_{2.}, g_{1.}$ and $g_{2.}$ as 1., 2., .1 and .2 respectively to avoid notational complexity. Hence, (3.80) yields:

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(v\,|,s.1)q^*(t\mid s,1.)[\bar{\tau}(t,1.)r(v,.1)-r(t,1.)\bar{\tau}(v,.1)]\\
&+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,1.)q^*(v\,|,s.2)[r(t,1.)\bar{\tau}(v,.2)-\bar{\tau}(t,1.)r(v,.2)]\\
&+\sum_{v=S_1+1}^{S_1+S_2}\sum_{v=S_1+1}^{S_1+S_2} q^*(v\mid s,.2)q^*(v\mid s,.2)[r(v,.1)\bar{\tau}(v,.2)-r(v,.2)\bar{\tau}(v,.1)]>0
\end{aligned}\tag{3.84}$$

(3.82) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,2.)q^*(v\,|,s.2)[r(v,.2)\bar{\tau}(t,2.)-\bar{\tau}(v,.2)r(t,2.)]\\
&+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,2.)q^*(v\,|,s.1)[\bar{\tau}(v,.1)r(t,2.)-\bar{\tau}(t,2.)r(v,.1)]\\
&+\sum_{v=S_1+1}^{S_1+S_2}\sum_{v=S_1+1}^{S_1+S_2} q^*(v\mid s,.1)q^*(v\mid s,.2)[r(v,.2)\bar{\tau}(v,.1)-\bar{\tau}(v,.2)r(v,.1)]>0
\end{aligned}\tag{3.85}$$

(3.81) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,2.)q^*(v\,|,s.1)[r(v,.1)\bar{\tau}(t,2.)-r(t,.2)\bar{\tau}(v,.1)]\\
&+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,1.)q^*(v\,|,s.1)[r(t,1.)\bar{\tau}(v,.1)-r(v,.1)\bar{\tau}(t,1.)]\\
&+\sum_{t=1}^{S_1}\sum_{t=1}^{S_1} q^*(t\mid s,2.)q^*(t\mid s,1.)[r(t,1.)\bar{\tau}(t,2.)-\bar{\tau}(t,1.)r(t,2.)]>0
\end{aligned}\tag{3.86}$$

(3.83) yields

$$\begin{aligned}
&\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,1.)q^*(v\,|,s.2)[r(v,.2)\bar{\tau}(t,1.)-r(t,1.)\bar{\tau}(v,.2)]\\
&+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} q^*(t\mid s,2.)q^*(v\,|,s.2)[r(t,2.)\bar{\tau}(v,.2)-r(v,.2)\bar{\tau}(t,2.)]\\
&+\sum_{t=1}^{S_1}\sum_{t=1}^{S_1} q^*(t\mid s,2.)q^*(t\mid s,1.)[r(t,2.)\bar{\tau}(t,1.)-r(t,1.)\bar{\tau}(t,2.)]>0
\end{aligned}\tag{3.87}$$

Similarly using the fact that $0 < q^*(s^{'} \mid s,a) < 1$, (where $s,s^{'} \in \{1,2,\cdots,z\}$, a is the action chosen by either player-I or II in the state s) and adding (3.84) and (3.85), we get

$$\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(\bar{\tau}(v,.2)r(t,1.)-r(v,.2)\bar{\tau}(t,1.))+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(r(v,.1)\bar{\tau}(t,1.)-\bar{\tau}(v,.1)r(t,1.))+ \\ \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(r(t,2.)\bar{\tau}(v,.1)-r(v,.1)\bar{\tau}(t,2.))+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} \\ (r(v,.2)\bar{\tau}(t,2.)-r(t,.2)\bar{\tau}(v,2.))>0 \tag{3.88}$$

Now adding (3.86) and (3.87) we get

$$\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(\bar{\tau}(t,1.)r(v,.2)-r(t,1.)\bar{\tau}(v,.2))+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(r(t,1.)\bar{\tau}(v,.1)-\bar{\tau}(t,1.)r(v,.1))+ \\ \sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2}(r(v,.1)\bar{\tau}(t,2.)-r(t,2.)\bar{\tau}(v,.1))+\sum_{t=1}^{S_1}\sum_{v=S_1+1}^{S_1+S_2} \\ (r(t,.2)\bar{\tau}(v,2.)-r(v,.2)\bar{\tau}(t,2.))>0 \tag{3.89}$$

From (3.88) and (3.89) we get a contradiction. Thus, every 2×2 sub-matrix has a pure saddle point and by theorem 2.1 by Shapley ([7], (page-6)), the matrix A has a pure saddle point and the game Γ has a pure stationary optimal strategy pair for each initial state. Suppose $(f_1, f_2, \cdots, f_z)$ be optimal pure stationary strategies for player-I when the initial states are $1,2,\cdots,z$ respectively and $(g_1, g_2, \cdots, g_z)$ be optimal pure stationary strategies for player-II when the initial states are $1,2,\cdots,z$ respectively. Now using the result by Mondal (2017) [32], which states that if a strategy is optimal among the set of pure semi-stationary strategies then it is optimal among the set of behavioural strategies also, we can state that $f^* = (f_1, f_2, \cdots, f_z)$ and $g^* = (g_1, g_2, \cdots, g_z)$ are the optimal pure semi-stationary strategies for player-I and II respectively in the perfect information semi-Markov game Γ. □

3.2.6 Calculating the Cesaro Limiting Matrix Of A Transition Matrix

Lazari et al.,(2020) [23] proposed an algorithm to compute the Cesaro limiting matrix of any Transition (Stochastic) matrix Q with n states. The algorithm runs as follows:

Input: Let the transition matrix $Q \in M_n(\mathbb{R})$ (where $M_n(\mathbb{R})$ is the set of $n \times n$ matrices over the field of real numbers).
Output: The Cesaro limiting matrix $Q^* \in M_n(\mathbb{R})$.
Step 1: Determine the characteristic polynomial $C_Q(z) =\mid Q - zI_n \mid$.
Step 2: Divide the polynomial $C_Q(z)$ by $(z-1)^{m(1)}$ (where $m(1)$ is the algebraic multiplicity of the eigenvalue $z_0 = 1$) and call it quotient $T(z)$.
Step 3: Compute the quotient matrix $W = T(Q)$.
Step 4: Determine the limiting matrix Q^* by dividing the matrix W by the sum of its elements of any arbitrary row.

3.2.7 Policy Improvement Algorithm For Solving An Undiscounted Zero-Sum Two Person Perfect Information Semi-Markov Game

Let Γ be an undiscounted zero-sum two person perfect information semi-Markov game. We consider the following policy-improvement algorithm to compute an optimal pure semi-stationary strategy pair of the players. In what follows we describe the major steps of the algorithm:
Step 1: Fix an initial state $s_0 \in S = \{1, 2, \cdots, s_0, \cdots, z\}$ in the PISMG Γ.
Step 2: Choose a random pure stationary strategy for player-II g_k in Γ.
Step 3: Find the best response optimal pure stationary strategy f_k (for the initial state s_0) of player-I in the SMDP $\Gamma(g_k)$ using the algorithm given by Mondal (2020,[33]) described in section 2.1.8.
Step 4: **if** g_k is an optimal pure stationary strategy (for the initial state s_0) for player-II in $\Gamma(f_k)$, set $(f^*_{s_0}, g^*_{s_0}) = (f_k, g_k)$ and stop.
Step 5: **else** find a best response optimal pure stationary strategy g_{k+1} for player-II (for the same initial state s_0) in the SMDP $\Gamma(f_k)$, set $k = k+1$ and go to step 2.
Step 6: Repeat the above process for all initial states $s_0 \in \{1, 2, \cdots, z\}$ and obtain an optimal pure-semi stationary strategy pair (f^*, g^*), where $f^* = (f^*_0, f^*_1, \cdots, f^*_{s_0}, \cdots, f^*_z)$ and $g^* = (g^*_0, g^*_1, \cdots, g^*_{s_0}, \cdots, g^*_z)$.

Note: As the number of pure stationary strategy pairs for both the players are finite in number, this algorithm stops in finite number of steps.

