

Julio B. Clempner
Alexander Poznyak

Optimization and Games for Controllable Markov Chains

Numerical Methods with Application
to Finance and Engineering

Studies in Systems, Decision and Control

Volume 504

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

The series “Studies in Systems, Decision and Control” (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control—quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Julio B. Clempner · Alexander Poznyak

Optimization and Games for Controllable Markov Chains

Numerical Methods with Application
to Finance and Engineering



Springer

Julio B. Clempner
Escuela Superior de Física y Matemáticas
(ESFM)
Instituto Politécnico Nacional
Mexico City, Mexico

Alexander Poznyak 
Department of Automatic Control
Center for Research and Advanced Studies
Mexico City, Mexico

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-031-43574-4 ISBN 978-3-031-43575-1 (eBook)
<https://doi.org/10.1007/978-3-031-43575-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

*This book is dedicated to my beloved wife
Erika and our three children, Michelle,
Alexander and Erick who took delight in my
triumphs and express great joy in my success.*

Julio B. Clempner

*This book is dedicated to my teachers and my
students.*

Alexander Poznyak

Preface

Markov systems have a wide range of applications, including modeling the behavior of macroeconomic systems, computer communication networks, manufacturing processes, and computer operating systems. These models have logically inferred the possibility of the Markov property. As a result, stochastic dynamic programming has been the focus of several journal articles and books as an optimization strategy that may be used in the Markov situation.

Here we consider the class of ergodic finite controllable Markov's chains. The core idea behind the method, described in this book, is to "immerse" the original discrete optimization problems (or game models) in the space of randomized formulations, where the variables stand in for the distributions (mixed strategies or preferences) of the original discrete (pure) strategies in the use. The following assumptions are made: a finite state space, a limited action space, continuity of the probabilities and rewards associated with the actions, and a necessity for accessibility. These hypotheses lead to the existence of an optimal policy. The best course of action is always stationary. It is either simple (i.e., non-randomized stationary) or composed of two non-randomized policies, which is equivalent to randomly selecting one of two simple policies throughout each epoch by tossing a biased coin. As an added bonus, the optimization procedure just has to repeatedly solve the time-average dynamic programming equation, making it theoretically feasible to choose the optimum course of action under the global restriction. In the ergodic cases the state distributions, generated by the corresponding transition equations, exponentially quickly converge to their stationary (final) values. This makes it possible to employ all widely used optimization methods (such as Gradient-like procedures, Lagrange's multipliers, and Tikhonov's regularization), including the related numerical techniques.

This book consists of 12 chapters:

- Chapter 1 describes a class of discrete-time, controllable, and ergodic Markov chains and to convert the nonlinear optimization Markov problem into a linear one and make the problem tractable, we propose to use an auxiliary c -variable.
- Chapter 2 presents a multi-objective Pareto front solution for a particular type of discrete-time ergodic controllable Markov chains considering the regularized

penalty method to identify the Pareto policies along the Pareto frontier based on Tikhonov's regularization method.

- Chapter 3 focuses on the design of an observer for a class of partially observable ergodic homogeneous finite Markov chains, where the goal is to derive the formulas for computing an observer and, as a consequence, the best control strategy.
- Chapter 4 extends the c -variable approach by adding a new linear constraint for continuous-time Markov chains, where this method's benefit is that it transforms the continuous-time Markov decision process problem into a discrete-time Markov decision process, such that the linear constraints make the problem computationally tractable.
- Chapter 5 provides an approach to locating the strong Nash and Stackelberg equilibrium based on a method that depends on identifying a scalar λ^* and the associated strategies $d^*(\lambda^*)$ fixing particular boundaries (*min* and *max*) that belong to the Pareto front, which refer to limits placed by the player over the Pareto front that form a specific decision region where the strategies can be selected.
- Chapter 6 demonstrates that the best-reply actions surely lead to an equilibrium point for a type of finite controlled Markov Chains dynamic games, providing a technique for creating a Lyapunov-like function that describes how players behave in a recurrent Markov-Lyapunov game that replaces the components of the ergodic system, which simulate players' anticipated behavior in one-shot games, for the recursive process.
- Chapter 7 suggests an analytical method for computing Bayesian incentive-compatible mechanisms where the private information is revealed following a class of controllable Markov games, including a new variable that denotes the outcome of the distribution vector, the strategies, and the mechanism design using a Reinforcement Learning methodology to calculate a mechanism that is nearly optimum and in equilibrium with the game's winning strategy.
- Chapter 8 offers a solution for Bayesian Partially Observable Markov Games (BPOMGs) supported by an AI strategy, based on a nucleus structure governed by three essential concepts: game theory, learning, and inference.
- Chapter 9 explores the theory behind bargaining games and offers a way for solving the game-theoretic models of bargaining put out by Nash and Kalai-Smorodinsky, which suggest a beautiful axiomatic solution to the problem based on several fairness standards.
- Chapter 10 provides a brand-new paradigm for combining game theory and the extraproximal approach to represent the multi-traffic signal-control synchronization problem, where the intersection's goal is to reduce queuing time, and finding the best signal timing strategy, or assigning a green period to each signal phase, which is a challenge for signal controllers.
- Chapter 11 presents a game that helps myopic participants achieve equilibrium as if they were forward-thinking agents. One of the game's main mechanics is that players are penalized for deviating from their prior best-reply plan as well as for the amount of time they spend to make decisions at each stage of play. Our chapter adds to existing research on typical myopic agent bargaining while also

broadening the class of processes and functions that may be used to define and apply Rubinstein's non-cooperative bargaining solutions.

- Chapter 12 suggests a solution to the transfer pricing problem. It analyzes a company with sequential transfers among its several divisions, where central management decides on the transfer price to maximize operational profitability. Throughout the negotiation process, the price shifting between divisions is a tool for bargaining.

This book is aimed at graduate students (Masters and Doctorate), who wish to learn more about how Markov chains theory solves different problems that arise in the real world.

Mexico City, Mexico

Julio B. Clempner
Alexander Poznyak

Acknowledgements

The presented material is based on more than 15 years of collaboration and teaching experience of the authors in Mexico Center for Research and Advanced Studies of the IPN (CINVESTAV), Automatic Control Department and School of Physics and Mathematics, National Polytechnic Institute, Mexico City, Mexico. The fundamental concepts of this course were created by world-renowned scientists such as John Nash, Harry Markowitz, John G. Kemeny and J. Laurie Snell, Ehud Kalai and Meir Smorodinsky, Andrey Tikhonov, and others.

The authors would like to express their wide thanks to the colleagues from the CINVESTAV and from the National Polytechnic Institute as well as their Mexican ex-Ph.D. students for their kind collaboration in the development of this book. Of course, we feel deep gratitude to our students over the years, without whose evident enjoyment and expressed appreciation the current book would not have been undertaken. Finally, we wish to acknowledge the editors at Springer for being so cooperative during the production process.

Mexico City, Mexico

Julio B. Clempner
Alexander Poznyak

Contents

1	Controllable Markov Chains	1
1.1	Finite Markov Chains	1
1.1.1	General Properties	1
1.1.2	Ergodic Markov Chains	4
1.1.3	Transition Equation and Invariant State Distribution	7
1.2	Controllable Markov Chain	7
1.2.1	Control Policy	7
1.2.2	Main Definition	8
1.3	Cost Functions for Markov Chains	9
1.3.1	Average Cost Function	9
1.3.2	Auxiliary c-Variables	9
1.4	Markov Decision Process as Linear a Programming Problem	10
1.4.1	Lineal Programming Formulation	10
1.4.2	Realization of Stationary Strategies in the State Space	12
1.4.3	Numerical Example	15
	References	15
2	Multiobjective Control	17
2.1	Introduction	17
2.1.1	Pareto Set	17
2.1.2	Parametrization of All Pareto Points	19
2.1.3	Utopia Point	21
2.2	Regularized Penalty Function Optimization Method (RPFOM)	23
2.2.1	Poly-Linear Optimization Problem Formulation	23
2.2.2	Penalty Functions Approach	23
2.2.3	Expected Property of RPFA	24

2.2.4	The Main Result on the Extremal Points of the Penalty Functions	24
2.2.5	Recurrent Algorithm	25
2.2.6	Special Selection of the Parameters	26
2.2.7	Numerical Example: Pareto Front Calculation	27
2.3	Portfolio Optimization Problem	29
2.3.1	Problem Formulation	29
2.3.2	Markowitz Function	31
2.3.3	Numerical Example: A Rational Investor	32
2.4	Appendix	33
	References	44
3	Partially Observable Markov Chains	47
3.1	Introduction	47
3.1.1	Brief Review	47
3.2	Partially Observable Markov Chains	49
3.3	Formulation of the Problem	51
3.4	Description in the c-Variables	52
3.5	Computation of the Estimated Value by Measurable Realization: Projection Stochastic Approximation Procedure	55
3.5.1	Adaptive Algorithm	57
3.6	Numerical Example: A Partially Observable Markowitz Portfolio	57
	References	62
4	Continuous-Time Markov Chains	65
4.1	Introduction	65
4.1.1	Related Work	65
4.2	Continuous-Time Markov Chains	67
4.3	Programming Solver for CTMDP	71
4.3.1	The c-Variable Method	71
4.3.2	Linear Programming Solver	72
4.4	Chemical Reaction Markov Models	74
4.4.1	Example 1. Formation of the Amidogen Radical	74
4.4.2	Example 2. Proton Transfer, Hydration and Tautomeric Reaction of Anthocyanin Pigments	78
	References	83
5	Nash and Stackelberg Equilibrium	85
5.1	Optimization and Equilibrium	85
5.2	ε -Nash Equilibrium and Tanaka's Function	87
5.2.1	Individual Cost Function	87
5.2.2	Regularized Lagrange Function	89
5.2.3	Tanaka's Function	90
5.2.4	ASG Continuous-Time Algorithm	91

5.3	Extraproximal Method	93
5.3.1	Proximal Format	93
5.4	Numerical Example: Strong Nash Equilibrium in Pareto Front	94
5.4.1	Euler Approach	94
5.4.2	Numerical Data	96
5.5	The Stackelberg-Nash Equilibrium Concept	100
5.5.1	Specific Features	100
5.5.2	Individual Aims and Tanaka's Representation	101
5.5.3	Extraproximal Procedure	102
5.6	Convergence Analysis	104
5.6.1	Auxiliary Results	104
5.6.2	Main Convergence Theorem	105
5.7	Application Example: Four Supermarkets Chain	108
	References	113
6	Best-Reply Strategies in Repeated Games	115
6.1	Introduction	115
6.2	Preliminaries	117
6.2.1	Controllable Markov Decision Process	117
6.2.2	Game Description	119
6.3	Problem Formulation	120
6.3.1	The State-Value Function	120
6.3.2	The Recursive Matrix Form	122
6.4	Construction of a Lyapunov-Like Function	124
6.4.1	Recurrent Form for the Cost Function	125
6.4.2	The Lyapunov Function Design	125
6.5	Examples	127
6.5.1	Example 1 (Banks Marketing Planning as Prisoner's Dilemma)	127
6.5.2	Example 2 (Duel Game)	129
	References	134
7	Mechanism Design	137
7.1	Introduction	137
7.1.1	Brief Review	137
7.2	Description of the Model	140
7.3	Mechanism and Equilibrium	142
7.4	Reinforcement Learning Approach	146
7.5	Risk-Averse Agents Strategies in Contracting Problem	150
	References	153

8 Joint Observer and Mechanism Design	155
8.1 Introduction	156
8.1.1 Brief Problem Analysis	156
8.1.2 Related Work	157
8.2 Markov Games	159
8.3 Problem Formulation	162
8.3.1 Initial Problem	162
8.3.2 Auxiliary Problem	162
8.4 Relation of Solutions for Initial and Auxiliary Problems	163
8.4.1 Ergodicity Condition	165
8.5 Reinforcement Learning Approach	167
8.5.1 Iterative Procedure	167
8.5.2 Learning Algorithm	169
8.6 Nash Equilibrium as a Solution of a Max-Min Problem	170
8.7 Application: Patrolling	172
8.7.1 Description of the Patrolling Problem	172
8.7.2 Solver	173
8.7.3 Random Walk Problem Formulation	174
8.7.4 Resulting Values	175
References	181
9 Bargaining Games or How to Negotiate	185
9.1 Introduction	185
9.1.1 Brief Review	185
9.1.2 Related Work	186
9.1.3 Nash Versus Kalai-Smorodinsky	189
9.2 Motivation	190
9.3 Preliminaries	192
9.4 The Nash Bargaining Model	196
9.5 The Kalai-Smorodinsky Bargaining Model	198
9.5.1 The n -Person Kalai-Smorodinsky Solution	199
9.6 The Bargaining Solver	201
9.6.1 The Nash Bargaining Solver	201
9.6.2 Kalai-Smorodinsky Solver	202
9.6.3 The Extraproximal Solver Method	204
9.7 The Model for the Disagreement Point	208
9.7.1 The Extraproximal Solver Method	210
9.8 Numerical Illustration	211
9.8.1 Computing the Disagreement Point	213
9.8.2 Computing the Nash Bargaining Solution	214
9.8.3 Computing the Kalai-Smorodinsky Bargaining Solution	216
References	218

10 Multi-traffic Signal-Control Synchronization	221
10.1 Introduction	222
10.1.1 Brief Review	222
10.1.2 Related Work on Traffic Control	222
10.2 Preliminaries	225
10.3 Nash Equilibrium	226
10.3.1 The Regularized Lagrange Principle Application	228
10.3.2 The Proximal Format	229
10.3.3 The Extraproximal Method	229
10.4 Traffic-Signal-Control Problem Formulation	230
10.4.1 Transition Matrix	231
10.4.2 Ergodicity	237
10.4.3 Cost Function	237
10.5 Gradient Solver	238
10.6 Application Example	240
References	245
11 Non-cooperative Bargaining with Unsophisticated Agents	249
11.1 Introduction	249
11.1.1 Related Literature	252
11.2 The Rubinstein's Alternating-Offers Model	254
11.3 Bargaining with Unsophisticated Players	256
11.4 An Extension to Continuous-Time Markov Chains	259
11.4.1 Solution Method	261
11.4.2 The Pareto Optimal Solution of the Bargaining Problem	262
11.4.3 The Non-cooperative Bargaining Solution	264
11.5 Numeric Simulations	266
11.5.1 Division of a Fix Resource	266
11.6 Extensions	268
11.6.1 Bargaining Under Different Discounting	268
11.6.2 Bargaining with Collusive Behavior	270
11.7 Appendix: Proofs	274
11.7.1 The Non-cooperative Bargaining Game	274
11.7.2 Formulation of the Problem	276
11.7.3 Convergence Analysis	277
11.7.4 Convergence Conditions of δ and α	285
References	287
12 Transfer Pricing as Bargaining	289
12.1 Introduction	289
12.1.1 Transfer Pricing Process	289
12.1.2 Brief Review	290
12.2 Preliminaries	293
12.2.1 Nash's Bargaining	293
12.2.2 Continuous-Time Bargaining	294

12.3	Transfer Pricing	296
12.4	The Transfer Pricing Nash Bargaining Solution	300
12.5	Transfer Price Bargaining Solver with Additional Constraints	303
12.5.1	Numerical Example for Nash's Bargaining Transfer Pricing	305
12.6	Continuous-Time Transfer Pricing	310
12.6.1	Revenue of a Passenger Between Members of an Airline Alliance	310
12.7	Extensions	314
12.7.1	Bargaining Under Different Discounting	314
12.7.2	Bargaining with Collusive Behavior	318
	References	325
	Index	329

Chapter 1

Controllable Markov Chains



Abstract In this chapter, we describe a class of discrete-time, controllable, and ergodic Markov chains. In these concepts, time and space are discrete. We start by outlining the fundamental model. Its single-step transition probabilities are then utilized to determine if a Markov chain is ergodic. In order to convert the nonlinear optimization Markov problem into a linear one and make the problem tractable, we propose to use an auxiliary c -variable. With such a setup, a discrete time Markov chain simulation is shown. A piece of the illustrative Markov chain laws for discrete time is constructed in the end.