3.2.8 Numerical Example

Example 3.3. *Consider a PISMG* Γ *with four states* $S=\{1,2,3,4\}$, $A(1)=\{1,2\}=A(2)$, $B(1)=B(2)=\{1\}$, $B(3)=B(4)=\{1,2\}$, $A(3)=A(4)=\{1\}$. *Player-II is the dummy player in the state* 1 *and* 2 *and player-I is the dummy player for the states* 3 *and* 4. *Rewards, transition probabilities and expected sojourn times for the players are given below*

State-1:

1.1 $(\frac{1}{2},\frac{1}{2},0,0)$ 1
1 $(\frac{1}{3},\frac{2}{3},0,0)$ 0.9

State-2:

3.1 $(\frac{1}{2},\frac{1}{2},0,0)$ 1
3 $(\frac{2}{3},\frac{1}{3},0,0)$ 1.1

State-3:

3 $(0,0,1,0)$ 1	5.8 $(0,0,1,0)$ 2

State-4:

4 $(\frac{1}{2},0,\frac{1}{2},0)$ 2	2 $(\frac{1}{2},0,\frac{1}{2},0)$ 1.1

Where a cell $\begin{array}{|r|}\hline r\\ (q_1,q_2,q_3,q_4)\\ \bar{\tau}\\ \hline\end{array}$ *represents that* r *is the immediate rewards of the players,* q_1, q_2, q_3, q_4 *represents that the next states are* 1, 2, 3 *and* 4 *respectively and* $\bar{\tau}$ *is the expected sojourn time if this cell is chosen at present. Here player-I is the row player and player-II is the column player. Player-I has the pure stationary strategies* $f_1=\{(1,0),(1,0),1,1\}$, $f_2=\{(1,0),(0,1),1,1\}$, $f_3=\{(0,1),(1,0),1,1\}$ *and* $f_4=\{(0,1),(0,1),1,1\}$. *Similarly the pure stationary strategies for player-II are* $g_1=\{1,1,(1,0),(1,0)\}$, $g_2=\{1,1,(1,0),(0,1)\}$, $g_3=\{1,1,(0,1),(1,0)\}$ *and* $g_4=\{1,1,(0,1),(0,1)\}$. *Now fix the initial state to be* 1 *in the PISMG* Γ. *We fix the strategy* $g_2=\{1,1,(1,0),(0,1)\}$ *for player-II in the PISMG* Γ. *Thus, we get a resultant SMDP* $\Gamma(g_2)$, *which is given below:*

$$\textit{State-1:}\ \begin{array}{|c|}\hline \qquad 1.1 \\ (\frac{1}{2},\frac{1}{2},0,0) \\ \qquad 1 \\ \hline \qquad 1 \\ (\frac{1}{3},\frac{2}{3},0,0) \\ \qquad 0.9 \\ \hline \end{array} \quad \textit{State-2:}\ \begin{array}{|c|}\hline \qquad 3.1 \\ (\frac{1}{2},\frac{1}{2},0,0) \\ \qquad 1 \\ \hline \qquad 3 \\ (\frac{2}{3},\frac{1}{3},0,0) \\ \qquad 1.1 \\ \hline \end{array} \quad \textit{State-3:}\ \begin{array}{|c|}\hline 3 \\ (0,0,1,0) \\ 1 \\ \hline \end{array}$$

$$\textit{State-4:}\ \begin{array}{|c|}\hline 2 \\ (\frac{1}{2},0,\frac{1}{2},0) \\ 1.1 \\ \hline \end{array}$$

Now, by the linear programming algorithm, given by Mondal (2020)[33], the linear programming problem with respect to the variables $x = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{41})$, $y = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{41})$ *and* t *is:*

$$\max R_1 = 1.1x_{11} + x_{12} + 3.1x_{21} + 3x_{22} + 3x_{31} + 2x_{41}$$

with respect to the constraints:

$$3x_{11} + 2x_{12} - 3x_{21} - 4x_{22} - 3x_{31} = 0 \tag{3.90}$$

$$-3x_{11} - 4x_{12} + 3x_{21} + 2x_{22} = 0 \tag{3.91}$$

$$x_{41} = 0 \tag{3.92}$$

$$12(x_{11} + x_{12}) + 6y_{11} + 8y_{12} - 6y_{21} - 8y_{22} - 3y_{41} - 12\delta(s_0, 1)t = 0 \tag{3.93}$$

$$12(x_{21} + x_{22}) - 6y_{11} - 8y_{12} + 6y_{21} + 4y_{22} - 12\delta(s_0, 2)t = 0 \tag{3.94}$$

$$2x_{31} - y_{41} - 2\delta(s_0, 3)t = 0 \tag{3.95}$$

$$x_{41} + y_{41} - \delta(s_0, 4)t = 0 \tag{3.96}$$

$$x_{11} + 0.9x_{12} + x_{21} + 1.1x_{22} + x_{31} + 1.1x_{42} = 1 \tag{3.97}$$

$$x, y, t \geq 0 \tag{3.98}$$

Where $\delta(s, s^{'})$ *is Kronecker delta function. Now solving the above LPP by dual-simplex method, we get* $x = (0, 0.3545, 0.4556, 0, 0.5, 0.5)$. *So, we get the optimal pure stationary strategy* f *for player-I in the SMDP* $\Gamma(g_2)$ *as* $f^1 = \{(0,1), (1,0), 1, 1\}$. *Now fix this strategy for player-I in the original PISMG* Γ. *The reduced SMDP* $\Gamma(f^1)$ *is given below:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline 1\\ (\frac{1}{3},\frac{2}{3},0,0)\\ 0.9\\ \hline\end{array}\quad \textit{State-2:}\ \begin{array}{|r|}\hline 3.1\\ (\frac{1}{2},\frac{1}{2},0,0)\\ 1\\ \hline\end{array}\quad \textit{State-3:}\ \begin{array}{|c|c|}\hline 3 & 5.8\\ (0,0,1,0) & (0,0,1,0)\\ 1 & 2\\ \hline\end{array}$$

$$\textit{State-4:}\ \begin{array}{|c|c|}\hline 4 & 2\\ (\frac{1}{2},0,\frac{1}{2},0) & (\frac{1}{2},0,\frac{1}{2},0)\\ 2 & 1.1\\ \hline\end{array}$$

Now solving again by the above algorithm, the linear programming problem with respect to the variables $x=(x_{11},x_{21},x_{31},x_{32},x_{41},x_{42})$, $y=(y_{11},y_{21},y_{31},y_{32},y_{41},y_{42})$ *and* t *is given by:*

$$\min R_2 = x_{11}+3.1x_{12}+3x_{31}+5.8x_{32}+4x_{41}+2x_{42}$$

with respect to the constraints:

$$4x_{11}+3x_{21}-3x_{41}-3x_{42}=0 \tag{3.99}$$
$$-4x_{11}+3x_{21}=0 \tag{3.100}$$
$$x_{41}+x_{42}=0 \tag{3.101}$$
$$6x_{11}+4y_{11}+3y_{21}-3y_{41}-3y_{42}-6\delta(s_0,1)t=0 \tag{3.102}$$
$$3x_{21}-2y_{11}+y_{21}-3\delta(s_0,2)t=0 \tag{3.103}$$
$$2(x_{31}+x_{32})-x_{41}-x_{42}-2\delta(s_0,3)t=0 \tag{3.104}$$
$$x_{41}+x_{42}+y_{41}+y_{42}-\delta(s_0,4)t=0 \tag{3.105}$$
$$0.9x_{11}+x_{12}+x_{31}+2x_{32}+2x_{41}+1.1x_{42}=1 \tag{3.106}$$
$$x,y,t\geq 0 \tag{3.107}$$

Solving the above LPP by dual-simplex method, and applying Mondal's algorithm (2020,[33]) we get the optimal pure stationary strategy of player-II in the SMDP $\Gamma(f^1)$ *is* $g^1=\{1,1,(0,1),(1,0)\}$. *Now fix this strategy for player-II in the PISMG* Γ. *The resulting SMDP* $\Gamma(g^1)$ *is given below:*

$$\textit{State-1:}\ \begin{array}{|r|}\hline 1.1\\ (\frac{1}{2},\frac{1}{2},0,0)\\ 1\\ \hline 1\\ (\frac{1}{3},\frac{2}{3},0,0)\\ 0.9\\ \hline\end{array}\quad \textit{State-2:}\ \begin{array}{|r|}\hline 3.1\\ (\frac{1}{2},\frac{1}{2},0,0)\\ 1\\ \hline 3\\ (\frac{2}{3},\frac{1}{3},0,0)\\ 1.1\\ \hline\end{array}\quad \textit{State-3:}\ \begin{array}{|c|}\hline 5.8\\ (0,0,1,0)\\ 2\\ \hline\end{array}\quad \textit{State-4:}\ \begin{array}{|c|}\hline 4\\ (\frac{1}{2},0,\frac{1}{2},0)\\ 2\\ \hline\end{array}$$

Now we solve the above SMDP by previously mentioned algorithm.The linear programming problem with respect to the variables $x = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{41})$*,* $y = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{41})$ *and* t *is given below:*

$$\max R_3 = 1.1x_{11} + x_{12} + 3.1x_{21} + 3x_{22} + 5.8x_{31} + 4x_{41}$$

with respect to the variables:

$$3x_{11} + 4x_{12} - 3x_{31} - 4x_{22} - 3x_{41} = 0 \tag{3.108}$$

$$-3x_{11} - 4x_{12} + 3x_{21} + 4x_{22} = 0 \tag{3.109}$$

$$x_{41} = 0 \tag{3.110}$$

$$2(x_{11} + x_{12}) + y_{11} + 4y_{12} - 3y_{21} - 4y_{22} - 3y_{41} - 2\delta(s_{0,1})t = 0 \tag{3.111}$$

$$3(x_{21} + x_{22}) - x_{11} - 2x_{12} + y_{21} + 2y_{22} - 3\delta(s_0, 2)t = 0 \tag{3.112}$$

$$2x_{31} - y_{41} - 2\delta(s_0, 3)t = 0 \tag{3.113}$$

$$x_{41} + y_{41} - \delta(s_0, 4)t = 0 \tag{3.114}$$

$$x_{11} + 0.9x_{12} + x_{21} + 1.1x_{22} + 2x_{31} + 2x_{41} = 1 \tag{3.115}$$

$$x, y, t \geq 0 \tag{3.116}$$

Solving the above LPP by dual-simplex method and applying previously mentioned algorithm we get the optimal pure stationary strategy of player-I in the above SMDP is $f^2 = \{(0,1), (1,0), 1, 1\}$*. Fixing the above strategy of player-I in the PISMG* Γ*, we get another resultant SMDP* $\Gamma(f^2)$ *as:*

State-1:	1 $(\frac{1}{3}, \frac{2}{3}, 0, 0)$ 0.9

State-2:	3.1 $(\frac{1}{2}, \frac{1}{2}, 0, 0)$ 1

State-3:	3 $(0, 0, 1, 0)$ 1	5.8 $(0, 0, 1, 0)$ 2

State-4:	4 $(\frac{1}{2}, 0, \frac{1}{2}, 0)$ 2	2 $(\frac{1}{2}, 0, \frac{1}{2}, 0)$ 1.1

Again applying Mondal's algorithm (2020,[33]), the linear programming problem with respect to the variables $x = (x_{11}, x_{21}, x_{31}, x_{32}, x_{41}, x_{42})$*,* $y = (y_{11}, y_{21}, y_{31}, y_{32}, y_{41}, y_{42})$ *and* t *is given by:*

$$\min R_4 = x_{11} + 3.1x_{21} + 3x_{31} + 5.8x_{32} + 4x_{41} + 2x_{42}$$

$$x_{11} + x_{21} - x_{41} - x_{42} = 0 \tag{3.117}$$

$$x_{11} - x_{21} = 0 \tag{3.118}$$

$$x_{41} + x_{42} = 0 \tag{3.119}$$

$$x_{11} + y_{11} - y_{21} - y_{41} - y_{42} - \delta(s_0, 1)t = 0 \tag{3.120}$$

$$x_{21} + y_{11} - y_{21} - \delta(s_0, 2)t = 0 \tag{3.121}$$

$$x_{31} + x_{32} + y_{41} + y_{42} - \delta(s_0, 3)t = 0 \tag{3.122}$$

$$x_{41} + x_{42} + y_{41} + y_{42} - \delta(s_0, 4)t = 0 \tag{3.123}$$

$$0.9x_{11} + x_{21} + x_{31} + 2x_{32} + 2x_{41} + 1.1x_{42} = 1 \tag{3.124}$$

$$x, y, t \geq 0 \tag{3.125}$$

Solving the above LP by dual-simplex method we get the optimal pure stationary strategy for player-II $g^2 = (1, 1, (0, 1), (1, 0))$ *and the objective functional value is* 2.2985*. Thus for the initial state* $s_0 = 1$*, the optimal pure stationary strategy pair of the players is* $(f_1^*, g_1^*) = (f_3, g_3)$*. Similarly by fixing the initial states* 2*,* 3 *and* 4 *in the PISMG* Γ *respectively and following the same process above, we get the optimal pure stationary strategy pairs as* (f_3, g_3)*,* (f_1, g_3) *and* (f_1, g_2)*. Thus the optimal pure semi-stationary strategy pair of the players in the PISMG* Γ *is given by* (f^*, g^*)*, where* $f^* = (f_3, f_3, f_1, f_1)$ *and* $g^* = (g_3, g_3, g_3, g_2)$ *and the value vector is given by* $(2.2985, 2.2985, 2.9, 0.9)$*.*

3.2.9 Concluding Remarks

In this chapter we proved the existence of the value and a pair of optimal pure stationary/semi-stationary strategies for the players in an undiscounted zero-sum two person stochastic/semi-Markov game. It is worthwhile to look into the corresponding non-zero sum version of undiscounted perfect information semi-Markov games.

Chapter 4

Undiscounted Semi-Markov Decision Processes With Countably Infinite Action Spaces

4.1 Introduction

In this section we investigate undiscounted semi-Markov decision processes with countably infinite action spaces where the state space is finite and the action space of the decision maker is (possibly) countably infinite. If there is any state where the number of actions is finite, then we can repeat a particular cell in that state to make the action space of that state countably infinite. However, we made no restrictions on the rewards. By undiscounted pay-offs we mean limiting ratio average pay-offs here. We prove that in such a model, the value of the decision process exists and the decision maker has near-optimal pure semi-stationary strategies. We have shown this result by exploiting the results of Sinha et. al [44] as the decision process can be regarded as a one player game. Thus the results can be applied for an undiscounted SMDP having finite states and countably many actions. We will refer to such SMDP as an infinite SMDP model. From the point of view of modelling real phenomena, the model with countably many actions can be used in a situation where the decision-maker has finitely many actions, but the cardinality of that set is not known appropriately. In this paper, we are considering maximisation problem for the infinite-SMDP. We organise this chapter as follows: section 4.1.1 contains the preliminaries and notations of an undiscounted infinite SMDP. In section 4.1.2 we define the undiscounted (limiting ratio average) pay-off criteria of an SMDP. In section 4.1.3, we define ϵN (near)-optimal strategy of the decision maker in an infinite SMDP and prove the existence of value and near-optimal pure semi-stationary strategies of an infinite SMDP. In section 4.1.4 we propose an algorithm to find the value of an undiscounted infinite SMDP and the near-optimal pure semi-stationary strategies of the decision maker. Also, later in that section, we propose an optimality equation of such an SMDP model under some recurrence/ergodicity conditions.