1.1 Finite Markov Chains

1.1.1 General Properties

Let $\mathbb{N} = \{0, 1, \dots\}$ and let \mathbb{R}^n be an n -dimensional Euclidean space, $\mathbb{R} = \mathbb{R}^1$. Let us define a probability space (Ω, \mathcal{F}, P) where Ω is a set of elementary events, \mathcal{F} is the minimal σ -algebra of the subsets of Ω , and P is a given probability measure defined on $\mathcal{A} \in \mathcal{F}$. Let us also consider the natural sequence $t = 0, 1, 2, \dots$ ($t \in \mathbb{N}$) as a time argument. Let S be a finite set consisting of states $\{s_1, \dots, s_N\}$, $N \in \mathbb{N}$, called the *state space*.

Definition 1.1 A *Markov chain*¹ [5, 6, 9] is a sequence of N -valued random variables $s(t)$, $t \in \mathbb{N}$, satisfying the **Markov condition**:

$$P(s(t+1) = s_j | s(t) = s_i, s(t-1) = s_{i_{t-1}}, \dots, s(0) = s_{i_0}) = P(s(t+1) = s_j | s(t) = s_i) := \pi_{j|i}(t), \quad (1.1.1)$$

¹ Markov chains are named in honor of the Russian mathematician, Alexander Markov (1856–1922), who did pioneering work in the definition and analysis of this class of processes.

namely, only the current state determines the conditional probability that the system will be in the state s_j at some future time under the specified prehistory. This phenomenon may be demonstrated with a physical example in which a particle's future dynamics depend only on this state and not on how it came to be in its current state. We call (1.1.1) the Markov chain condition. The whole theory of Markov chains flows out of this condition.

Remark 1.1 In particular, (1.1.1) a general property called the *Markov property*, which implies that at any time t the past states $s(t-1), \dots, s(0)$ and future state $s(t+1)$ of the process are conditionally independent, given the present $s(t)$.

The (1.1.1) is a consequence of the simple identity given by $P(A \cap B) = P(A|B)P(B)$, such that $A = \{s(t+1) = s_j\}$ and $B = \{s(0) = s_{i_0}, \dots, s(t) = s_i\}$. If t is thought of as the present time, the expression

$$P(s(t+1) = s_j | s(t) = s_i, s(t-1) = s_{i_{t-1}}, \dots, s(0) = s_{i_0}) \text{ for } t \geq 0,$$

in (1.1.1) is a conditional probability for the value of the process one step ahead into the future, given its entire past history. We call it a *one-step-ahead conditional probability*. As we see, from (1.1.1), these conditional probabilities link the joint distribution of $(s(0) = s_{i_0}, \dots, s(t) = s_i, s(t+1) = s_j)$ to that of $(s(0) = s_{i_0}, \dots, s(t) = s_i)$, and thus they determine how joint distributions evolve as the time goes forward. Then, we can model a stochastic process by specifying its one-step-ahead conditional probabilities. In practice, they can often be deduced directly from assumptions on the physical nature of the process.

Definition 1.2 If the probability $\pi_{j|i}(t) = P(s(t+1) = s_j | s(t) = s_i)$ does not depend on t , i.e. $\pi_{j|i}$, then the Markov chain is said to be **time-homogeneous**.

The random variables $s(t)$ are defined on the probability space (Ω, \mathcal{F}, P) and take values in S . The stochastic process $\{s(t), t \in \mathbb{N}\}$ is assumed to be a Markov chain. The Markov chain can be represented by a complete graph (see Fig. 1.1) whose nodes are the states, where each edge $(s_i, s_j) \in S^2$ is labeled by the transition probability. The sum of the probabilities of $\pi_{j|i}$ over all states equals 1, that is,

$$\sum_{j=1}^N \pi_{j|i}(t) = \sum_{j=1}^N P(s(t+1) = s_j | s(t) = s_i) = 1.$$

(1.1.2)

A *stochastic matrix* is any square matrix of non-negative entries each of whose rows sums to 1. Then, state transition matrices are stochastic matrices. Conversely, any stochastic matrix is a valid model for the transition probabilities of a Markov chain.

The matrix $\Pi = (\pi_{j|i})_{(s_i, s_j) \in S} \in [0, 1]^{N \times N}$ determines the evolution of the time-homogeneous chain: for each $t \in \mathbb{N}$, the power Π^t has in each entry (s_i, s_j) the probability of going from state s_i to state s_j in exactly t steps. In particular, we have

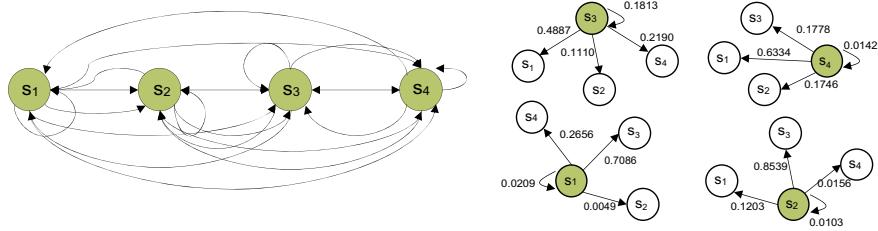


Fig. 1.1 Markov chain graph

that $\Pi^1 = \Pi$. The probability, starting in state i , of going to state j in two steps is the sum over l of the probability of going first to l and then to j . Using the Markov condition [7] we have

$$\pi_{j|i} = \sum_{l=1}^N (\pi_{j|l}) (\pi_{l|i}).$$

It can be seen that this is just the ij term of the product of the matrix Π with itself, this means that $(\pi_{j|i})^2$ is the (i, j) element of the matrix Π^2 .

Proposition 1.1 *The Chapman-Kolmogorov equation is given by*

$$\boxed{\Pi^{t+n} = \Pi^t \Pi^n}, \quad (1.1.3)$$

where Π^t is the t -step transition probability matrix and Π is the one-step transition matrix.

Proof By induction, for $t = 0$, we have that $\Pi^0 \Pi = I \Pi$. Let us suppose that the proposition is true for time t . Then, for time $t + 1$ we have

$$\Pi^{t+1} = \sum_{l=1}^N (\pi_{j|l})^t (\pi_{l|i})^1 = \sum_{l=1}^N (\pi_{j|l})^t \pi_{l|i} = \Pi^t \Pi,$$

and for n we have that

$$(\pi_{j|l})^{t+n} = \sum_{l=1}^N (\pi_{j|l})^t (\pi_{l|i})^n,$$

or in matrix form we have

$$\Pi^{t+n} = \Pi^t \Pi^n.$$

□

1.1.2 Ergodic Markov Chains

Our results are based on the following Theorems and Lemmas (for the proof see [9]).

Theorem 1.1 (the ergodic theorem) *Let for some state $j \in (1, \dots, N)$ of a homogeneous (stationary) Markov chain with the transition matrix Π and some $t > 0$, $\xi \in (0, 1)$ for all $i \in \mathcal{G}$*

$$\tilde{\pi}_{j|i}(t) := P(s(t) = s_{j_0} | s(0) = s_i) = \prod_{\tau=1}^t \pi_{s(\tau)|s(\tau-1)} \geq \xi > 0. \quad (1.1.4)$$

Then for any initial state distribution $P\{s(0) = s_i\}$ and for any $i, j = 1, \dots, N$ there exists the limit

$$p_j^* := \lim_{t \rightarrow \infty} \tilde{\pi}_{j|i}(t) > 0, \quad (1.1.5)$$

such that for any $t \geq 0$ this limit is reachable with an exponential rate, namely,

$$|\tilde{\pi}_{j|i}(t) - p_j^*| \leq (1 - \xi)^t = e^{-\alpha t}, \quad (1.1.6)$$

where $\alpha := |\ln(1 - \xi)|$.

Proof (a) For any $t \geq 0$ define

$$q_j(t) := \inf_{i=1, N} \pi_{i,j}(t) \text{ and } Q_j(t) := \sup_{i=1, N} \pi_{i,j}(t),$$

which evidently satisfy

$$q_j(t) \leq \pi_{i,j}(t) \leq Q_j(t)$$

for any $i, j = 1, \dots, N$ and any $t \geq 0$. Show that $q_j(t)$ monotonically increases and $Q_j(t)$ monotonically decreases such that

$$Q_j(t) - q_j(t) \xrightarrow{t \rightarrow \infty} 0 \quad (1.1.7)$$

since, having (1.1.7), we obtain (1.1.5). Using the property of Markov chain we have

$$q_j(r+t) := \inf_{i=1, N} \sum_{k=1}^N \pi_{i,k}(r) \pi_{k,j}(t) \geq q_j(t) \inf_{i=1, N} \sum_{k=1}^N \pi_{i,k}(t) = q_j(t),$$

$$Q_j(r+t) := \sup_{i=1, N} \sum_{k=1}^N \pi_{i,k}(r) \pi_{k,j}(t) \leq Q_j(t) \sup_{i=1, N} \sum_{k=1}^N \pi_{i,k}(r) = Q_j(t).$$

Next, for $0 \leq n \leq t$

$$\begin{aligned}
Q_j(t) - q_j(t) &= \sup_{i=1,N} \pi_{i,j}(t) + \sup_{l=1,N} [-\pi_{l,j}(t)] = \sup_{i,l=1,N} [\pi_{i,j}(t) - \pi_{l,j}(t)] = \\
\sup_{i,l=1,N} \sum_{k=1}^N [\pi_{i,k}(h) - \pi_{l,k}(h)] \pi_{k,j}(t-h) &= \sup_{i,l=1,N} \left\{ \begin{array}{l} \sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)] \pi_{k,j}(t-h) + \\ \sum_{k \in \mathcal{G}_-} [\pi_{i,k}(h) - \pi_{l,k}(h)] \pi_{k,j}(t-h) \end{array} \right\} \leq \\
\sup_{i,l=1,N} \left\{ Q_j(t-h) \sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)] + q_j(t-h) \sum_{k \in \mathcal{G}_-} [\pi_{i,k}(h) - \pi_{l,k}(h)] \right\}.
\end{aligned}$$

Here

$$\mathcal{G}_+ := \{k = 1, \dots, N : \pi_{i,k}(h) - \pi_{l,k}(h) \geq 0\},$$

$$\mathcal{G}_- := \{k = 1, \dots, N : \pi_{i,k}(h) - \pi_{l,k}(h) < 0\}.$$

So, evidently that

$$\begin{aligned}
\sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)] + \sum_{k \in \mathcal{G}_-} [\pi_{i,k}(h) - \pi_{l,k}(h)] &= \\
\sum_{k \in \mathcal{G}_+} \pi_{i,k}(h) + \sum_{k \in \mathcal{G}_-} \pi_{i,k}(h) - \sum_{k \in \mathcal{G}_+} \pi_{l,k}(h) - \sum_{k \in \mathcal{G}_-} \pi_{l,k}(h) &= \\
\sum_{k \in \mathcal{G}} \pi_{i,k}(h) - \sum_{k \in \mathcal{G}} \pi_{l,k}(h) &= 1 - 1 = 0,
\end{aligned}$$

and therefore

$$Q_j(t) - q_j(t) \leq [Q_j(t-h) - q_j(t-h)] \sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)]. \quad (1.1.8)$$

Now notice that if $j_0 \notin \mathcal{G}_+$ then

$$\sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)] \leq \sum_{k \in \mathcal{G}_+} \pi_{i,k}(h) \leq 1 - \pi_{i,j}(h_0) \leq 1 - \delta,$$

and if $j_0 \in \mathcal{G}_+$ then

$$\sum_{k \in \mathcal{G}_+} [\pi_{i,k}(h) - \pi_{l,k}(h)] \leq \sum_{k \in \mathcal{G}_+} \pi_{i,k}(h) - \pi_{i,j_0}(h) \leq 1 - \delta,$$

For $h = 1$ (1.1.8) leads to

$$Q_j(t) - q_j(t) \leq (1 - \delta) [Q_j(t-1) - q_j(t-1)].$$

Iterating back this inequality t -times and using the estimate we get

$$Q_j(t) - q_j(t) \leq (1 - \delta)^t, \quad (1.1.9)$$

which proves (1.1.7) and, consequently, (1.1.5).

(b) In view of the inequality

$$|\pi_{i,j}(t) - p_j^*| \leq Q_j(t) - q_j(t),$$

and using (1.1.9) we obtain (1.1.6). Theorem is proven. \square

Corollary 1.1 (on a stationary state distribution) *Suppose that (1.1.4) holds. Then for any $j = 1, \dots, N$ and for any*

$$p_j(t) := P\{s(t) = s_j\}, \quad (1.1.10)$$

the following property holds

$$|p_j(t) - p_j^*| \leq (1 - \delta)^t, \quad (1.1.11)$$

where p_j^* as in (1.1.5).

Proof The existence of \mathbf{p}^* follows from Theorem 1.1, and the formula (1.1.11) results from

$$\begin{aligned} |\mathbf{p}_t(t) - \mathbf{p}_j^*| &= \left| \sum_{i \in \mathcal{G}} \pi_{i,j}(t) p_i(0) - p_j^* \right| = \left| \sum_{i=1}^N [\pi_{i,j}(t) - p_j^*] p_i(0) \right| \leq \\ &\sum_{i=1}^N |\pi_{i,j}(t) - p_j^*| p_i(0) \leq (1 - \delta)^t \sum_{i=1}^N p_i(0) = (1 - \delta)^t. \end{aligned}$$

which proves the corollary. \square

Corollary 1.2 Since $\tilde{\pi}_{j_0|i}(t) = (\Pi^t)|_{ij_0}$, to verify the property (1.1.4) it is sufficient to multiply Π by itself t times up to the moment when all elements of at least one row will be positive.

Definition 1.3 For a homogeneous finite Markov chain with transition matrix $\Pi = [\pi_{j|i}]_{i,j=1,\dots,N}$ the parameter $k_{erg}(t_0)$ defined by

$$k_{erg}(t_0) := 1 - \frac{1}{2} \max_{i,j=1,\dots,N} \sum_{m=1}^N |(\tilde{\pi}_{im}(t_0)) - (\tilde{\pi}_{jm}(t_0))| \in [0, 1]. \quad (1.1.12)$$

is said to be **coefficient of ergodicity** of this Markov chain at time t_0 , where

$$(\tilde{\pi}_{im}(t_0)) = P\{s(t_0) = s_m | s(1) = s_i\},$$

is the probability to evolve from the initial state $s_1 = s_i$ to the state $s_{t_0} = s_m$ after t_0 transitions.

Lemma 1.1 *The coefficient of ergodicity $k_{erg}(t_0)$ can be estimated from below as*

$$k_{erg}(t_0) \geq \min_{i=1,\dots,N} \max_{j=1,\dots,N} \tilde{\pi}_{j|i}(t_0).$$

Remark 1.2 If all the elements $\tilde{\pi}_{j|i}(t_0)$ of the transition matrix Π^{t_0} are positive, then the coefficient of ergodicity $k_{erg}(t_0)$ is also positive. Notice that there exist ergodic Markov chains with elements $\tilde{\pi}_{j|i}(t_0)$ equal to zero, but with a positive coefficient of ergodicity $k_{erg}(t_0)$. So, the coefficient of ergodicity is strictly positive, then the corresponding Markov's chain is said to be ergodic.

1.1.3 Transition Equation and Invariant State Distribution

The state probabilities $p_j(t)$ (1.1.10) satisfies the following transition equation

$$p_j(t+1) = \sum_{i=1}^N \pi_{ij}(t) p_i(t),$$

which in matrix form is

$$\boxed{p(t+1) = \Pi^\top(t) p(t)}, \quad (1.1.13)$$

where

$$p(t) = (p_1(t), \dots, p_N(t))^\top.$$

For ergodic homogeneous (stationary) Markov chain the final distribution $p^* = \lim_{t \rightarrow \infty} p(t)$ (1.1.5) and satisfies the relation

$$\boxed{p^* = \Pi^\top p^*.} \quad (1.1.14)$$

Remark 1.3 For ergodic homogeneous (stationary) Markov chain the final distribution p^* is unique.

Remark 1.4 Theorem 1.1 ensures that Π^* has a unique everywhere positive invariant distribution p^* and, it is equivalent to the existence of some t_0 , such that $\pi_{ij}^*(t_0) > 0$.

1.2 Controllable Markov Chain

1.2.1 Control Policy

Let A be a finite set consisting of actions $\{a_1, \dots, a_M\}$, $M \in \mathbb{N}$, called the **action space**. A trajectory of a Markov chain is a sequence $s(0), a(0), s(1), a(1), \dots$

Definition 1.4 • A *policy* $\{\phi(t)\}_{t \in \mathbb{N}}$ is a sequence of finite collections

$$\boxed{\phi(t) = (s(0), a(0), \dots, s(t-1), a(t-1), s(t))} \quad (1.2.1)$$

- A policy $\{d(t)\}_{t \in \mathbb{N}}$ is referred to as a *randomized control policy* if the matrix $d(t) := [d_{k|i}(t)] \in \mathbb{R}^{M \times N}$ has elements

$$\boxed{d_{k|i}(t) := P(a(t) = a_k | s(t) = s_i),} \quad (1.2.2)$$

which means the probability to apply the action $a(t) = a_k$ in the state $s(t) = s_i$. Evidently, the elements $d_{k|i}(t)$ satisfy for $t \in \mathbb{N}$ the condition

$$\boxed{d(t) \in \mathcal{D} = \left\{ d(t) \in \mathbb{R}^{M \times N} \mid d_{k|i}(t) \geq 0, \sum_{k=1}^M d_{k|i}(t) = 1 \right\}} \quad (1.2.3)$$

- If $d(t) = d = \text{Const}$, then $d \in \mathbb{R}^{M \times N}$ is said to be a *stationary randomize control policy*.

If each row of matrix d contains only one element equal to 1 and others elements equal to 0, then such policy becomes to be a non-randomized policy coinciding with (1.2.1).

1.2.2 Main Definition

Definition 1.5 A *controllable Markov chain* [1–4, 7, 8, 10] is a four-tuple

$$MC = (S, A, \{d(t)\}_{t \in \mathbb{N}}, \{\Pi(t)\}_{t \in \mathbb{N}}), \quad (1.2.4)$$

where S is a finite set of *states*, $S \subset \mathbb{N}$, A is a finite set of actions, $\{d(t)\}_{t \in \mathbb{N}}$ is a randomized control policy, and $\Pi(t) = [\pi_{j|ik}(t)]_{i,j=\overline{1,N}, k=\overline{1,M}}$ is a *controlled transition matrix* (3 dimensional) where

$$\boxed{\pi_{j|ik}(t) := P(s(t+1) = s_j | s(t) = s_i, a(t) = a_k),} \quad (1.2.5)$$

$\forall t \in \mathbb{N}$ represents the probability associated with the transition from state s_i to state s_j ($i, j = \overline{1, N}$) under the action a_k ($k = \overline{1, K}$).

Obviously, for all $i, j = \overline{1, N}, k = \overline{1, M}$

$$\boxed{\pi_{j|ik}(t) \geq 0, \sum_{j=1}^N \pi_{j|ik}(t) = 1.} \quad (1.2.6)$$

In view of the definitions above, we have

$$\boxed{\pi_{j|i}(t|d(t)) = \sum_{k=1}^M \pi_{j|ik}(t) d_{k|i}(t).} \quad (1.2.7)$$

The transition equation (1.1.13) in this case becomes

$$\boxed{p(t+1) = \Pi^\top(t|d(t)) p(t),} \quad (1.2.8)$$

where

$$\Pi^\top(t|d(t)) = [\pi_{j|i}(t|d(t))]. \quad (1.2.9)$$

1.3 Cost Functions for Markov Chains

1.3.1 Average Cost Function

Let us define v_{ijk} as the cost for changing from state i to state j applying action k .

Definition 1.6 Let us define $J(d)$ the **average cost function** for homogeneous Markov chain with transition matrix $[\pi_{j|ik}]$ and final distribution $p_i(d)$ under stationary randomize control policy $d = [d_{k|i}]$ as follows

$$\boxed{J(d) = \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M v_{ijk} \pi_{j|ik} d_{k|i} p_i(d),} \quad (1.3.1)$$

where $p_i(d)$ satisfies for all $i = \overline{1, N}$

$$\boxed{p_i(d) = \sum_{l=1}^N \sum_{k=1}^M \pi_{i|lk} d_{k|l} p_l(d).} \quad (1.3.2)$$

Remark 1.5 As it follows from (1.3.1) the average cost function $J(d)$ is an extremely nonlinear function of the elements $d_{k|i}$ of matrix d .

1.3.2 Auxiliary c -Variables

Let us denote by $c \in \mathbb{R}^{N \times M}$ the matrix with elements c_{ik} defined as

$$\boxed{c_{ik} = d_{k|i} p_i(d).} \quad (1.3.3)$$

The average cost function in terms of c-variable is

$$J(c) = \sum_{i=1}^N \sum_{k=1}^M \sum_{j=1}^N v_{ijk} \pi_{j|ik} c_{ik} = \sum_{i=1}^N \sum_{k=1}^M w_{ik} c_{ik}, \quad (1.3.4)$$

where

$$w_{ik} = \sum_{j=1}^N v_{ijk} \pi_{j|ik}. \quad (1.3.5)$$

Remark 1.6 The average cost function $J(c)$ is now a *linear function* of c .

Notice that in view of (1.2.3) the variables p_i and $d_{k|i}$ can be recovered by the following relations

$$p_i(d) = \sum_{k=1}^M c_{ik} \text{ and } d_{ik} = c_{ik} / \sum_{l=1}^M c_{il}. \quad (1.3.6)$$

In the ergodic case $\sum_{k=1}^M c_{ik} > 0$ and

$$c \in C_{adm} = \left\{ \begin{array}{l} \sum_{i=1}^N \sum_{k=1}^M c_{ik} = 1, c_{ik} > 0, \\ \sum_{k=1}^M c_{jk} = \sum_{i=1}^N \sum_{k=1}^M \pi_{j|ik} c_{ik}. \end{array} \right. \quad (1.3.7)$$

1.4 Markov Decision Process as Linear a Programming Problem

1.4.1 Lineal Programming Formulation

Definition 1.7 A *Markov Decision Process* (MDP) for homogeneous Markov chain models with stationary randomize control policy d (1.2.3) is a pair

$$MDP = (MC, J),$$

where MC is a controllable Markov chain (1.2.4) and J is a *cost function* (1.3.4).

Definition 1.8 We say that a MDP is *optimal* if it is a solution of the following optimization problem

$$J(c) \rightarrow \min_{c \in C_{adm}}, \quad (1.4.1)$$

under possible additional constraints

$$\boxed{\left. \begin{aligned} \Phi_{eq,l}(c) &= \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \tilde{v}_{ijk,l}^{eq} \pi_{j|ik} d_{k|i} p_i(d) = \\ &\quad \sum_{i=1}^N \sum_{k=1}^M \tilde{w}_{ik,l}^{eq} c_{ik} = b_{eq,l}, l = \overline{1, L_{eq}}, \\ \Phi_{ineq,m}(c) &= \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \tilde{v}_{ijk,m}^{ineq} \pi_{j|ik} d_{k|i} p_i(d) = \\ &\quad \sum_{i=1}^N \sum_{k=1}^M \tilde{w}_{ik,m}^{ineq} c_{ik} \leq b_{ineq,m}, m = \overline{1, L_{ineq}}. \end{aligned} \right\}} \quad (1.4.2)$$

The optimization problem given by (1.4.1) and (1.4.2) may be represented in a Lineal Programming (LP) format

$$\min_x f^\top x \text{ subject to } \begin{cases} A_{eq}x = b_{eq} \\ A_{ineq}x \leq b_{ineq} \\ lb \leq x \leq ub, \end{cases} \quad (1.4.3)$$

where

$$x = col[c_{ik}] := (c_{11}, c_{12} \dots, c_{1M}; c_{21}, c_{22} \dots, c_{2M}; \dots c_{N1}, c_{N2} \dots, c_{NM})^\top, \quad (1.4.4)$$

$$f = col[w_{ik}] := (w_{11}, w_{12} \dots, w_{1M}; w_{21}, w_{22} \dots, w_{2M}; \dots w_{N1}, w_{N2} \dots, w_{NM})^\top, \quad (1.4.5)$$

$$A_{eq} = \begin{bmatrix} BL_{eq}^{(1)} \\ BL_{eq}^{(2)} \\ BL_{eq}^{(3)} \end{bmatrix}, \quad (1.4.6)$$

$$BL_{eq}^{(1)} = [\delta_{(j,i)} - \pi_{j|ik}]_{j=\overline{1,N} i=\overline{1,N}} = \begin{bmatrix} 1 - \pi_{(1,11)} & \dots & -\pi_{(N,1,M)} \\ \dots & \dots & \dots \\ -\pi_{(N|11)} & \dots & 1 - \pi_{(N|N,M)} \end{bmatrix} \in \mathbb{R}^{N \times (NM)}, \quad (1.4.7)$$

$$BL_{eq}^{(2)} = [1 \dots 1] \in \mathbb{R}^{1 \times (NM)}, \quad (1.4.8)$$

$$BL_{eq}^{(3)} = \begin{bmatrix} \tilde{w}_{11,1}^{eq}, \tilde{w}_{12,1}^{eq} \dots, \tilde{w}_{1M,1}^{eq}; & \dots & \tilde{w}_{N1,1}^{eq}, \tilde{w}_{N2,1}^{eq} \dots, \tilde{w}_{NM,1}^{eq} \\ \tilde{w}_{11,L_{eq}}^{eq}, \tilde{w}_{12,L_{eq}}^{eq} \dots, \tilde{w}_{1M,L_{eq}}^{eq}; & \dots & \tilde{w}_{N1,L_{eq}}^{eq}, \tilde{w}_{N2,L_{eq}}^{eq} \dots, \tilde{w}_{NM,L_{eq}}^{eq} \end{bmatrix} \in \mathbb{R}^{L_{eq} \times (NM)}, \quad (1.4.9)$$

$$\begin{bmatrix} \tilde{w}_{11,1}^{ineq}, \tilde{w}_{12,1}^{ineq} \dots, \tilde{w}_{1M,1}^{ineq}; & \dots & \tilde{w}_{N1,1}^{ineq}, \tilde{w}_{N2,1}^{ineq} \dots, \tilde{w}_{NM,1}^{ineq} \\ \tilde{w}_{11,L_{ineq}}^{ineq}, \tilde{w}_{12,L_{ineq}}^{ineq} \dots, \tilde{w}_{1M,L_{ineq}}^{ineq}; & \dots & \tilde{w}_{N1,L_{ineq}}^{ineq}, \tilde{w}_{N2,L_{ineq}}^{ineq} \dots, \tilde{w}_{NM,L_{ineq}}^{ineq} \end{bmatrix} \in \mathbb{R}^{L_{ineq} \times (NM)}, \quad (1.4.10)$$

$$b_{eq} = [0, \dots 0; 1; b_{eq,N+2} \dots b_{eq,N+L_{eq}}]^\top \in \mathbb{R}^{N+1+L_{eq}}, \quad (1.4.11)$$

$$b_{ineq} \in \mathbb{R}^{L_{ineq}}. \quad (1.4.12)$$

The matrix block $BL_{eq}^{(1)}$ (1.4.7) corresponds to the constraints defined in (1.3.7) rewritten as

$$\sum_{k=1}^M \sum_{i=1}^N (c_{jk} - \pi_{j|ik} c_{ik}) = 0.$$

The matrix block $BL_{eq}^{(2)}$ (1.4.8) corresponds to the relation defined in (1.3.7)

$$\sum_{i=1}^N \sum_{k=1}^M c_{ik} = 1.$$

The lower bound and the upper bound for c-variables are given by

$$lb = 0 \leq c_{ik} \leq ub = 1. \quad (1.4.13)$$

Remark 1.7 All the constraint in LP problem (1.4.3) are *consistent*, that is, the set of arguments satisfying this constraints is not empty.

1.4.2 Realization of Stationary Strategies in the State Space

To realize a random process $s(t)$ that corresponds, for example, to the transition matrix

$$\pi_{j|i} = \begin{bmatrix} 0 & 0.6 & 0.4 \\ 1 & 0 & 0 \\ 0.5 & 0.25 & 0.25 \end{bmatrix}$$

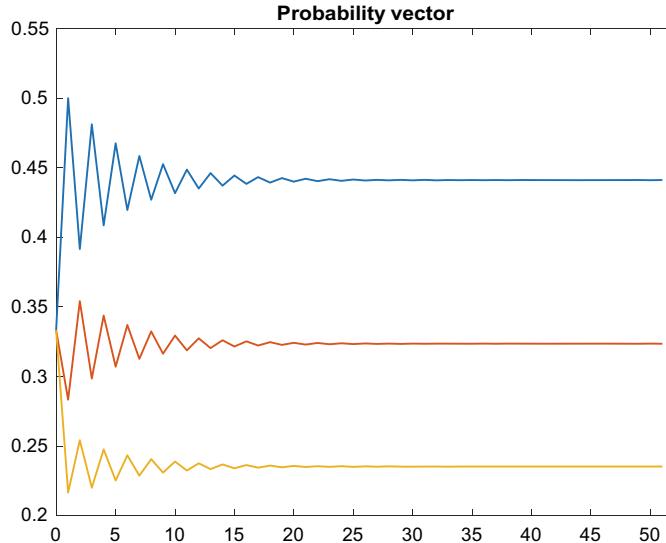


Fig. 1.2 Probability vector

with initial state distribution

$$p = [0.3333 \ 0.3333 \ 0.3333]$$

and recurrent equation given by (1.1.13)

$$p(t+1) = \Pi^\top(t)p(t),$$

we need to compute

$$p^* = \lim_{t \rightarrow \infty} p(t) = [0.4412; 0.3235; 0.2353].$$

The convergence of vector $p(t)$ is given in Fig. 1.2. At time $t = 1, \dots$ to generate the state vector $s(t)$ we use the generator of uniform distribution in the interval $[0, 1]$. Then, the output of the generator is projected in the interval $[0, 1]$ with subintervals proportional to the components of the $p(t)$ (see Fig. 1.3). The state s_i is associated with the number of the subinterval.