4.1.1 Preliminaries And Notations

An undiscounted semi-Markov decision process with countably infinite action space, considered in this process, is denoted by $\Gamma_\infty =< S, \{A(s) : s \in S\}$, $\{r(s,a) : s \in S, a \in A(s)\}, \{q(s^{'} \mid s,a) : s \in S, a \in A(s)\}, \{\bar{\tau}(s,a) : s \in S, a \in A(s)\} >$ which consists of a finite state space $S = \{1, 2, \cdots, z\}$. For each $s \in S$ we associate a countable set of actions $A(s) = \{1, 2, \dots\}$. The process is controlled by a decision maker. When the system is in the state s, the controller chooses an action $a \in A(s)$ and gets an immediate reward $r(s,a)$. Let us denote $K = \{(s,a) : s \in S, a \in A(s),\}$ to be the set of admissible tuples. For each $(s,a) \in K$, we denote $q(. \mid s,a)$ to be the transition law of the game. Given $(s,a) \in K$ and $s^{'} \in S$, let $\tau^a_{ss^{'}}$ be the transition time random variable which denotes the time for a transition to a state $s^{'}$ from a state s by the action $a \in A(s)$. Let $P^a_{ss^{'}} = Prob(\tau^a_{ss^{'}} \le t)$ for each $(s,a) \in K, s^{'} \in S$ be a probability distribution function on $[0, \infty)$ and it is called the conditional transition time distribution function. The process moves to a state $s^{'} \in S$ with probability $q(s^{'} \mid s,a)$, where $q(s^{'} \mid s,a) \ge 0$ and $\sum_{s^{'} \in S} q(s^{'} \mid s,a) = 1$. Thus the process repeats over infinite time with s replaced by $s^{'}$.

Here the strategies of the decision maker is restricted to strategies with finite support, i.e., for each $\pi \in \Pi$ ***there is*** $k \in \mathbb{N}$***, such that with probability*** 1 ***the decision maker chooses one of the first*** k ***cells in all states, for all histories. Mathematically, the decision maker chooses mixed actions from*** $\cup_{k \in \mathbb{N}} A_k$***; where*** $A_k = \{a \in \mathbb{R}^\infty : a \ge 0, \sum_{i=1}^k a_i = 1; a_i = 0$ ***for*** $i > k\}$.

We now discuss about the definitions of strategy spaces of the decision maker in an undiscounted SMDP, where the state and the action space are finite. By a strategy (behavioural) π of the decision maker, we mean a sequence $\{(\pi)_n(. \mid hist_n)\}_{n=1}^\infty$, where $(\pi)_n$ specifies which action is to be chosen on the n-th decision epoch by associating with each history $hist_n$ of the system up to nth decision epoch (where $hist_n$=$(s_1, a_1, s_2, a_2, \cdots, s_{n-1}, a_{n-1}, s_n)$ for $n \ge 1$, $hist_1 = (s_1)$ and $(s_k, a_k) \in K$ are respectively the state and actions of the players at the k-th decision epoch) a probability distribution $(\pi)_n(. \mid hist_n)$ on $A(s_n)$. Generally by any unspecified strategy, we mean behavioural strategy here. We denote Π to be the set of strategy (behavioural) spaces of decision maker. A strategy $g = \{g_n\}_{n=1}^{\infty}$ is called Markov if g_n depends on the history $hist_n$, through the current state s_n and the decision epoch number n. A strategy $\pi = \{\pi_n\}_{n=1}^{\infty}$ is called a stationary strategy if $\exists$ a map $f : S \to \mathbb{P}(A) = \{\mathbb{P}(A(s)) : s \in S\}$, where $\mathbb{P}(A(s))$ is the set of probability distribution on $A(s)$ such that $\pi_n = f$ for all n and $f(s) \in \mathbb{P}(A(s))$. A stationary strategy for the decision maker is defined as z tuple $f = (f(1), f(2), \cdots, f(z))$, where each $f(s)$ is the probability distribution on $A(s)$ given by $f(s) = (f(s,1), f(s,2), \cdots, f(s,m_s))$ (where $m_s =| A(s) |, s \in S,$). Where, $f(s,a)$ denotes the probability of choosing action a in the state s by the decision maker. Let us denote F^s to be the set of stationary strategies for the decision maker.

A stationary strategy is called pure if any player selects a particular action with probability 1 while visiting a state s. We denote F^{sp} to be the set of pure stationary strategies of the decision maker. A strategy $g = \{g_n\}_{n=1}^{\infty}$ is called semi-stationary strategy if there exists a map $\xi : S \times S \to \mathbb{P}(A)$ such that (i) $g_n(. \mid s_1 = s, \cdots, s_n = s^{'}) = \xi(s, s^{'})$ for all $s_1, s_2, \cdots, s_n \in S$, $n \in \mathbb{N}$ and (ii) $\xi(s, s^{'})$ has a support in $A(s^{'})$ for each pair $(s, s^{'}) \in S \times S$. So, it can also be viewed as a semi-Markov strategy which is independent of the decision epoch n, i.e., for a initial state s_1 and present state s_n, if a semi-Markov strategy $g(s_1, s_n, n)$ turns out to be independent of n, then we call it a semi-stationary strategy. Let ξ^s and ξ^{sp} denote the set of semi-stationary and pure semi-stationary strategies respectively for the decision maker.

For each undiscounted SMDP with countably infinite action space, defined as Γ_∞, we associate a truncated SMDP Γ_n, where Γ_n is defined as $\Gamma_n =< S, \{A_n(s) : s \in S\}$, $\{r(s,a) : s \in S, a \in A_n(s)\}, \{q(s^{'} \mid s,a) : s \in S, a \in A_n(s)\} >$ where $A_n(s) \subset A(s)$ consists of the actions for which $a < n$. Thus, the truncated game can be found from Γ by deleting all cells in Γ for which $a > n$ where $n \in \mathbb{N}$. We show that the value of the infinite SMDP model exists in $\mathbb{R}_{+}^{z}$, where $\mathbb{R}_{+}^{z} = (\mathbb{R} \cup \{+\infty\})^z$. In an undiscounted infinite SMDP, any strategy with finite support $n \in \mathbb{N}$, can be identified with the strategies in Γ_n. By completion with zeros in the cells of the n-truncated SMDP , the strategies for the decision maker in n-truncated SMDP Γ_n, can be identified with finite support strategies of the decision maker in Γ_∞. Similarly by deleting zeros from the cells of the infinite SMDP, finite support strategies of the decision maker in Γ_∞ can be identified with strategies of the decision maker in Γ_n for n sufficiently large.

4.1.2 Undiscounted Pay-Off Criterion In An SMDP Model With Finite State And Action Spaces

Suppose $\Gamma =< S, A, q, P, r >$ be an undiscounted SMDP with finite state space S action space A. Let $(X_1, A_1, X_2, A_2, \cdots)$ be a co-ordinate sequence in $S \times (A \times S)^{\infty}$. Given a strategy $\pi \in \Pi$ and an initial state $s \in S$, there exists a unique probability measure $\mathbb{P}_{\pi}(. \mid X_1 = s)$, hence an expectation $\mathbb{E}_{\pi}(. \mid X_1 = s)$ on the product σ-field of $S \times (A \times S)^{\infty}$ by Kolmogorov's extension theorem. For a behavioural strategy $\pi \in \Pi$, the limiting ratio average pay-off for the decision maker can be defined as:

$$\phi^1(s,\pi) = \liminf_{n\to\infty} \frac{\mathbb{E}_{\pi}\sum_{m=1}^{n}[r(X_m,A_m)|X_1=s]}{\mathbb{E}_{\pi}\sum_{m=1}^{n}[\bar{\tau}(X_m,A_m)|X_1=s]} \text{ for all } s \in S.$$

and

$$\phi^2(s,\pi) = \limsup_{n\to\infty} \frac{\mathbb{E}_{\pi}\sum_{m=1}^{n}[r(X_m,A_m)|X_1=s]}{\mathbb{E}_{\pi}\sum_{m=1}^{n}[\bar{\tau}(X_m,A_m)|X_1=s]} \text{ for all } s \in S.$$

Where $\bar{\tau}(s,a) = \sum_{s' \in S} q(s^{'} \mid s,a) \int_0^{\infty} t dP_{ss'}(t \mid a)$ is the expected sojourn time in the state s when decision maker chooses the action $a \in A(s)$. We assume that $\bar{\tau}(s,a)$ is bounded away from zero for each $(s,a) \in K$, where $K = \{(s,a) : s \in S, a \in A(s)\}$