The corresponding Markov chain state process $s(t)$ is presented in Fig. 1.4.

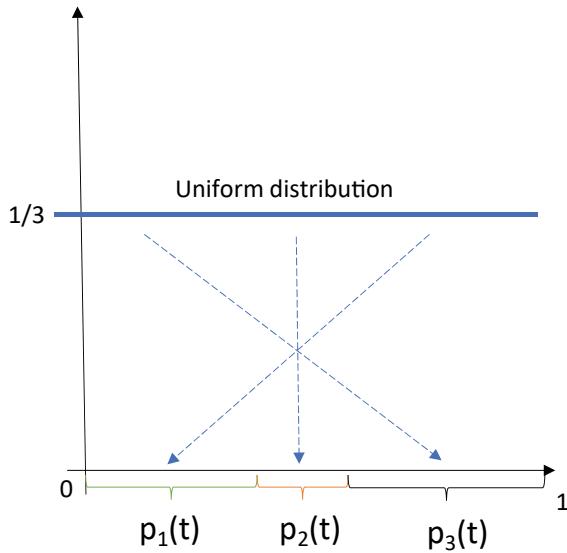


Fig. 1.3 State generator

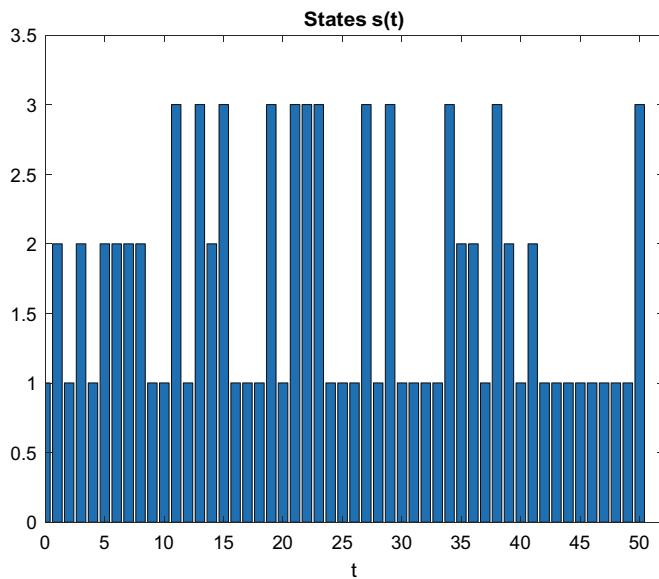


Fig. 1.4 States of the Markov chain

1.4.3 Numerical Example

Consider an environment where total number of states is $N = 3$ and the total number of actions is $M = 2$ for which the transition matrices are given by

$$\pi_{j|i1} = \begin{bmatrix} 0.7094 & 0.4984 & 0.0060 \\ 0.7547 & 0.1597 & 0.6991 \\ 0.2760 & 0.3404 & 0.0909 \end{bmatrix}, \quad \pi_{j|i2} = \begin{bmatrix} 0.4694 & 0.1656 & 0.0838 \\ 0.4119 & 0.6020 & 0.2290 \\ 0.3371 & 0.2630 & 0.9133 \end{bmatrix},$$

and the cost matrices are given by

$$v_{ij1} = \begin{bmatrix} 2.9411 & 7.8237 & 9.9358 \\ 2.9411 & 7.8237 & 9.9358 \\ 2.9411 & 7.8237 & 9.9358 \end{bmatrix}, \quad v_{ij2} = \begin{bmatrix} 4.3204 & 6.5513 & 8.5479 \\ 4.3204 & 6.5513 & 8.5479 \\ 4.3204 & 6.5513 & 8.5479 \end{bmatrix},$$

We get for optimal policy

$$p_i^* = \begin{bmatrix} 0.4514 \\ 0.4508 \\ 0.0979 \end{bmatrix}, \quad c_{ik}^* = \begin{bmatrix} 0.4514 & 0.0000 \\ 0.0000 & 0.4508 \\ 0.0979 & 0.0000 \end{bmatrix}, \quad d_{k|i}^* = \begin{bmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \\ 1.0000 & 0.0000 \end{bmatrix},$$

and $J(c) = 5.6395$.

References

1. Altman, E.: Constrained Markov Decision Processes: Stochastic Modeling. Routledge (1999)
2. Clempner, J.B.: A lyapunov approach for stable reinforcement learning. Comput. Appl. Math. **41**, 279 (2022)
3. Clempner, J.B., Poznyak, A.S.: Analysis of best-reply strategies in repeated finite markov chains games. In: 52nd IEEE Conference on Decision and Control (CDC), pp. 568–573. Firenze, Italy (2013)
4. Clempner, J.B., Poznyak, A.S.: Simple computing of the customer lifetime value: a fixed local-optimal policy approach. J. Syst. Sci. Syst. Eng. **23**(4), 439–459 (2014)
5. Howard, R.A.: Dynamic Programming and Markov Processes (1960)
6. Kemeny, J.G., Snell, J.L.: Finite Markov Chains: With a New Appendix “Generalization of a Fundamental Matrix”. Springer (1983)
7. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker, New York (2000)
8. Puterman, M.L.: Markov decision processes. Handbooks in Operations Research and Management Science, vol. 2, pp. 331–434 (1990)
9. Rozanov, Y.: Probability Theory: A Concise Course (1977)
10. Sragovich, V.G.: Mathematical Theory of Adaptive Control, vol. 4. World Scientific (2005)

Chapter 2

Multiobjective Control



Abstract A multi-objective Pareto front solution is presented in this chapter for a particular type of discrete-time ergodic controllable Markov chains. We offer a technique that, given specific boundaries, chooses the best multi-objective option for the Pareto frontier as a decision support system. We only consider a class of finite, ergodic, and controllable Markov chains while addressing this issue. The regularized penalty method utilizes a projection-gradient strategy to identify the Pareto policies along the Pareto frontier and is based on Tikhonov's regularization method. The goal is to make the parameters as efficient as possible while still maintaining the original form of the functional. After setting the initial value, we gradually reduce it until each policy closely resembles the Pareto policy. In this sense, we specify the precise direction of the parameter tendencies toward zero and establish the convergence of the gradient regularized penalty algorithm. The matching picture in the objective space receives a Pareto frontier of only Pareto policies thanks to our policy-gradient multi-objective algorithms, which, on the other hand, use a gradient-based strategy. In order to improve security when transporting cash and valuables, we empirically validate the technique by providing a numerical example of a genuine alternative solution to the vehicle routing planning problem. In addition, we describe a portfolio optimization and represent the Pareto frontier. The decision-making techniques investigated in this paper are consistent with the most widely used computational intelligent models in the *Artificial Intelligence* research field.

2.1 Introduction

2.1.1 Pareto Set

In practical areas arises the problem that several objective functions (outcomes) have to be optimized concurrently. For example, companies may take different approaches to maximize profit, minimize loss, minimize cost and maximize quality, etc. This example presents a dilemma: different objectives contradict each other and therefore do not have the same optima. The problem that comes up is how to compute all the

optimal compromises of this ***Multi-objective Optimization Problem*** (MOP) [4, 6, 14, 19, 20].

We will consider controllable homogeneous Markov chains under stationary policy. In a MOP there are given n objective functions J^1, \dots, J^n which have to be minimized:

$$J(d) = \min_{d \in \mathcal{D}} (J^1(d), \dots, J^n(d)) \quad (2.1.1)$$

over the class of all admissible policies (strategies) $d \in \mathcal{D}$ (see Chap. 1). There is no single solution for a nontrivial multi-objective optimization problem that concurrently optimizes all of the objectives.

Definition 2.1 The goal functions are said to be at *conflict* in that situation. If none of the objective functions can have their value decreased without increasing some of the other objective values, the solution is said to be nondominated, ***Pareto optimum***, Pareto efficient, or noninferior. There may be a (potentially infinite) number of Pareto optimum solutions, all of which are regarded as equally acceptable, without extra subjective preference information. If d^* minimizes $J(d)$ in the sense of Pareto, then d^* is said to be a *Pareto policy*.

A two dimensional case is represented in Fig. 2.1. If the internal part of the cone \mathcal{K} with vertex P those not contains any points from set Ω of all possible functions value, then this vertex is a Pareto optimum point. For instance, it is \mathcal{K}_2 with vertex P_2 . If it is not the case (\mathcal{K}_1 with vertex P_1), then this vertex point is not Pareto optimal.

In multi-objective optimization (2.1.1), there does not typically exist a feasible solution that minimizes all objective functions simultaneously. Therefore, attention is paid to Pareto optimal solutions, that is, solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives. In mathematical terms, a feasible solution $d_1 \in \mathcal{D}$ is said to (Pareto) dominate another solution $d_2 \in \mathcal{D}$, if

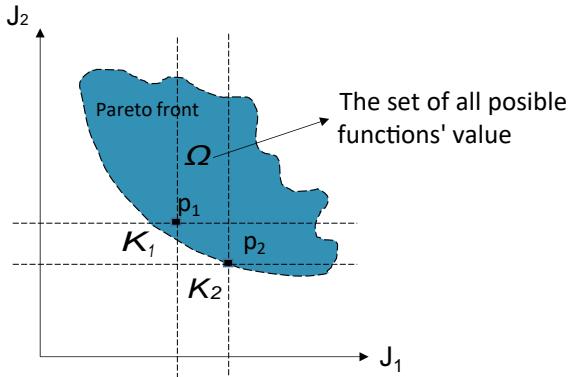
$$\boxed{\begin{aligned} \forall i \in \{1, \dots, n\}, J^i(d_1) &\leq J^i(d_2), \text{ and} \\ \exists i \in \{1, \dots, n\}, J^i(d_1) &< J^i(d_2). \end{aligned}} \quad (2.1.2)$$

A local Pareto optimal policy is a policy such that no improvement in all the objectives can be achieved by moving to a neighboring feasible point. The MOP is different from single-objective nonlinear programming because the set of Pareto optimal points is usually a continuum that may have disjoint components.

Definition 2.2 A solution $d^* \in \mathcal{D}$ (and the corresponding outcome $J(d^*)$) is called ***Pareto optimal*** if there does not exist another solution that dominates it. The set of Pareto optimal outcomes, denoted \mathcal{D}^* , is often called the ***Pareto front***, Pareto frontier, or Pareto boundary.

In this chapter we study discrete-time multi-objective Markov chains problem.

- Presents a formulation of the problem in terms of a nonlinear programming problem implementing the Lagrange principle.

Fig. 2.1 Pareto front

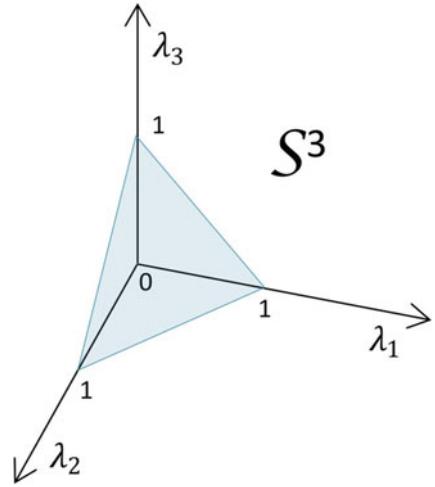
- Highlights a fundamental feature of the Pareto set where the search space is in most of the cases nonconvex (strict convex) and a complicated set because there is not a unique solution. For solving the existence and characterization of strong Pareto policies we employ the Tikhonov's regularization method [30, 31].
- Formulates the original problem considering several constraints:
 - (a) we employ the c -variable method for the introduction of linear constraints over the nonlinear problem and,
 - (b) we restrict the cost-functions allowing points in the Pareto front to have a small distance from one another.
- Solves the optimization problem using the projected gradient method.
- Suggests the convergence conditions and compute the estimate rate of convergence of variables θ and δ corresponding to the Lagrange principle and the Tikhonov's regularization respectively. We describe the dependence for the regularized Lagrange function on the regularizing parameters θ and δ , and analyses the asymptotic behavior when $\theta_n \downarrow 0$ and the fact that $\frac{\theta_n}{\delta_n} \downarrow 0$ also holds.
- Provides the details needed to implement the proposed method in an efficient way.

2.1.2 Parametrization of All Pareto Points

A different important problem arises when the individual cost-functions $J^1(d), \dots, J^n(d)$ are ranked in a hierarchical order: an optimal policy became a particular Pareto policy.

One of the fundamental problems are the existence and characterization of Pareto policies. This can be obtained via the usual *linear scalarization*¹ approach, in which

¹ By far most of the methods for the computation of single Pareto points or the entire Pareto set are based on a “scalarization” of the MOP (see e.g. [9, 10, 27, 29]).

Fig. 2.2 Simplex

the multi-objective Markov chain is reduced to a single-objective with a “weighted” objective function of the form

$$\Phi(\lambda, d) := \lambda^\top J(d) = \lambda_1 J^1(d) + \cdots + \lambda_n J^n(d), \quad (2.1.3)$$

where the vector λ is from n -dimensional simplex, that is,

$$\boxed{\lambda \in \mathcal{S}^n = \{\lambda_i \geq 0 | \sum_{i=1}^n \lambda_i = 1\}} \quad (2.1.4)$$

(see Fig. 2.2 for $n = 3$). Scalarizing a multi-objective optimization problem (2.1.1) is an a priori method, which means formulating a single-objective optimization problem such that optimal solutions to the single-objective optimization problem are Pareto optimal solutions to the multi-objective optimization problem.