Definition 4.1. *A strategy $\pi^* \in \Pi$ is called an optimal strategy under limiting ratio average pay-off if*

$$\phi^1(s,\pi^*) \geq \phi^1(s,\pi) \ (\phi^2(s,\pi^*) \geq \phi^2(s,\pi)) \text{ for all } \pi \in \Pi \text{ and all } s \in S.$$

Let $v^1(s) = \sup_{s\in S}(\phi^1(s,\pi))$ and $v^2(s) = \sup_{s\in S}(\phi^2(s,\pi))$. Then $v^1(s)$ (resp. $v^2(s)$) is called the limiting ratio average value of the undiscounted SMDP for the initial state $s \in S$. For a stationary strategy $f \in F^s$ and $s \in S$, let $r(s,f) = \sum_{a\in A(s)} r(s,a).f(s,a)$, $\bar{\tau}(s,f) = \sum_{a\in A(s)} \bar{\tau}(s,a) f(s,a)$ and $q(s^{'} \mid s,f) = \sum_{a\in A(s)} q(s^{'} \mid s,a) f(s,a)$. Then the reward vector, sojourn time vector and transition probability matrix can be defined respectively as: $r(f) = [r(s,f)]_{z\times 1}$, $\bar{\tau}(f) = [\bar{\tau}(s,f)]_{z\times 1}$ and $Q(f) = [q(s^{'} \mid s,f)]_{z\times z}$. Let $Q^m(f) = [q^m(s^{'} \mid s,f)]_{z\times z}$ where $q^m(s^{'} \mid s,f)$ is the m-step transition probability from state s to $s^{'}$ under the stationary strategy f.

Lemma 4.1.1. *(Doob, theorem* 2.1*, page-* 175*) [6] Let $Q = [q(s^{'} \mid s)]_{z\times z}$ be a transition probability matrix. Then $\exists$ a stochastic matrix $Q^* = [q^*(s^{'} \mid s)]_{z\times z}$, which is called the Cesaro-limiting matrix of Q, is defined as:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} q^m(s^{'} \mid s) = q^*(s^{'} \mid s) \ s, s^{'} \in S$$

From the above lemma and Sinha et al. [43], we have the following result:

Proposition 4.1.2. *Let $f \in F^s$ be a stationary strategy, then:*

$$\phi(s,f) = \phi^1(s,f) = \phi^2(s,f) = \tfrac{[Q^*(f)r(f)](s)}{[Q^*(f)\bar{\tau}(f)](s)} \text{ for all } s \in S$$

where $Q^*(f)$ is the Cesaro-limiting matrix of the transition matrix $Q(f)$ and $\phi(s,f)$ is the undiscounted pay-off of the decision maker. Define $v(s) = \sup_{s\in S}(\phi(s,f))$. Then $v(s)$ is called the limiting ratio average value of the undiscounted SMDP for the initial state $s \in S$.

For a stationary strategy $f \in F^s$ and all $s \in S$, we denote the limiting ratio average pay-off and the value of the n-truncated SMDP associated with an infinite undiscounted SMDP as $\phi_n(s,f)$ and v_n. Also for a stationary strategy $f \in F^s$ (with finite support) and for all $s \in S$, the pay-off function (limiting ratio average pay-off) for the undiscounted SMDP with countable actions is denoted as $\phi_\infty(s,f)$.

4.1.3 ϵN-Optimal Strategies For The Decision Maker In An Infinite Undiscounted SMDP Model

In this section we define the ϵN-optimal or near-optimal strategies for the decision maker in an infinite undiscounted SMDP model, assuming the infinite SMDP is a maximisation problem.

Definition 4.2. *A strategy $\pi^*_{\epsilon N}$ is called an ϵN-optimal strategy if there exists a $v_\infty \in \mathbb{R}^z_+$ such that for all $\epsilon > 0$ and $N > 0 (\in \mathbb{N})$ and for all $s \in S$, the following inequality holds:*

$$\phi_\infty(s, \pi^*_{\epsilon N}) \geq \begin{cases} v_\infty(s) - \epsilon & \text{if } v_\infty(s) \in \mathbb{R} \\ N & \text{if } v_\infty(s) = +\infty \end{cases}$$

Then v_∞ is called the value vector of the SMDP Γ_∞.

Now we prove the following theorem to establish the fact that there exists an ϵN-optimal pure semi-stationary strategy of the decision maker in the undiscounted SMDP model.

Theorem 4.1.3. *(i)* $\lim_{n\to\infty} v_n$ *exists and equals to* v_∞.
(ii) The decision maker has a near-optimal pure semi-stationary strategy in Γ_∞.

Proof. We know from Sinha et.,al ([43], 2017), for an undiscounted (limiting ratio average) SMDP, the value and optimal pure semi-stationary strategy exists. Now the semi-stationary strategy space of Γ_{n+1} contains that of Γ_n. Thus the value vector v_{n+1}, i.e., the maximal value in Γ_{n+1} will lie above v_n, that of Γ_n. So, $\{v_n\}_{n=1}^{\infty}$ is an increasing sequence and hence it converges to a point $w \in \mathbb{R}_+^z$ co-ordinate wise (where $\mathbb{R}_+ = \mathbb{R} \cup \{\infty\}$). We prove that w is the value of Γ_∞, i.e., $w = v_\infty$.
Let π be any strategy with finite support for the decision maker in Γ_∞. Then it can be seen as a strategy for the decision maker in Γ_n for some $n \in \mathbb{N}$. Now we fix $n_0 \in \mathbb{N}$ to be large enough. Now considering the n_0-truncated SMDP Γ_{n_0}, we know by Sinha et al.(2017) [43] the decision maker has an optimal pure semi-stationary strategy in Γ_{n_0}. Suppose $f_{n_0}^*$ is the optimal pure semi-stationary strategy of the decision maker in Γ_{n_0}. Now by theorem 1 of [43], we have $v_{n_0} = \phi_{n_0}(f_{n_0}^*) \geq \phi_{n_0}(\pi)$ co-ordinate wise, for all $\pi \in \Pi$. Since $\{v_n\}_{n=1}^{\infty}$ is an increasing sequence converging to $w \in \mathbb{R}_+^z$, we have $w \geq v_{n_0} \geq \phi_{n_0}(\pi)$ for all $\pi \in \Pi$. Thus we have $w \geq v_\infty$. Now take $\epsilon > 0$. As $\{v_n\}_{n=1}^{\infty}$ converges to w, there exists an $n_1 \in \mathbb{N}$ such that $v_{n_1}(s) \geq w(s) - \epsilon$ if $w(s)$ is finite and $v_{n_1}(s) \geq N$ if $w(s) = +\infty$. Now considering the truncated SMDP Γ_{n_1}, we get a pure semi-stationary strategy $f_{n_1}^*$ of the decision maker, having a value vector v_{n_1}. Thus we have

$$\phi_\infty(s, f_{n_1}^*) = v_{n_1}(s) \geq \begin{cases} w(s) - \epsilon & \text{if } w(s) \in \mathbb{R} \\ N & \text{if } w(s) = +\infty \end{cases}$$

Combination of the above arguments proves our theorem 4.1.3. □

4.1.4 Algorithm To Compute Value And A Near-Optimal Pure Semi-Stationary Strategy In An Undiscounted SMDP With Countably Infinite Action Spaces