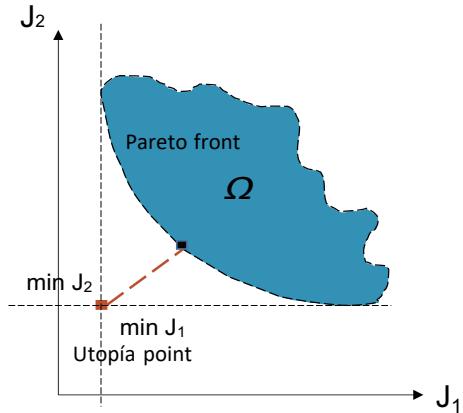
Theorem 2.1 ([2, 14, 19]) *For any weights combination $\lambda \in \mathcal{S}^n$ the point*

$$d^*(\lambda) \in \operatorname{Arg} \min_{d \in \mathcal{D}} \Phi(\lambda, d) \quad (2.1.5)$$

is a Pareto optimal, i.e., $d^(\lambda) \in \mathcal{D}^*$.*

Remark 2.1 If in the optimization problem (2.1.5) there exists a unique solution then

$$d^*(\lambda) = \arg \min_{d \in \mathcal{D}} \Phi(\lambda, d) \in \mathcal{D}^*. \quad (2.1.6)$$

Fig. 2.3 Utopia point

2.1.3 Utopia Point

If we let

$$J^{i*} = \inf_d J^i(d)$$

and define the *utopia minimum* as $J^*(d) = (J^{1*}(d), \dots, J^{n*}(d))$ (infeasible in general) the resulting problem is to find the Pareto policies d^* whose cost vector $J(d^*)$ is the “closest” to $J^*(d)$ in the usual Euclidean norm (see Fig. 2.3).

Rigorously, this problem can be formulated as the following optimization problem

$$\Phi(\lambda) := \sum_{i=1}^n [J^i(d^*(\lambda)) - J^{i*}]^2 \rightarrow \min_{\lambda \in \mathcal{S}^n},$$

(2.1.7)

where $d^*(\lambda)$ is defined by (2.1.6). The solution of the problem (2.1.7) can be obtained by applying the numerical Kiefer–Wolfowitz procedure [15] for the case when there is no noise in function observations:

$$\lambda^{(k)} = \mathcal{P}_{\mathcal{S}^n} \left\{ \lambda^{(k-1)} - \frac{\gamma}{\alpha} \sum_{j=1}^n [\Phi(\lambda^{(k-1)} + \alpha e_j) - \Phi(\lambda^{(k-1)})] \right\}, \quad (2.1.8)$$

where, γ, α are small positive parameters, $e_j = [\underbrace{0, \dots, 0}_j, 1, 0, \dots, 0]^T$ are test vectors and $k = 1, 2, \dots$. In (2.1.8) the operator $\mathcal{P}_{\mathcal{S}^n}$ is the projector to the simplex \mathcal{S}^n , which is a finite step procedure realized by the following procedure [24]:

Algorithm 2.1 Projector

```

function[ $\lambda$ ] = Projector( $\lambda$ )
n ← length( $\lambda$ );
e ← 0.0001;
R ← e * n;
AR ← 1 - R;
AK ← n;
SQ1 ← 1 - sum( $\lambda$ );
Mon ← zeros(1, n);
Mon ← zeros(1, n);
for i = 1 : n do
     $\lambda(i)$  ← ( $\lambda(i)$  - e) / AR;
end for
SQ1 ← SQ1 / AR;
 $\lambda$  ← Calc $\lambda$ (n,  $\lambda$ , AK, SQ1, Mon);
for i = 1 : n do
     $\lambda(i)$  ← e + AR *  $\lambda(i)$ ;
end for
end

```

Algorithm 2.2 Calc

```

function[ $\lambda$ ] = Calc $\lambda$ (n,  $\lambda$ , AK, SQ1, Mon)
 $\lambda$  = CalcNomalization $\lambda$ (n,  $\lambda$ , AK, SQ1, Mon);
for i = 1 : n do
    if Mon(i) == 0 &&  $\lambda(i)$  < 0.0000 then
        SQ1 ←  $\lambda(i)$ ;
         $\lambda(i)$  ← 0;
        Mon(i) ← 1;
        AK ← AK - 1;
         $\lambda$  ← Calc $\lambda$ (n,  $\lambda$ , AK, SQ1, Mon);
    end if
end for
end

```

Algorithm 2.3 CalcNomalization

```

function[ $\lambda$ ] = CalcNomalization $\lambda$ (n,  $\lambda$ , AK, SQ1, Mon)
for i = 1 : n do
    if Mon(i) == 0 then
         $\lambda(i)$  =  $\lambda(i)$  + SQ1 / AK;
    end if
end for
end

```

2.2 Regularized Penalty Function Optimization Method (RPFOM)

2.2.1 Poly-Linear Optimization Problem Formulation

Consider the following poly-linear programming problem

$$\begin{aligned}
 f(x) = & \alpha_1 \sum_{j_1=1}^N c_{j_1} x_{j_1} + \alpha_2 \sum_{j_1=1}^N \sum_{j_2=1}^N c_{j_1, j_2} x_{j_1} x_{j_2} + \\
 & \alpha_3 \sum_{j_1=1}^N \sum_{j_2=1}^N \sum_{j_3=1}^N c_{j_1, j_2, j_3} x_{j_1} x_{j_2} x_{j_3} + \dots + \\
 & \alpha_{N-1} \sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_{N-1}=1}^N c_{j_1, \dots, j_{N-1}} x_{j_1} \dots x_{j_{N-1}} + \\
 & \alpha_N \sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_N=1}^N c_{j_1, \dots, j_N} x_{j_1} \dots x_{j_N} \rightarrow \min_{x \in X_{adm}}
 \end{aligned} \tag{2.2.1}$$

$\alpha_j = \{0; 1\}$ ($j = 1, \dots, N$) are binary variables

$X_{adm} := \{x \in \mathbb{R}^N : x \geq 0, V_0 x = b_0 \in \mathbb{R}^{M_0}, V_1 x \leq b_1 \in \mathbb{R}^{M_1}\}$ is a bounded set.

Introducing the “slack” vectors $u \in \mathbb{R}^{M_1}$ with nonnegative components, that is, $u_j \geq 0$ for all $j = 1, \dots, M_1$, the original problem (2.2.1) can be rewritten as

$$\left. \begin{array}{l} \min_{x \in X_{adm}, u \geq 0} f(x) \\ \text{subject to} \\ X_{adm} := \{x \in \mathbb{R}^N : x \geq 0, V_0 x = b_0, V_1 x - b_1 + u = 0\} \end{array} \right\} \tag{2.2.2}$$

Note that this problem may have non-unique solution and $\det(V_0^\top V_0) = 0$. Define by $X^* \subseteq X_{adm}$ the set of all solutions of the problem (2.2.2).

2.2.2 Penalty Functions Approach

Following [11, 35] consider the *penalty function* [7, 8]

$$\boxed{\tilde{\mathcal{F}}_{k, \delta}(x, u) := f(x) + k \left[\frac{1}{2} \|V_0 x - b_0\|^2 + \frac{1}{2} \|V_1 x - b_1 + u\|^2 + \frac{\delta_1}{2} (\|x\|^2 + \|u\|^2) \right]} \tag{2.2.3}$$

where the parameters k and δ_1 are positive. Obviously, the unconstraint on x the optimization problem

$$\min_{x \in X_{adm}, u \geq 0} \tilde{\mathcal{F}}_{k,\delta}(x, u) \quad (2.2.4)$$

has a unique solution since the optimized function (2.2.3) is *strongly convex* if $\delta > 0$. Note also that

$$\arg \min_{x \in X_{adm}, u \geq 0} \tilde{\mathcal{F}}_{k,\delta}(x, u) = \arg \min_{x \in X_{adm}, u \geq 0} \mathcal{F}_{\mu,\delta}(x, u),$$

where $\mu := k^{-1} > 0$ and

$$\begin{aligned} \mathcal{F}_{\mu,\delta}(x, u) := & \mu f(x) + \frac{1}{2} \|V_0 x - b_0\|^2 + \\ & \frac{1}{2} \|V_1 x - b_1 + u\|^2 + \frac{\delta}{2} (\|x\|^2 + \|u\|^2). \end{aligned} \quad (2.2.5)$$

2.2.3 Expected Property of RPFA

Proposition 2.1 *If the penalty parameter μ and δ tends to zero by a particular manner, then we may expect that $x^*(\mu, \delta)$ and $u^*(\mu, \delta)$, which are the solutions of the optimization problem*

$$\min_{x \in X_{adm}, u \geq 0} \mathcal{F}_{\mu,\delta}(x, u)$$

tend to the set X^ of all solutions of the original optimization problem (2.2.2), that is,*

$$\rho \{x^*(\mu, \delta), u^*(\mu, \delta); X^*\} \xrightarrow[\mu \downarrow 0]{} 0 \quad (2.2.6)$$

where $\rho \{a; X^*\}$ is the Hausdorff distance defined as

$$\rho \{a; X^*\} = \min_{x^* \in X^*} \|a - x^*\|^2.$$

Below we define exactly how the parameters μ and δ should tend to zero to provide the property (2.2.6).

2.2.4 The Main Result on the Extremal Points of the Penalty Functions

Theorem 2.2 *Let us assume that*

- (1) *the bounded set X^* of all solutions of the original optimization problem (2.2.2) is not empty and the Slater's condition holds, that is, there exists a point $\hat{x} \in X_{adm}$ such that*

$$V_1 \dot{x} < b_1. \quad (2.2.7)$$

(2) The parameters μ and δ are time-varying, i.e.,

$$\mu = \mu_n, \delta = \delta_n \ (n = 0, 1, 2, \dots)$$

such that

$$0 < \mu_n \downarrow 0, \frac{\mu_n}{\delta_n} \downarrow 0 \text{ when } n \rightarrow \infty. \quad (2.2.8)$$

Then

$$\left. \begin{aligned} x_n^* &:= x^*(\mu_n, \delta_n) \xrightarrow{n \rightarrow \infty} x^{**} \\ u_n^* &:= u^*(\mu_n, \delta_n) \xrightarrow{n \rightarrow \infty} u^{**} \end{aligned} \right\} \quad (2.2.9)$$

where $x^{**} \in X^*$ is the solution of the original problem (2.2.2) with the minimal weighted norm which is unique, i.e.,

$$\|x^{**}\| \leq \|x^*\| \text{ for all } x^* \in X^* \quad (2.2.10)$$

and

$$u^{**} = b_1 - V_1 x^{**}. \quad (2.2.11)$$

Proof See Appendix 2.4. □

We also need the following lemma.

Lemma 2.1 Under the assumptions of the Theorem above there exist positive constants C_μ and C_δ such that

$$\|x_n^* - x_m^*\| + \|u_n^* - u_m^*\| \leq C_\mu |\mu_n - \mu_m| + C_\delta |\delta_n - \delta_m|. \quad (2.2.12)$$

Proof See Appendix 2.4. □

2.2.5 Recurrent Algorithm

General case

Consider the following recurrent procedure for finding the extremal point $z^{**} = \begin{pmatrix} x^{**} \\ u^{**} \end{pmatrix}$:

$$\left. \begin{aligned} z_n &= \left[z_{n-1} - \gamma_n \frac{\partial}{\partial z} \mathcal{P}_{\mu_n, \delta_n}(z_{n-1}) \right]_+, \quad z := \begin{pmatrix} x \\ u \end{pmatrix}, \\ \frac{\partial}{\partial z} \mathcal{P}_{\mu_n, \delta_n}(z_{n-1}) &= \begin{pmatrix} \frac{\partial}{\partial x} \mathcal{P}_{\mu_n, \delta_n}(x_{n-1}, u_{n-1}) \\ \frac{\partial}{\partial u} \mathcal{P}_{\mu_n, \delta_n}(x_{n-1}, u_{n-1}) \end{pmatrix} = \\ &\quad \begin{pmatrix} \mu_n \frac{\partial}{\partial x} f(x_{n-1}) + V_0^\top [V_0 x_{n-1} - b_0] + \\ V_1^\top [V_1 x_{n-1} - b_1 + u_{n-1}] + \delta_n x_{n-1} \\ V_1 x_{n-1} - b_1 + (1 + \delta_n) u_{n-1} \end{pmatrix}. \end{aligned} \right\} \quad (2.2.13)$$

Theorem 2.3 (on the convergence of the projection gradient method) *If*

$$\left. \begin{aligned} \sum_{n=0}^{\infty} \gamma_n \delta_n &= \infty, \quad \frac{\gamma_n}{\delta_n} \xrightarrow{n \rightarrow \infty} 0, \\ \frac{|\mu_n - \mu_{n-1}| + |\delta_n - \delta_{n-1}|}{\gamma_n \delta_n} &\xrightarrow{n \rightarrow \infty} 0, \end{aligned} \right\} \quad (2.2.14)$$

then

$$W_n := \|z_n - z_n^*\|^2 \xrightarrow{n \rightarrow \infty} 0. \quad (2.2.15)$$

Proof See the Appendix 2.4. □

2.2.6 Special Selection of the Parameters

Let us select the parameters of the algorithm (2.2.13) as follows:

$$\begin{aligned} \delta_n &= \begin{cases} \delta_0 & \text{if } n \leq n_0 \\ \delta_0 \frac{[1+\ln(n-n_0)]}{(1+n-n_0)^\delta} & \text{if } n > n_0 \end{cases}, \quad \mu_n = \begin{cases} \mu_0 & \text{if } n < n_0 \\ \frac{\mu_0}{(1+n-n_0)^\mu} & \text{if } n \geq n_0 \end{cases}, \\ \gamma_n &= \begin{cases} \gamma_0 & \text{if } n < n_0 \\ \frac{\gamma_0}{(1+n-n_0)^\gamma} & \text{if } n \geq n_0 \end{cases}, \quad \delta, \mu, \gamma > 0, \delta_0, \mu_0, \gamma_0 > 0. \end{aligned} \quad (2.2.16)$$

To guarantee the convergence of the suggested procedure, by the property $\frac{\mu_n}{\delta_n} \xrightarrow{n \rightarrow \infty} 0$ and by the conditions (2.2.14) we should have

$$\delta \leq \mu, \gamma \geq \delta, \gamma + \delta \leq 1. \quad (2.2.17)$$

2.2.7 Numerical Example: Pareto Front Calculation

This example presents a real alternative solution of the vehicle routing planning problem to increase security in transportation of cash and valuables (see, [1, 3, 5] for multi-objective Markov decision in urban modeling). Companies in the arena are focused in the physical transfer of cash, jewels, coins, etc. As a result of the type of the transported goods the security vehicles are frequently exposed to attacks along their routes. Crime is a significant challenge. In addition, the risk rates and the losses are different from sector to sector. A conflict arises because higher risk exposures allow a reduction of the travel cost. We suggest a bi-objective formulation using the proposed method with the goal of reducing both the risk and the travel cost. The problem is weighted according to which a maximum amount of valuables can be transported the security vehicle.