We propose an efficient algorithm to compute the value and a near-optimal pure semi-stationary strategy of the decision maker in an undiscounted SMDP with countably infinite action space. We follow the algorithm proposed by Mondal (2020) [33] to compute optimal pure semi-stationary strategy of the decision maker and value vector in each truncated SMDP. The steps of our algorithm are given below:

```
Step 1: Fix $\epsilon > 0$ and a large positive integer $N \in \mathbb{N}$.
Step 2: Given $n \in \mathbb{N}$, consider the truncated SMDPs $\Gamma_n$ and $\Gamma_{n+1}$.
Step 3: Fix an initial state $s_0 \in S$.
Step 4: Apply the algorithm by Mondal ([33], 2020) discussed in section 2.1.8, to compute optimal pure stationary strategies $f_n^{s_0*}$ and $f_{n+1}^{s_0*}$ for the initial state $s_0$ of the SMDPs $\Gamma_n$ and $\Gamma_{n+1}$ respectively and also the values $v_n(s_0)$ and $v_{n+1}(s_0)$, for given $n$.
Step 5: if
(i)$\mid v_{n+1}(s_0) - v_n(s_0) \mid < \epsilon$ then $v_{n+1}(s_0) = v_\infty(s_0)$ and $f_{n+1}^{s_0*}$ is a near-optimal pure stationary strategy for the initial state $s_0$.
    do Repeat step 4 and 5($i$) for all $s_0 \in S = \{1,2,\cdots,z\}$ to get $f_{n+1}^* = \{f_{n+1}^{1*}, f_{n+1}^{2*}, \cdots, f_{n+1}^{s_0*}, \cdots, f_{n+1}^{z*}\}$ as a near-optimal pure semi-stationary strategy.
    end do
or
else if
(ii)$\mid v_{n+1}(s_0) \mid > N$ for all $s \in S$, then $v_{n+1}(s_0) = +\infty = v_\infty(s_0)$ and $f_{n+1}^{s_0*}$ is a near-optimal pure stationary strategy for the initial state $s_0$.
    do Repeat step 4 and 5($ii$) for all $s_0 \in S = \{1,2,\cdots,z\}$ to get $f_{n+1}^* = \{f_{n+1}^{1*}, f_{n+1}^{2*}, \cdots, f_{n+1}^{s_0*}, \cdots, f_{n+1}^{z*}\}$ as a near-optimal pure semi-stationary strategy.
    end do
or
else if (iii)$\mid v_{n+1}(s_0) - v_n(s_0) \mid = 0$ then $v_{n+1}(s_0) = v_\infty(s_0)$ and $f_{n+1}^{s_0*}$ is a pure stationary strategy for the initial state $s_0$.
    do Repeat step 4 and 5($iii$) for all $s_0 \in S = \{1,2,\cdots,z\}$ to get $f_{n+1}^* = \{f_{n+1}^{1*}, f_{n+1}^{2*}, \cdots, f_{n+1}^{s_0*}, \cdots, f_{n+1}^{z*}\}$ as an optimal pure semi-stationary strategy.
    end do
end if
Step 6: Replace $n$ by $n+1$ and return to step 3.
```

4.1.5 Numerical Example

Example 4.1. *Consider an undiscounted infinite SMDP Γ_∞ with three states $S = \{1,2,3\}$, $A(1) = A(2) = A(3) = \{1,2,\cdots\}$. Thus in all states, the decision maker has action sets whose cardinality is not finite. Rewards, transition probabilities and expected sojourn times for the decision maker are given below:*

State-1:	*State*-2:	*State*-3:
1 $(1,0,0)$ 1	1 $(1,0,0)$ 2.1	1 $(\frac{1}{2},\frac{1}{4},\frac{1}{4})$ 3.4
$2^{0.86}$ $(1,0,0)$ 1	$(1+\frac{1}{4})$ $(\frac{1}{2},0,\frac{1}{2})$ 2.1	$(1+\frac{1}{8})$ $(\frac{3}{4},\frac{1}{8},\frac{1}{8})$ 3.4
$3^{0.86}$ $(1,0,0)$ 1	$(1+\frac{1}{4}+\frac{1}{9})$ $(\frac{1}{3},0,\frac{2}{3})$ 2.1	$(1+\frac{1}{8}+\frac{1}{27})$ $(\frac{5}{6},\frac{1}{12},\frac{1}{12})$ 3.4
$\vdots$	$\vdots$	$\vdots$
$n^{0.86}$ $(1,0,0)$ 1	$(1+\frac{1}{4}+\cdots\frac{1}{n^2})$ $(\frac{1}{n},0,1-\frac{1}{n})$ 2.1	$(1+\frac{1}{8}+\cdots\frac{1}{n^3})$ $(1-\frac{1}{2n},\frac{1}{4n},\frac{1}{4n})$ 3.4
$\vdots$	$\vdots$	$\vdots$

We can write the reward function as:
$r(1,n)=n^{0.86}$, $r(2,n)=\sum_{n=1}^{\infty}(1/n^2)$ *and* $r(3,n)=\sum_{n=1}^{\infty}(1/n^3)$ *for* $n\in\mathbb{N}$.
The transition probabilities can be written as:
$p(1\mid 1,n)=1,\ p(2\mid 1,n)=0,\ p(3\mid 1,n)=0.$
$p(1\mid 2,n)=\frac{1}{n},\ p(2\mid 2,n)=0,\ p(3\mid 2,n)=1-\frac{1}{n}.$
$p(1\mid 3,n)=1-\frac{1}{2n},\ p(2\mid 3,n)=\frac{1}{4n},\ p(3\mid 3,n)=\frac{1}{4n}.$
The mean sojourn times are given as:
$\bar{\tau}(1,n)=1,\ \bar{\tau}(2,n)=2.1,\ \bar{\tau}(3,n)=3.4.$

For each $n\in\mathbb{N}$ *we write the linear programming problem with respect to the variables* $x=(x_{1n},x_{2n},x_{3n}),y=(y_{1n},y_{2n},y_{3n})$ *and* t *for this model as:*

$$\max[\textstyle\sum_{n=1}^{\infty}n^{0.86}x_{1n}+\sum_{n=1}^{\infty}\frac{1}{n^2}x_{2n}+\sum_{n=1}^{\infty}\frac{1}{n^3}x_{3n}]$$

With respect to the constraints:
(i) $-\frac{1}{n}x_{2n}+(\frac{1}{2n}-1)x_{3n}=0$
(ii) $x_{2n}-\frac{1}{4n}x_{3n}=0$
(iii) $(\frac{1}{n}-1)x_{2n}+(1-\frac{1}{4n})x_{3n}=0$
(iv) $x_{1n}+-\frac{1}{n}y_{2n}+(\frac{1}{2n}-1)y_{3n}-\delta_{s_01}t=0$
(v) $x_{2n}+x_{2n}-\frac{1}{4n}x_{3n}-\delta_{s_02}t=0$
(vi) $x_{3n}+(\frac{1}{n}-1)x_{2n}+(1-\frac{1}{4n})x_{3n}-\delta_{s_03}t=0$
(vii)$[x_{1n}+2.1x_{2n}+3.4x_{3n}]=1$

where $x_{1n}, x_{2n}, x_{3n}, y_{1n}, y_{2n}, y_{3n}, t \geq 0$ *for all* $n \in \mathbb{N}$. *Let us fix* $N = 100$ *in this case. The values of* v_{n+1} *and* v_n *for each* $n \in \mathbb{N}$ *are given below.*

Value of n	v_n	v_{n+1}
1	$(1,1,1)$	$(1.828427, 1.828427, 1.828427)$
2	$(1.828427, 1.828427, 1.828427)$	$(2.196152, 2.196152, 2.196152)$
3	$(2.196152, 2.196152, 2.196152)$	$(3,3,3)$
4	$(3,3,3)$	$(4.180340, 4.180340, 4.180340)$
5	$(4.180340, 4.180340, 4.180340)$	$(6.696938, 6.696938, 6.696938)$
6	$(6.696938, 6.696938, 6.696938)$	$(8.520259, 8.520259, 8.520259)$
7	$(8.520259, 8.520259, 8.520259)$	$(12.627417, 12.627417, 12.627417)$
8	$(12.627417, 12.627417, 12.627417)$	$(17,17,17)$
9	$(17,17,17)$	$(23.622777, 23.622777, 23.622777)$
10	$(23.622777, 23.622777, 23.622777)$	$(30.482873, 30.482873, 30.482873)$
11	$(30.482873, 30.482873, 30.482873)$	$(37.569219, 37.569219, 37.569219)$
12	$(37.569219, 37.569219, 37.569219)$	$(45.872167, 45.872167, 45.872167)$
13	$(45.872167, 45.872167, 45.872167)$	$(53.383203, 53.383203, 53.383203)$
14	$(53.383203, 53.383203, 53.383203)$	$(59.094750, 59.094750, 59.094750)$
15	$(59.094750, 59.094750, 59.094750)$	$(67,67,67)$
16	$(67,67,67)$	$(77.092796, 77.092796, 77.092796)$
17	$(77.092796, 77.092796, 77.092796)$	$(85.367532, 85.367532, 85.367532)$
18	$(85.367532, 85.367532, 85.367532)$	$(92.819080, 92.819080, 92.819080)$
19	$(92.819080, 92.819080, 92.819080)$	$(99.442719, 99.442719, 99.442719)$
20	$(99.442719, 99.442719, 99.442719)$	$(108.234090, 108.234090, 108.234090)$
$\vdots$	$\vdots$	$\vdots$
1000	$(1851.73, 1851.73, 1851.73)$	$(1866.5, 1866.5, 1866.5)$

Now applying our algorithm, we conclude $v_\infty = (+\infty, +\infty, +\infty)$. *Now for the initial state* 1 *the optimal pure stationary strategy is* $f_1^* = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0)$. *Similarly for the initial state* 2 *and* 3, *the optimal pure stationary strategies are* $f_2^* = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)$ *and* $f_3^* = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)$. *Thus* $f^* = (f_1^*, f_2^*, f_3^*)$ *is the pure near-optimal semi-stationary strategy.*

Note: In this undiscounted SMDP model with countably infinite action space, we have not put any restriction on the reward function. However if we assume the reward function to be uniformly bounded over all state-action pair and further assuming some ergodicity conditions on the SMDP model, we can establish an optimality equation of an undiscounted SMDP model with countable action space. We state the ergodicity condition below:

C1: There is a state $x_0 \in S$ and a finite number T such that $\mathbb{E}_f[T_{x_0} \mid X_0 = y] \leq T$ for all $y \in S$ and for all stationary strategies $f \in F^s$ with finite support (where $T_{x_0} = \inf[n \geq 1 \mid X_n = X_0]$).

To establish the optimality equation, we use a data transformation introduced by Schweitzer (1971) [41]. Let $\Gamma_\infty =< S, A, q, r, \bar{\tau} >$ be an undiscounted SMDP with countable action space. Here, the mean sojourn time is $\bar{\tau}(s,a)$ in a state s, when the decision maker chooses an action $a \in A(s)$. We assume that $\bar{\tau}(s,a)$ is bounded away from zero for each $(s,a) \in K$, where $K = \{(s,a) : s \in S, a \in A(s)\} \subseteq S \times A$. Now, we associate the semi-Markov model a discrete time Markov decision process, with state space $\hat{S} = S$, countable action space $\hat{A}(s) = \{\hat{A}(s) = A(s) : s \in \hat{S}\}$, one step reward $\hat{r}(s,a)$, one-step transition time $\hat{\bar{\tau}}(s,a) = 1$ and one-step transition probabilities $\hat{q}(s^{'} \mid s,a)$, where $s, s^{'} \in S$, and $a \in A(s)$. We define them in the following:

$$\hat{r}(s,a) = \tfrac{r(s,a)}{\bar{\tau}(s,a)}$$
$$\text{and}$$
$$\hat{q}(s^{'} \mid s,a) = \eta \tfrac{(q(s^{'}|s,a) - \delta(s,s^{'}))}{\bar{\tau}(s,a)} + \delta(s,s^{'}),$$

where $\delta(s,s^{'})$ is Kronecker delta function defined as $\delta(s,s^{'}) = 1$ if $s = s^{'}$ and 0 otherwise. The parameter η is chosen such that:

$$0 < \eta < \inf_{(s,a)}\{\tfrac{\bar{\tau}(s,a)}{(1-q(s|s,a))}\} \text{ if } q(s \mid s,a) < 1 \text{ for some } (s,a)K.$$
$$\text{and } \eta = 1 \text{ if } q(s \mid s,a) = 1 \text{ for all } (s,a) \in K.$$

The associated β-discounted Markov decision process with countable action space is defined by $\hat{\Gamma}_\infty^\beta =< \hat{S}, \hat{A}, \hat{q}, \hat{r} >$. For any $\beta \in (0,1)$, we know that the Markov decision process with countable action space, has a value $v_\infty^\beta(s) \in \mathbb{R}_+^z$ (considering maximisation problem) for $s \in \hat{S}$ and has a near-optimal pure stationary strategy. Following similar methods, as described in [44], we can also verify that the value of the above said

Markov decision process, satisfies the optimality equation, i.e.,

$$v_{\infty}^{\beta}(s) = \max_{a \in \hat{A}(s)} [\hat{r}(s,a) + \beta \sum_{s' \in S} \hat{q}(s^{'} \mid s,a) v_{\infty}^{\beta}(s)] \text{ for all } s \in \hat{S}. \tag{4.1}$$

Assumption: To establish the optimality equation in the SMDP model, we assume that the reward function of the Markov decision process as well as the semi-Markov decision process is uniformly bounded, i.e., $\mid \hat{r}(s,a) \mid \leq M$**, for all** $(s,a) \in \hat{K} = \{(s,a) : s \in \hat{S}, a \in \hat{A}(s)\}$**, where** $\hat{K} \subseteq \hat{S} \times \hat{A}$ **and** $M > 0$ **is a positive real number.**

Now we prove the following lemma to establish the optimality equation of our SMDP model.

Lemma 4.1.4. *Let* $s_0 \in \hat{S}$ *be fixed. Define* $H_{\infty}^{\beta}(s) = v_{\infty}^{\beta}(s) - v_{\infty}^{\beta}(s_0)$*. Suppose* $\mid H_{\infty}^{\beta}(s) \mid < K$ *for all* $s \in \hat{S}$ *and all* $\beta \in (0,1)$*. Then*
(i)There exist bounded functions $h(s)$ *and a constant* v^* *satisfying:*

$$h(s) = \max_{a \in \hat{A}(s)} [\hat{r}(s,a) - v^* + \textstyle\sum_{s' \in S} \hat{q}(s^{'} \mid s,a) h(s^{'})]$$

(ii)For some sequence $\beta_n \uparrow 1$*,* $h(s) = \lim_{n \to \infty} H_{\infty}^{\beta_n}(s)$*.*
(iii)$\lim_{n \to \infty}(1-\beta_n) v_{\infty}^{\beta_n}(s_0) = v^*$*.*

Proof. Define $V_{\infty}^{\beta}(s) = (1-\beta) v_{\infty}^{\beta}(s)$. Then,

$$V_{\infty}^{\beta}(s_0) + H_{\infty}^{\beta}(s) = (1-\beta) v_{\infty}^{\beta}(s_0) + v_{\infty}^{\beta}(s) - v_{\infty}^{\beta}(s_0) \tag{4.2}$$

$$= v_{\infty}^{\beta}(s) - \beta v_{\infty}^{\beta}(s_0) \tag{4.3}$$

$$= \max_{a \in \hat{A}(s)} [\hat{r}(s,a) + \beta \sum_{s' \in S} \hat{q}(s^{'} \mid s,a) H_{\infty}^{\beta}(s)] \quad \text{(from(4.1))} \tag{4.4}$$

Now as the state space $\hat{S}$ is finite and $H_{\infty}^{\beta}(s)$ is bounded for all $\beta \in (0,1)$, there exists a sub-sequence $\beta_n \uparrow 1$ such that $H_{\infty}^{\beta_n}(s)$ converges to some $h(s)$ for all $s \in S$. Also for $\beta \in (0,1)$, $V_{\infty}^{\beta}(s_0) = (1-\beta) v_{\infty}^{\beta}(s_0) \leq (1-\beta)M$, as $\mid \hat{r}(s,a) \mid \leq M$ for all $(s,a) \in \hat{K}$. Thus without any loss of generality, we assume that $V_{\infty}^{\beta_n}(s_0)$ converges to v^* as $\beta_n \uparrow 1$. Thus the lemma is proved. □

From the above lemma, we have the following equation

$$h(s) = \max_{a \in \hat{A}(S)} [\hat{r}(s,a) - v^* + \sum_{s' \in S} \hat{q}(s^{'} \mid s,a) h(s^{'})], s \in \hat{S}. \tag{4.5}$$

Which implies

$$\eta h(s) = \max_{a \in A(s)} [r(s,a) - v^* \bar{\tau}(s,a) + \sum_{s^{'} \in S} q(s^{'} \mid s,a) \eta h(s^{'})], \quad s \in S. \tag{4.6}$$

Thus putting $\eta h(s) = h(s)$, we have the following optimality equation for an undiscounted SMDP with countable action spaces:

$$h(s) = \max_{a \in A(s)} [r(s,a) - v^* \bar{\tau}(s,a) + \sum_{s^{'} \in S} q(s^{'} \mid s,a) h(s^{'})]. \tag{4.7}$$

4.1.6 Concluding Remarks

We conclude this chapter by showing that for an undiscounted (limiting ratio average) SMDP with countably infinite action space, instead of an optimal pure semi-stationary strategy, the decision maker has a near-optimal pure semi-stationary strategy. Also, under some ergodicity conditions we can establish the optimality equation of such SMDP model. By our proposed algorithm we can efficiently find the value and a near-optimal pure semi-stationary strategy of the decision maker. However, it is interesting to look into an undiscounted zero-sum two person perfect information semi-Markov game, where a non-fixed player in each state has countably infinite actions. The existence of near-optimal pure semi-stationary strategies in such scenarios is worth studying.