It is important to note that each customer must be assigned to exactly one of the routes and the vehicle capacity must not be exceeded. Assuming that a vehicle picks up cash and valuables at certain places i visited along route, a risk index φ for each criminal l can be defined as follows

$$\sum_k \varphi_{ik}^{-l} c_{ik}^{-l*}(\lambda) = \Phi_i^{-l},$$

where φ_{ik}^{-l} is the risk of criminal $-l$ to attack place i and action k along a given route and, Φ_i^{-l} is the maximum risk able to undertaken by a criminal $-l$ to attack place i . The risk method assesses the probabilities for the actions of the attackers and it is defined as follows

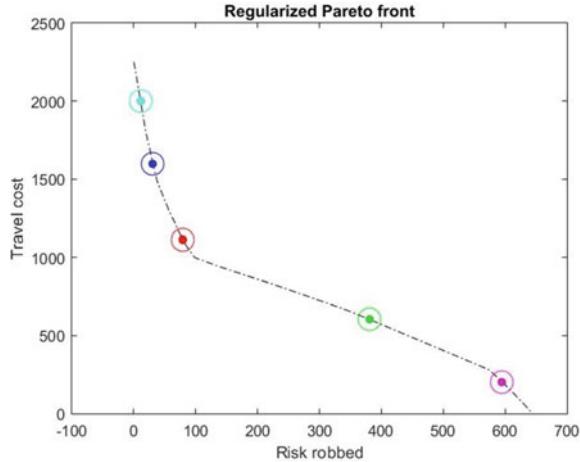
$$\begin{aligned} & \arg \min_{\lambda \in \Lambda^N} J^l(c^{l*}(\lambda)) \\ & \Phi_{prev}^{-l} \leq J^{-l}(c^{-l*}(\lambda)) \text{ and} \\ & J^{-l}(c^{-l*}(\lambda)) \leq \Phi^{-l} \text{ for } -l = \{\overline{1, q}\} \setminus \{l\}. \end{aligned} \tag{2.2.18}$$

It is interesting to remark that in our case we used i and j as the sequence between two consecutive places (states) that measure the probability of an attack on a specific roadway segment. In terms of Markov chains we observe the current state $i \in S$. Then, by optimizing the risk using Eq. (2.2.18) is selected an optimal action $a(n) = a_{(k)} \in A(s)$. Then two things happen: a cost J_{ijk} is incurred and, the system at time $n + 1$ moves to a new state $j \in S$ with probability $\pi_{j|ik}$.

Let us consider $N = 6$ and $M = 2$. Then, fixing $\delta = 0.1$ and $\mu = 0.001$ the Pareto front for 1000 points is shown in Fig. 2.4 which represents the routing plans that should both be safe and efficient. The Pareto front allows the decision-maker to deal with two critical problems:

- (a) the minimization of the traveled cost-time of a security vehicle and,
- (b) the reduction of the expected exposure of the transported goods to robberies.

Fig. 2.4 Regularized Pareto front: security routing against vehicle attack



This is not a simple assignment. It involves the conflict between objectives and the difficulty of selecting the routes which implicate to visit and to collect valuables along several places every day. The corresponding value of the vector λ and the joint strategy $c_{ik}^{(l)}$ for $l = 1, 2$ is as follows:

Route 1 (color cyan)

$$\lambda_1 = 0.8120, \quad \lambda_2 = 0.1880,$$

$$c_{ik}^{(1)} = \begin{bmatrix} 0.1164 & 0.1059 \\ 0.1115 & 0.0904 \\ 0.1356 & 0.0122 \\ 0.1487 & 0.0070 \\ 0.0711 & 0.1039 \\ 0.0957 & 0.0000 \end{bmatrix}, \quad c_{ik}^{(2)} = \begin{bmatrix} 0.1901 & 0.0198 \\ 0.0756 & 0.1177 \\ 0.1215 & 0.0005 \\ 0.0985 & 0.0020 \\ 0.0582 & 0.0659 \\ 0.2046 & 0.0437 \end{bmatrix}.$$

Route 2 (color blue)

$$\lambda_1 = 0.5770, \quad \lambda_2 = 0.4230$$

$$c_{ik}^{(1)} = \begin{bmatrix} 0.1305 & 0.0949 \\ 0.1221 & 0.0697 \\ 0.1176 & 0.0308 \\ 0.1469 & 0.0125 \\ 0.0677 & 0.1082 \\ 0.0969 & 0.0000 \end{bmatrix}, \quad c_{ik}^{(2)} = \begin{bmatrix} 0.1888 & 0.0264 \\ 0.1036 & 0.0950 \\ 0.1168 & 0.0028 \\ 0.0982 & 0.0055 \\ 0.0631 & 0.0660 \\ 0.1766 & 0.0551 \end{bmatrix}.$$

Route 3 (color red)

$$\lambda_1 = 0.3580, \quad \lambda_2 = 0.6420,$$

$$c_{ik}^{(1)} = \begin{bmatrix} 0.1220 & 0.0986 \\ 0.1138 & 0.0686 \\ 0.0993 & 0.0527 \\ 0.1138 & 0.0463 \\ 0.0451 & 0.1328 \\ 0.1006 & 0.0033 \end{bmatrix}, \quad c_{ik}^{(2)} = \begin{bmatrix} 0.1673 & 0.0579 \\ 0.1170 & 0.0819 \\ 0.0957 & 0.0237 \\ 0.0904 & 0.0268 \\ 0.0603 & 0.0730 \\ 0.1209 & 0.0826 \end{bmatrix}.$$

Route 4 (color green)

$$\lambda_1 = 0.1650, \quad \lambda_2 = 0.8350,$$

$$c_{ik}^{(1)} = \begin{bmatrix} 0.1167 & 0.0937 \\ 0.1074 & 0.0572 \\ 0.0852 & 0.0588 \\ 0.0814 & 0.0635 \\ 0.0057 & 0.1660 \\ 0.0389 & 0.1212 \end{bmatrix}, \quad c_{ik}^{(2)} = \begin{bmatrix} 0.1326 & 0.1028 \\ 0.1181 & 0.0603 \\ 0.0637 & 0.0605 \\ 0.0757 & 0.0473 \\ 0.0016 & 0.1480 \\ 0.0667 & 0.1187 \end{bmatrix}.$$

Route 5 (color magenta)

$$\lambda_1 = 0.0520 \quad \lambda_2 = 0.9480$$

$$c_{ik}^{(1)} = \begin{bmatrix} 0.1147 & 0.0895 \\ 0.1079 & 0.0454 \\ 0.0789 & 0.0615 \\ 0.0643 & 0.0755 \\ 0.0021 & 0.1677 \\ 0.0030 & 0.1847 \end{bmatrix}, \quad c_{ik}^{(2)} = \begin{bmatrix} 0.1279 & 0.1149 \\ 0.1351 & 0.0467 \\ 0.0385 & 0.0816 \\ 0.0691 & 0.0569 \\ 0.0000 & 0.1563 \\ 0.0308 & 0.1370 \end{bmatrix}.$$

In Fig. 2.5 we show the security routing bounds for decision making over the Pareto front. Bounds correspond to restrictions imposed by the decision maker over the Pareto front that establish a specific decision area where the strategies can be selected. By computing the minimum distance to the utopian point we have that Route 3 (color red) fulfill the requirements.

2.3 Portfolio Optimization Problem

2.3.1 Problem Formulation

The *portfolio optimization* issue compares the expected (mean) return of a portfolio with the standard deviation of the same portfolio using the mean-variance theory [12, 13, 21, 22, 25, 26, 32]. The expected return is plotted on the vertical axis in Fig. 2.6,

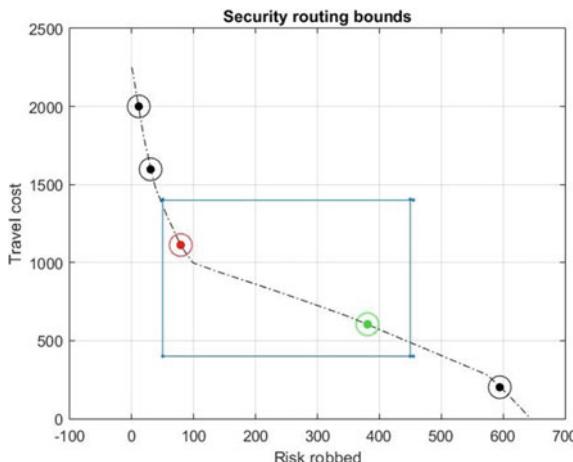


Fig. 2.5 Security routing bounds for decision making over the Pareto front

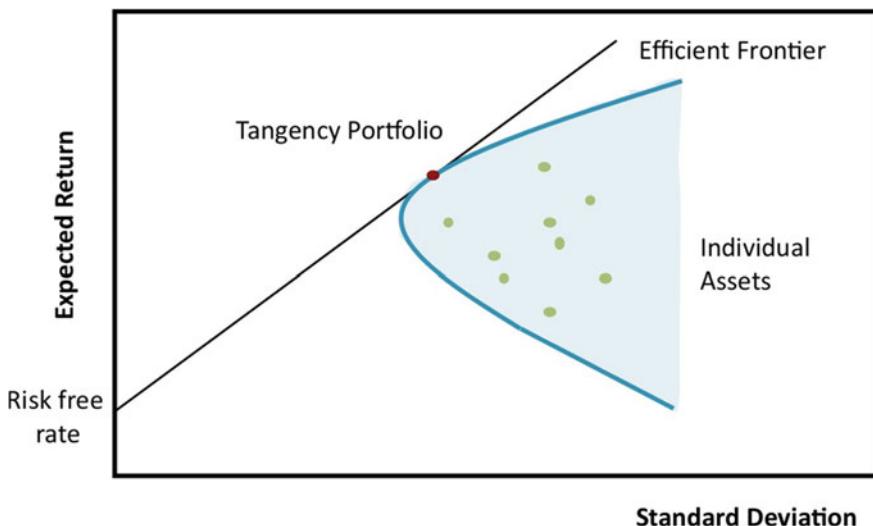


Fig. 2.6 Efficient Frontier. No risk-free asset is available

while the standard deviation is plotted on the horizontal axis (volatility). Standard deviation serves as a measure of risk and describes *volatility* [16, 18].

The space of “expected return vs risk” is another name for the return-standard deviation relationship. This risk-expected return space allows for the plotting of any conceivable combination of risky assets, and the area in this space defined by the collection of all such potential portfolios. In the absence of a risk-free asset, the higher portion of the region’s left parabolic border serves as the efficient frontier (sometimes called “the Markowitz bullet”). Combinations near this top edge reflect

portfolios with the lowest risk for a specific level of projected return (including no holdings of the risk-free asset). The combination that offers the best feasible expected return for a given risk level is equivalently represented by a portfolio that is located on the efficient frontier. The capital allocation line is tangent to the top portion of the parabolic boundary.

- The *expected return function* $U(c)$ is

$$\boxed{U(c) = \sum_{i=1}^N \sum_{k=1}^M \sum_{j=1}^N u_{ijk} \pi_{j|ik} d_{k|i} p_i(d) = \sum_{i=1}^N \sum_{k=1}^M w_{ik} c_{ik}, \\ c_{ik} = d_{k|i} p_i(d), \quad w_{ik} = \sum_{j=1}^N u_{ijk} \pi_{j|ik},} \quad (2.3.1)$$

where u_{ijk} are the elements of the utility tensor, which mean the utility of transfer from state (asset of portfolio) i to state j applying action (one of possible versions of scenario assets distributions) k under the applied strategy distribution c_{ik} . The variables p_i and $d_{k|i}$ are defined in Chap. 1.

- The *portfolio return variance* $\text{Var}(U(c))$ is:

$$\boxed{\text{Var}(U(c)) := \sum_{i=1}^N \sum_{k=1}^M [w_{ik} - U(c)]^2 c_{ik} = \sum_{i=1}^N \sum_{k=1}^M w_{ik}^2 c_{ik} - U^2(c).} \quad (2.3.2)$$

So, the portfolio optimization problem may be formulated as designing the strategies c_{ik} (defined in Chap. 1), which solve the following two-criteria optimization problem:

$$\boxed{U(c) \rightarrow \max_{c \in C_{adm}} \text{ or } [-U(c)] \rightarrow \min_{c \in C_{adm}}, \\ \text{Var}(U(c)) \rightarrow \min_{c \in C_{adm}}.} \quad (2.3.3)$$

2.3.2 Markowitz Function

There are known two possible approaches for the solution of the problem (2.3.3):

- transformation of the initial problem into a quadratic optimization with linear constraints,
- representation of the initial problem as a Pareto front optimization.

1. Let us introduce the additional requirement that the expected return would be no less than some required label U_0 , that is,

$$U(c) \rightarrow \max_{c \in C_{adm}} \geq U_0 \text{ or } [-U(c)] \rightarrow \min_{c \in C_{adm}} \leq [-U_0]. \quad (2.3.4)$$

Then, reformulate the problem (2.3.3) as

$$\boxed{\begin{aligned} \text{Var}(U(c)) &\rightarrow \min_{c \in C_{adm}}, \\ [-U(c)] &\rightarrow \min_{c \in C_{adm}} \leq [-U_0]. \end{aligned}} \quad (2.3.5)$$

Notice that the problem (2.3.5) is a quadratic optimization problem where $\text{Var}(U(c))$ is a quadratic function and $U(c)$ as well as $c \in C_{adm}$ are linear constraints, which can be resolved as in [17].

2. As it follows from Theorem 2.1, the set of all simultaneously not improvable policies (Pareto front points) can be represented as

$$\boxed{c^*(\lambda) = \arg \min_{c \in C_{adm}} [\lambda_1 \text{Var}(U(c)) - \lambda_2 U(c)] \in C_{adm}} \quad (2.3.6)$$

With fixed weights λ_1 and $\lambda_2 = 1 - \lambda_1$ the problem (2.3.6) is again a quadratic optimization problem subject to linear constraints.

To select desirable weights λ_1, λ_2 may be applied the following methods:

- Projection of the utopia point to Pareto front (2.1.7) and (2.1.8);
- Lexicographic Goal Programming method [33];
- Wierzbicki's achievement scalarization [34];
- Sen's Multi-Objective Programming [28];
- Hypervolume/Chebyshev Scalarization [36].

2.3.3 Numerical Example: A Rational Investor

Gaining a specific return is the investor's principal objective. A rational investor makes an effort to locate the portfolio with the lowest risk that meets this objective. We provide every potential portfolio for achieving this goal displaying the expected returns and risk (variance) of a portfolio's hazardous assets as a mean-variance diagram, where the points stand for the expected returns (\mathcal{U}) and risk (Var). Also, we refer to the collection of all points that have the shape of a hyperbola (Pareto typeface) as the efficient frontier. For more information, see Fig. 2.7.

If for a given volatility there is no portfolio with a higher return such that $\mathcal{U}(c^*) \leq \mathcal{U}(c)$ and $\text{Var}(c^*) \geq \text{Var}(c)$. In the mean-variance diagram from Fig. 2.7, it is the top boundary of every portfolio. According to our paradigm, the rational investor is specifically seeking the following set of portfolios: They both increase the expected return while minimizing the risk for a given return.

Considering

$$\pi_{j|i,1} = \begin{bmatrix} 0.3145 & 0.0936 & 0.0402 & 0.2803 & 0.2953 & 0.1823 \\ 0.3496 & 0.2510 & 0.4819 & 0.1980 & 0.1121 & 0.0547 \\ 0.0490 & 0.2545 & 0.0002 & 0.2320 & 0.3233 & 0.4762 \\ 0.0052 & 0.2906 & 0.0998 & 0.0126 & 0.1706 & 0.0794 \\ 0.2441 & 0.0530 & 0.2966 & 0.0174 & 0.0241 & 0.1670 \\ 0.0376 & 0.0573 & 0.0813 & 0.2597 & 0.0745 & 0.0404 \end{bmatrix},$$

$$\pi_{j|i,2} = \begin{bmatrix} 0.0698 & 0.0300 & 0.1055 & 0.2260 & 0.0665 & 0.0686 \\ 0.1494 & 0.0436 & 0.3801 & 0.1869 & 0.3306 & 0.0561 \\ 0.1612 & 0.3607 & 0.1894 & 0.1134 & 0.1374 & 0.2171 \\ 0.2543 & 0.2750 & 0.0401 & 0.0767 & 0.1414 & 0.3745 \\ 0.2645 & 0.0667 & 0.1788 & 0.3519 & 0.1178 & 0.0592 \\ 0.1008 & 0.2239 & 0.1061 & 0.0452 & 0.2064 & 0.2244 \end{bmatrix},$$

the optimal portfolio $\pi_{k|i}^*$ and the distribution vector P_i^* are given by

$$d_{k|i}^* = \begin{bmatrix} 0.5832 & 0.4168 \\ 0.4168 & 0.5832 \\ 0.5832 & 0.4168 \\ 0.6663 & 0.3337 \\ 0.5000 & 0.5000 \\ 0.5012 & 0.4988 \end{bmatrix}, \quad P_i^* = \begin{bmatrix} 0.1358 \\ 0.0984 \\ 0.1150 \\ 0.2565 \\ 0.0012 \\ 0.3932 \end{bmatrix}.$$

The mean-variance portfolio is the one on the efficient frontier with the lowest volatility (red circle in Fig. 2.7). The set of mean-variance efficient portfolios created by the risk-free and risky assets is the tangency point on the efficient frontier (blue circle in Fig. 2.7) if there is a risk-free asset (asset with zero volatility). Figure 2.8 plots the convergence of the functional.

2.4 Appendix

Proof (Theorem 2.2)

- (a) First, let us prove that the Hessian matrix H associated with the penalty function (2.2.5) is strictly positive definite for any positive μ and δ , i.e., we prove that for all $x \in \mathbb{R}^N$ and $u \in \mathbb{R}^{M_1}$

$$H = \begin{bmatrix} \frac{\partial^2}{\partial x^2} \mathcal{F}_{\mu,\delta}(x, u) & \frac{\partial^2}{\partial u \partial x} \mathcal{F}_{\mu,\delta}(x, u) \\ \frac{\partial^2}{\partial x \partial u} \mathcal{F}_{\mu,\delta}(x, u) & \frac{\partial^2}{\partial u^2} \mathcal{F}_{\mu,\delta}(x, u) \end{bmatrix} > 0. \quad (2.4.1)$$

To prove that, by the Schur lemma, it is necessary and sufficient to prove that

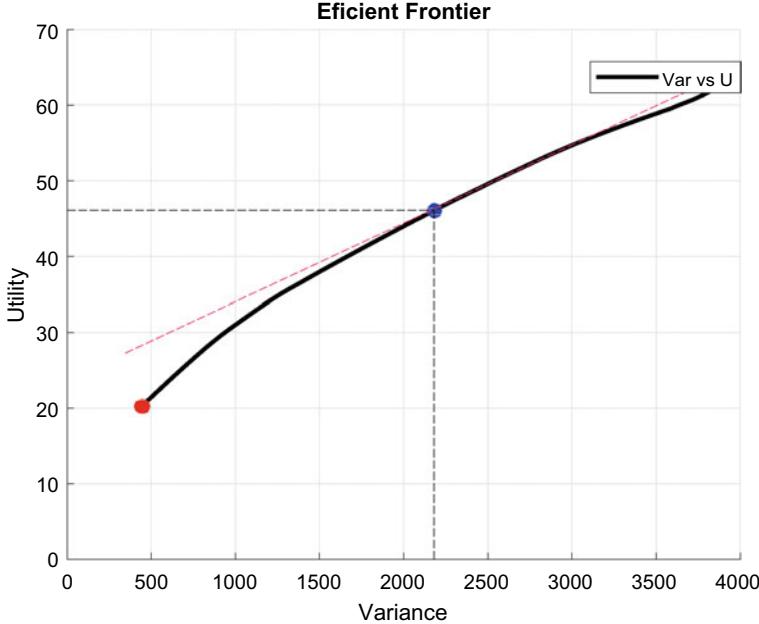


Fig. 2.7 Efficient Frontier

$$\begin{aligned} \frac{\partial^2}{\partial x^2} \mathcal{F}_{\mu, \delta}(x, u) &> 0, \quad \frac{\partial^2}{\partial u^2} \mathcal{F}_{\mu, \delta}(x, u) > 0 \\ \frac{\partial^2}{\partial x^2} \mathcal{F}_{\mu, \delta}(x, u) &> \frac{\partial^2}{\partial u \partial x} \mathcal{F}_{\mu, \delta}(x, u) \left[\frac{\partial^2}{\partial u^2} \mathcal{F}_{\mu, \delta}(x, u) \right]^{-1} \frac{\partial^2}{\partial x \partial u} \mathcal{F}_{\mu, \delta}(x, u) \end{aligned} \quad (2.4.2)$$

We have

$$\begin{aligned} \frac{\partial^2}{\partial x^2} \mathcal{F}_{\mu, \delta}(x, u) &= \mu \frac{\partial^2}{\partial x^2} f(x) + V_0^\top V_0 + V_1^\top V_1 + \delta I_{N \times N} \\ &\geq \mu \frac{\partial^2}{\partial x^2} f(x) + \delta I_{N \times N} \geq \delta \left(1 + \frac{\mu}{\delta} \lambda^- \right) I_{N \times N} > 0, \\ \forall \delta_n > 0 \end{aligned}$$

$$\lambda^- := \min_{x \in X_{adm}} \lambda_{\min} \left(\frac{\partial^2}{\partial x^2} f(x) \right),$$

$$\frac{\partial^2}{\partial u^2} \mathcal{F}_{\mu, \delta}(x, u) = I_{M_1 \times M_1} > 0.$$

By the Schur lemma (see for example [23])

$$\begin{aligned} \frac{\partial^2}{\partial x^2} \mathcal{F}_{\mu, \delta}(x, u) &= \mu \frac{\partial^2}{\partial x^2} f(x) + V_0^\top V_0 + V_1^\top V_1 + \delta I_{N \times N} > \\ \frac{\partial^2}{\partial u \partial x} \mathcal{F}_{\mu, \delta}(x, u) \left[\frac{\partial^2}{\partial u^2} \mathcal{F}_{\mu, \delta}(x, u) \right]^{-1} \frac{\partial^2}{\partial x \partial u} \mathcal{F}_{\mu, \delta}(x, u) &= (1 + \delta)^{-1} V_1^\top V_1 \end{aligned}$$

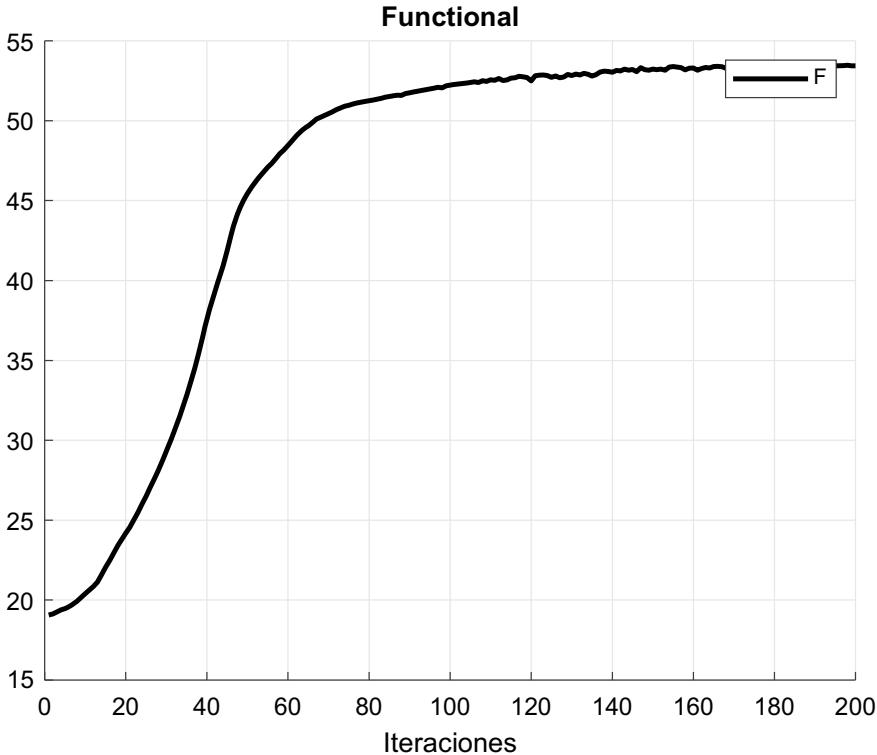


Fig. 2.8 Functional

implying $\mu \frac{\partial^2}{\partial x^2} f(x) + V_0^\top V_0 + \frac{\delta}{1+\delta} V_1^\top V_1 + \delta I_{N \times N} > 0$, which holds for any $\delta > 0$ by the condition (2.2.8) since

$$\begin{aligned} (\mu \lambda^- + \delta) I_{N \times N} + V_0^\top V_0 + \frac{\delta}{1+\delta} V_1^\top V_1 &\geq \\ \delta \left(1 + \frac{\mu}{\delta} \lambda^-\right) I_{N \times N} &= \delta (1 + o(1)) I_{N \times N} > 0. \end{aligned}$$

So, $H > 0$ which means that the penalty function (2.2.5) is strongly convex and, hence, has a unique minimal point defined below as $x^*(\mu, \delta)$ and $u^*(\mu, \delta)$.

- (b) By the strictly convexity property (2.4.1) for any $z := \begin{pmatrix} x \\ u \end{pmatrix}$ and any vector $z_n^* := \begin{pmatrix} x_n^* = x^*(\mu_n, \delta_n) \\ u_n^* = u^*(\mu_n, \delta_n) \end{pmatrix}$ for the function $\mathcal{F}_{\mu, \delta}(x, u) = \mathcal{F}_{\mu, \delta}(z)$ we have

$$\left. \begin{aligned} 0 &\geq (z_n^* - z)^\top \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_n^*) = \mu_n (x_n^* - x)^\top \frac{\partial}{\partial x} f(x_n^*) + \\ &[V_0(x_n^* - x)]^\top [V_0 x_n^* - b_0] + [V_1(x_n^* - x)]^\top [V_1 x_n^* - b_1 + u_n^*] + \\ &\delta_n (x_n^* - x)^\top x_n^* + (u_n^* - u)^\top [V_1 x_n^* - b_1 + (1 + \delta) u_n^*]. \end{aligned} \right\} \quad (2.4.3)$$

Selecting in (2.4.3) $x := x^* \in X^*$ (x^* is one of admissible solutions such that $V_0 x^* = b_0$) and $u := b_1 - V_1 x_n^*$ we obtain

$$\begin{aligned} 0 &\geq \mu_n (x_n^* - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) + \|V_0(x_n^* - x^*)\|^2 + \|V_1(x_n^* - x^*)\|^2 + \\ &\delta_n (x_n^* - x^*)^\top x_n^* + (1 + \delta)^{-1} \|V_1 x_n^* - b_1 + (1 + \delta) u_n^*\|^2 + \\ &\delta_n (u_n^* - b_1 - V_1 x_n^*)^\top u_n^*. \end{aligned}$$

Dividing both sides of this inequality by δ_n we get

$$\begin{aligned} 0 &\geq \frac{\mu_n}{\delta_n} (x_n^* - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) + \\ &\frac{1}{\delta_n} \left(\|V_0 x_n^* - b_0\|^2 + \|V_1(x_n^* - x^*)\|^2 \right. \\ &\quad \left. + \|V_1 x_n^* - b_1 + (1 + \delta) u_n^*\|^2 \right) + \\ &(x_n^* - x^*)^\top x_n^* + (u_n^* - b_1 - V_1 x_n^*)^\top u_n^*. \end{aligned} \quad (2.4.4)$$

Notice also that from (2.4.3), taking $x = x_n^*$ and $u = 0$, it follows

$$0 \geq \left[\left\| \sqrt{1 + \delta} u_n^* + \frac{(V_1 x_n^* - b_1)}{2\sqrt{1 + \delta}} \right\|^2 - \left\| \frac{(V_1 x_n^* - b_1)}{2\sqrt{1 + \delta}} \right\|^2 \right],$$

implying

$$1 \geq \left\| e + 2(1 + \delta) u_n^* \right\| \left(V_1 x_n^* - b_1 \right)^{-1} \left\| e \right\|^2, \|e\| = 1.$$

which means that the sequence $\{u_n^*\}$ is bounded. In view of this and taking into account that by the supposition (2.2.8) $\frac{\mu_n}{\delta_n} \xrightarrow{n \rightarrow \infty} 0$, from (2.4.4) it follows

$$\left. \begin{aligned} \text{Const} &= \limsup_{n \rightarrow \infty} (|(x_n^* - x^*)^\top x_n^*| + |(u_n^* - b_1 - V_1 x_n^*)^\top u_n^*|) \geq \\ &\limsup_{n \rightarrow \infty} \frac{1}{\delta_n} \left(\|V_0 x_n^* - b_0\|^2 + \right. \\ &\quad \left. \|V_1(x_n^* - x^*)\|^2 + (1 + \delta_n)^{-1} \|V_1 x_n^* - b_1 + (1 + \delta_n) u_n^*\|^2 \right). \end{aligned} \right\} \quad (2.4.5)$$

From (2.4.5) we may conclude that

$$\left. \begin{aligned} & \|V_0 x_n^* - b_0\|^2 + \|V_1 (x_n^* - x^*)\|^2 + \\ & (1 + \delta_n)^{-1} \|V_1 x_n^* - b_1 + (1 + \delta_n) u_n^*\|^2 = O(\delta_n) \end{aligned} \right\} \quad (2.4.6)$$

and

$$V_0 x_\infty^* - b_0 = 0,$$

$$V_1 x_\infty^* - V_1 x^* = V_1 x_\infty^* - b_1 + u_\infty^* = 0,$$

where $x_\infty^* \in X^*$ is a partial limit of the sequence $\{x_n^*\}$ which, obviously, may be not unique. The vector u_∞^* is also a partial limit of the sequence $\{u_n^*\}$.