References

[1] Bellman, R. (1957). Dynamic programming, princeton univ. *Press Princeton, New Jersey.*

[2] Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726.

[3] Chatterjee, K., Majumdar, R., and Henzinger, T. A. (2006). Markov decision processes with multiple objectives. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 325–336. Springer.

[4] Derman, C. (1962). On sequential decisions and markov chains. *Management Science*, 9(1):16–24.

[5] Derman, C. (1964). On sequential control processes. *The Annals of Mathematical Statistics*, 35(1):341–349.

[6] Doob, J. L. (1953). *Stochastic processes.* John Wiley & Sons.

[7] Dresher, M., Berkovitz, L., Aumann, R., Shapley, L., Davis, M., and Tucker, A. (1964). *Advances in Game Theory.* Annals of Mathematics Studies. Princeton University Press.

[8] Dresher, M., Shapley, L. S., and Tucker, A. W. (pp. 6; 2016). *Advances in Game Theory.(AM-52), Volume 52.*

[9] Federgruen, A., Hordijk, A., and Tijms, H. C. (1978). A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices. *Journal of Applied Probability*, 15(4):842–847.

[10] Filar, J. and Vrieze, K. (2012). *Competitive Markov decision processes.* Springer Science & Business Media.

[11] Fudenberg, D. and Maskin, E. (2009). The folk theorem in repeated games with discounting or with incomplete information. In *A long-run collaboration on long-run games*, pages 209–230. World Scientific.

[12] Gillette, D. (1957). Stochastic games with zero stop probabilities. *Contributions to the Theory of Games*, 3:179–187.

[13] Hoffman, A. J. and Karp, R. M. (1966). On nonterminating stochastic games. *Management Science*, 12(5):359–370.

[14] Hordijk, A., Dekker, R., and Kallenberg, L. C. M. (1985). Sensitivity-analysis in discounted markovian decision problems. *Operations-Research-Spektrum*, 7(3):143–151.

[15] Hordijk, A. and Kallenberg, L. (1979). Linear programming and markov decision chains. *Management Science*, 25(4):352–362.

[16] Howard, R. A. (1963). Semi-markovian decision-processes. *Bulletin of the International Statistical Institute*, 40(2):625–652.

[17] Howard, R. A. (1971). Semi-markov and decision processes. *(No Title)*.

[18] Jewell, W. S. (1963a). Markov-renewal programming. i: Formulation, finite return models. *Operations Research*, 11(6):938–948.

[19] Jewell, W. S. (1963b). Markov-renewal programming. ii: Infinite return models, example. *Operations Research*, 11(6):949–971.

[20] Jianyong, L. and Xiaobo, Z. (2004). On average reward semi-markov decision processes with a general multichain structure. *Mathematics of Operations Research*, 29(2):339–352.

[21] Kemeny, J. G. and Snell, J. L. (1983). *Finite Markov chains: with a new appendix" Generalization of a fundamental matrix"*. Springer.

[22] Lal, A. K. and Sinha, S. (1992). Zero-sum two-person semi-markov games. *Journal of applied probability*, 29(1):56–72.

[23] Lazari, A. and Lozovanu, D. (2020). New algorithms for finding the limiting and differential matrices in markov chains. *Buletinul Academiei de Ştiinţe a Moldovei. Matematica*, 92(1):75–88.

[24] Liggett, T. M. and Lippman, S. A. (1969). Stochastic games with perfect information and time average payoff. *Siam Review*, 11(4):604–607.

[25] Lippman, S. A. et al. (1971). Maximal average-reward policies for semi-markov decision processes with arbitrary state and action space. *The Annals of Mathematical Statistics*, 42(5):1717–1726.

[26] Loomis, L. H. (1946). On a theorem of von neumann. *Proceedings of the National Academy of Sciences*, 32(8):213–215.

[27] Luque-Vásquez, F. (2002). Zero-sum semi-markov games in borel spaces: discounted and average payoff. *Bol. Soc. Mat. Mexicana*, 8:227–241.

[28] Luque-Vásquez, F. and Hernández-Lerma, O. (1999). Semi-markov control models with average costs. *Applicationes mathematicae*, 26(3):315–331.

[29] Maitra, A. (1965). Dynamic programming for countable state systems. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 241–248.

[30] Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science*, 6(3):259–267.

[31] Mertens, J. F. and Neyman, A. (1981). Stochastic games. *International Journal of Game Theory*, 10:53–66.

[32] Mondal, P. (2017). On zero-sum two-person undiscounted semi-markov games with a multichain structure. *Advances in Applied Probability*, 49(3):826–849.

[33] Mondal, P. (2020). Computing semi-stationary optimal policies for multichain semi-markov decision processes. *Annals of Operations Research*, 287(2):843–865.

[34] Mondal, P. and Sinha, S. (2015). Ordered field property for semi-markov games when one player controls transition probabilities and transition times. *International Game Theory Review*, 17(02):1540022.

[35] Neumann, J. v. (1945). A model of general economic equilibrium. *The Review of Economic Studies*, 13(1):1–9.

[36] Neumann, J. v. (1971). A model of general economic equilibrium. In *Readings in the Theory of Growth*, pages 1–9. Springer.

[37] Owen, G. (1995). Game theory academic press. *San Diego*.

[38] Puterman, M. L. (1990). Markov decision processes. *Handbooks in operations research and management science*, 2:331–434.

[39] Raghavan, T. and Syed, Z. (2003). A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. *Mathematical Programming*, 95(3):513–532.

[40] Ross, S. M. (2013). *Applied probability models with optimization applications*. Courier Corporation.

[41] Schweitzer, P. J. (1971). Iterative solution of the functional equations of undiscounted markov renewal programming. *Journal of Mathematical Analysis and Applications*, 34(3):495–501.

[42] Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100.

[43] Sinha, S. and Mondal, P. (2017). Semi-markov decision processes with limiting ratio average rewards. *Journal of Mathematical Analysis and Applications*, 455(1):864–871.

[44] Sinha, S., Thuijsman, F., and Tijs, S. H. (1991). Semi-infinite stochastic games. In *Stochastic Games And Related Topics: In Honor of Professor LS Shapley*, pages 71–83. Springer.

[45] Thuijsman, F. and Raghavan, T. E. (1997). Perfect information stochastic games and related classes. *International Journal of Game Theory*, 26(3):403–408.

[46] Vega-Amaya, O. (2003). Zero-sum average semi-markov games: fixed-point solutions of the shapley equation. *SIAM journal on control and optimization*, 42(5):1876–1894.

[47] Von Neumann, J. (1928). On the theory op games op strategy1. *Mathematische Annalen*, 100:295–320.

[48] Von Neumann, J. and Morgenstern, O. (2007). Theory of games and economic behavior. In *Theory of games and economic behavior*. Princeton university press.

[49] Yang, P. and Catthoor, F. (2003). Pareto-optimization-based run-time task scheduling for embedded systems. In *Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, pages 120–125.