- (c) Denote by \hat{x}_n the projection of x_n^* to the set X_{adm} , namely,

$$\hat{x}_n = \text{Pr}_{X_{adm}}(x_n^*), \quad (2.4.7)$$

and show that

$$\|x_n^* - \hat{x}_n\| \leq C\sqrt{\delta_n}, \quad C = \text{const} > 0. \quad (2.4.8)$$

Developing, we have

$$\|V_1 x_n^* - b_1 + u_n^*\| \leq C_1 \sqrt{\delta_n}, \quad C_1 = \text{const} > 0,$$

implying

$$V_1 x_n^* - b_1 \leq C_1 \sqrt{\delta_n} e - u_n^* \leq C_1 \sqrt{\delta_n} e, \quad \|e\| = 1.$$

where the vector inequality is treated in component-wise sense. Therefore

$$\|x_n^* - \hat{x}_n\|^2 \leq \max_{V_1 x - b_1 \leq C_1 \sqrt{\delta_n} e, x \in X_{adm}} \min_{y \in X_{adm}} \|x - y\|^2 := d(\delta_n).$$

Introduce the new variable

$$\tilde{x} := (1 - \nu_n) x + \nu_n \dot{x} \in X_{adm} \quad (2.4.9)$$

where by the Slater condition (2.2.7) $0 < \nu_n := \frac{C_1 \sqrt{\delta_n}}{C_1 \sqrt{\delta_n} + \min_{j=1, \dots, M_1} |(V_1 \dot{x} - b_1)_j|} < 1$. For

new variable $x = \frac{\tilde{x} - \nu_n \dot{x}}{1 - \nu_n}$ we have

$$\begin{aligned} V_1 \tilde{x} - b_1 &\leq \frac{C_1 \sqrt{\delta_n}}{C_1 \sqrt{\delta_n} + \min_{j=1, \dots, M_1} |(V_1 \dot{x} - b_1)_j|} \cdot \\ &\left(\min_{j=1, \dots, M_1} |(V_1 \dot{x} - b_1)_j| e + (V_1 \dot{x} - b_1) \right) \leq 0. \end{aligned}$$

and therefore

$$\begin{aligned} d(\delta_n) &\leq \max_{V_1\tilde{x}-b_1 \leq 0, \tilde{x} \in X_{adm}} \left\| \frac{\tilde{x} - \nu_n \dot{x}}{1 - \nu_n} - \tilde{x} \right\|^2 = \\ &= \frac{\nu_n^2}{(1 - \nu_n)^2} \max_{V_1\tilde{x}-b_1 \leq 0, \tilde{x} \in X_{adm}} \|\tilde{x} - \dot{x}\|^2 \leq C_2 \delta_n, \\ &0 < C_2 < \infty. \end{aligned}$$

In view of that $\|x_n^* - \hat{x}_n\| \leq \sqrt{d(\delta_n)} \leq \sqrt{C_2} \sqrt{\delta_n}$ which proves (2.4.8).

(d) The last step is to prove the inequality

$$0 \geq (x_\infty^* - x^*)^\top x_\infty^* \text{ for any } x_\infty^* \leq X^*. \quad (2.4.10)$$

From (2.4.4) we get

$$\left. \begin{aligned} 0 &\geq (x_n^* - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) + \\ &\frac{1}{\mu_n} \left(\|V_0 x_n^* - b_0\|^2 + \|V_1(x_n^* - x^*)\|^2 \right) + \frac{\delta_n}{\mu_n} (x_n^* - x^*)^\top x_n^* + \\ &\frac{1}{\mu_n} \|V_1 x_n^* - b_1 + u_n^*\|^2 \geq \\ &(x_n^* - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) + \frac{\delta_n}{\mu_n} (x_n^* - x^*)^\top x_n^*. \end{aligned} \right\} \quad (2.4.11)$$

By the strong convexity property we have

$(x - y)^\top \left(\frac{\partial}{\partial x} f(x) - \frac{\partial}{\partial x} f(y) \right) \geq 0$ for any $x, y \in \mathbb{R}^N$ which, in view of the property (2.4.8), implies

$$\begin{aligned} (x_n^* - \hat{x}_n)^\top \frac{\partial}{\partial x} f(x_n^*) &= O(\sqrt{\delta_n}), \\ (\hat{x}_n - x^*)^\top \frac{\partial}{\partial x} f(\hat{x}_n) &\geq (\hat{x}_n - x^*)^\top \frac{\partial}{\partial x} f(x^*) \geq 0, \end{aligned}$$

and

$$\begin{aligned} (x_n^* - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) &= (x_n^* - \hat{x}_n)^\top \frac{\partial}{\partial x} f(x_n^*) + \\ &(\hat{x}_n - x^*)^\top \frac{\partial}{\partial x} f(x_n^*) = O(\sqrt{\delta_n}) + (\hat{x}_n - x^*)^\top \left(\frac{\partial}{\partial x} f(x_n^*) - \frac{\partial}{\partial x} f(\hat{x}_n) \right) + \\ &(\hat{x}_n - x^*)^\top \frac{\partial}{\partial x} f(\hat{x}_n) \geq O(\sqrt{\delta_n}) - \|\hat{x}_n - x^*\| \left\| \frac{\partial}{\partial x} f(x_n^*) - \frac{\partial}{\partial x} f(\hat{x}_n) \right\|. \end{aligned}$$

Since any polynomial function is Lipschitz continuous on any bounded compact set, we can conclude that

$$\left\| \frac{\partial}{\partial x} f(x_n^*) - \frac{\partial}{\partial x} f(\hat{x}_n) \right\| \leq \text{Const} \|x_n^* - \hat{x}_n\| = O(\sqrt{\delta_n}),$$

which gives $(x_n^* - \hat{x}^*)^\top \frac{\partial}{\partial x} f(x_n^*) = O(\sqrt{\delta_n})$, which by (2.4.11) leads to

$$\left. \begin{aligned} 0 &\geq (x_n^* - \hat{x}_n)^\top \frac{\partial}{\partial x} f(x_n^*) + \frac{\delta_n}{\mu_n} (x_n^* - x^*)^\top x_n^* = \\ &O(\sqrt{\delta_n}) + \frac{\delta_n}{\mu_n} (x_n^* - x^*)^\top x_n^*. \end{aligned} \right\} \quad (2.4.12)$$

Dividing both side of the inequality (2.4.12) by $\frac{\mu_n}{\delta_n}$ and in view (2.4.8) we finally obtain

$$0 \geq O\left(\frac{\mu_n}{\sqrt{\delta_n}}\right) + (x_n^* - x^*)^\top x_n^* = o(1)\sqrt{\delta_n} + (x_n^* - x^*)^\top x_n^*. \quad (2.4.13)$$

This, by (2.2.8), for $n \rightarrow \infty$ leads to (2.4.10). Finally, for any $x^* \leq X^*$ it implies

$$0 \geq (x_\infty^* - x^*)^\top x_\infty^* = \|x_\infty^* - x^*\|^2 + (x_\infty^* - x^*)^\top x^* \geq (x_\infty^* - x^*)^\top x^*.$$

This inequality exactly represents the necessary and sufficient condition that the point x^* is the minimum point of the function $\|x_\infty^*\|^2$ on the set X^* . Obviously, this point is unique and has a minimal norm among all possible partial limits x_∞^* .

Theorem is proven. \square

Proof (Lemma 2.1) The necessary and sufficient conditions for the points $x_n^* = x^*(\mu_n, \delta_n)$, $u_n^* = u^*(\mu_n, \delta_n)$ to the extremal points of the function $\mathcal{F}_{\mu_n, \delta_n}(x, u)$ are as follows:

$$\left. \begin{aligned} 0 &= \frac{\partial}{\partial x} \mathcal{L}_{\mu_n, \delta_n}(x_n^*, u_n^* | \lambda_{n,x}^*, \lambda_{n,u}^*) = \\ \mu_n \frac{\partial}{\partial x} f(x_n^*) + V_0^\top (V_0 x_n^* - b_0) + V_1^\top (V_1 x_n^* - b_1 + u_n^*) + \delta_n x_n^* - \lambda_{n,x}^*, \\ 0 &= \frac{\partial}{\partial u} \mathcal{L}_{\mu_n, \delta_n}(x_n^*, u_n^* | \lambda_{n,x}^*, \lambda_{n,u}^*) = V_1 x_n^* - b_1 + (1 + \delta) u_n^* - \lambda_{n,u}^*, \\ \lambda_{n,x}^*(i) x_n^*(i) &= \lambda_{n,u}^*(j) u_n^*(j) = 0 \text{ for all } i, j - \\ &\text{complementary slackness condition,} \end{aligned} \right\} \quad (2.4.14)$$

where $\mathcal{L}_{\mu_n, \delta_n}(x, u | \lambda_{n,x}, \lambda_{n,u})$ is the Lagrange function for the problem (2.2.4), defined for $\lambda_{n,x}(i) \geq 0$, $\lambda_{n,u}(j) \geq 0$ as

$$\mathcal{L}_{\mu_n, \delta_n}(x, u | \lambda_{n,x}, \lambda_{n,u}) := \mathcal{F}_{\mu_n, \delta_n}(x, u) - \sum_{i=1}^N \lambda_{n,x}.$$

Multiplying the first equation in (2.4.14) by x_n^* and the second one by u_n^* , in view of the complementary slackness conditions we derive

$$\begin{aligned} 0 &= \mu_n (x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) + (V_0 x_n^*)^\top (V_0 x_n^* - b_0) + \\ &\quad (V_1 x_n^*)^\top (V_1 x_n^* - b_1 + u_n^*) + \delta_n \|x_n^*\|^2, \\ 0 &= (u_n^*)^\top (V_1 x_n^* - b_1 + (1 + \delta) u_n^*), \end{aligned}$$

implying

$$\left. \begin{aligned} 0 &= \mu_n (x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) + (V_0 x_n^*)^\top (V_0 x_n^* - b_0) + \\ &\quad \|V_1 x_n^* - b_1 + u_n^*\|^2 + b_1^\top (V_1 x_n^* - b_1 + u_n^*) + \delta_n (\|x_n^*\|^2 + \|u_n^*\|^2) \end{aligned} \right\} \quad (2.4.15)$$

By the construction of the regularized penalty function it follows that

$$V_1 x_n^* - b_1 + u_n^* = 0. \quad (2.4.16)$$

Indeed, if it is not the case, then u_n^* can not be the optimal point since the function $\mathcal{F}_{\mu,\delta}(x_n^*, u_n^*)$ is more than its value when

$$u_n^* = - (V_1 x_n^* - b_1).$$

In view of this, the identity (2.4.15) is equal to

$$0 = \mu_n (x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) + (V_0 x_n^*)^\top (V_0 x_n^* - b_0) + \delta_n (\|x_n^*\|^2 + \|u_n^*\|^2), \quad (2.4.17)$$

implying

$$\begin{aligned} 0 &= \mu_n (x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) + \\ &\quad \left\| (V_0^\top V_0 + \delta_n I_{N \times N})^{1/2} x_n^* - \left[(V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2} V_0^\top b_0 / 2 \right] \right\|^2 - \\ &\quad \left\| (V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2} V_0^\top b_0 / 2 \right\|^2 + \delta_n \|u_n^*\|^2 \end{aligned}$$

and

$$\begin{aligned} \mu_n \left[(x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) \right] + \left\| (V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2} V_0^\top b_0 / 2 \right\|^2 = \\ \left\| (V_0^\top V_0 + \delta_n I_{N \times N})^{1/2} x_n^* - \left[(V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2} V_0^\top b_0 / 2 \right] \right\|^2. \end{aligned}$$

The last identity can be represented as

$$\begin{aligned} x_n^* &= \left(V_0^\top V_0 + \delta_n I_{N \times N} \right)^{-1} V_0^\top b_0 / 2 + \\ \rho_n(x_n^*) &\left(V_0^\top V_0 + \delta_n I_{N \times N} \right)^{-1/2} e_n(x_n^*). \end{aligned} \quad (2.4.18)$$

where $\|e_n(x_n^*)\| = 1$ and

$$\rho_n(x_n^*) := \left(\mu_n \left[(x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) \right] + \left\| \left(V_0^\top V_0 + \delta_n I_{N \times N} \right)^{-1/2} V_0^\top b_0 / 2 \right\|^2 \right)^{1/2}.$$

Notice that in the last identity, by the boundedness property of X_{adm} , $\left| (x_n^*)^\top \frac{\partial}{\partial x} f(x_n^*) \right| \leq c = \text{const}$ and the matrix $(V_0^\top V_0 + \delta_n I_{N \times N})^{-1}$ have following structure

$$\frac{(V_0^\top V_0 + \delta_n I_{N \times N})^{-1}}{\det[V_0^\top V_0 + \delta_n I_{N \times N}]} = \left[\begin{pmatrix} \sum_{k=0}^{N-1} b_{ij|k} \delta_n^k \\ \sum_{k=0}^N a_{ij|k} \delta_n^k \end{pmatrix}_{i,j} \right]_{i,j=\overline{1,N}},$$

where $\text{adj}A = [A_{ji}]^\top$ is the matrix adjoined to A and A_{ji} is the cofactor to the element a_{ij} . Since this matrix is nonsingular, it follows that $a_{ij|k} \neq 0$. The matrix $(V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2}$ has the same structure, namely,

$$(V_0^\top V_0 + \delta_n I_{N \times N})^{-1/2} = \left[\begin{pmatrix} \sum_{k=0}^{N-1} \bar{b}_{ij|k} \delta_n^k \\ \sum_{k=0}^N \bar{a}_{ij|k} \delta_n^k \end{pmatrix}_{i,j} \right]_{i,j=\overline{1,N}}, \bar{a}_{ij|k} \neq 0.$$

In view of that the vector x_n^* (2.4.18) has the following structure

$$\begin{aligned} (x_n^*)_i &= \sum_{j=1}^N d_j \frac{\sum_{k=0}^{N-1} b_{ij|k} \delta_n^k}{\sum_{k=0}^N a_{ij|k} \delta_n^k} + \\ &\left(\sqrt{\mu_n c(x_n^*) + c_0} \right) \sum_{j=1}^N \bar{d}_j(x_n^*) \frac{\sum_{k=0}^{N-1} \bar{b}_{ij|k} \delta_n^k}{\sum_{k=0}^N \bar{a}_{ij|k} \delta_n^k} \end{aligned} \quad (2.4.19)$$

where $|c(x_n^*)| \leq c < \infty$ and $|\bar{d}_j(x_n^*)| \leq c_1 < \infty$. The structure (2.4.19) together with the equality (2.4.16) directly implies (2.2.12). \square

Proof (Theorem 2.3) In view of (2.2.13) it follows

$$\left. \begin{aligned} W_n &= \left\| \left[z_{n-1} - \gamma_n \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) \right]_+ - z_n^* \right\|^2 \leq \\ &\quad \left\| (z_{n-1} - z_{n-1}^*) - \gamma_n \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) + (z_{n-1}^* - z_n^*) \right\|^2 = \\ &\quad W_{n-1} + \gamma_n^2 \left\| \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) \right\|^2 + \left\| (z_{n-1}^* - z_n^*) \right\|^2 - \\ &\quad 2\gamma_n (z_{n-1} - z_{n-1}^*)^\top \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) + \\ &\quad 2(z_{n-1} - z_{n-1}^*)^\top (z_{n-1}^* - z_n^*) - 2\gamma_n (z_{n-1}^* - z_n^*)^\top \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}). \end{aligned} \right\} \quad (2.4.20)$$

By the inequalities (see the inequalities (21.17) and (21.36) in [23]) we can conclude that

$$\left\| \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) \right\|^2 \leq (1 + \vartheta_n) L_\nabla W_{n-1} + (1 + \vartheta_n^{-1}) d, \quad (2.4.21)$$

where

$$\left\| \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}^*) \right\|^2 \leq d. \quad (2.4.22)$$

We also have

$$\begin{aligned} (z_{n-1} - z_{n-1}^*)^\top \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) &\geq l_n W_{n-1}, \quad l_n = (\mu_n \lambda^- + \delta_n) \\ |(z_{n-1} - z_{n-1}^*)^\top (z_{n-1}^* - z_n^*)| &\leq \|z_{n-1}^* - z_n^*\|, \\ \sqrt{W_{n-1}} \left| (z_{n-1}^* - z_n^*)^\top \frac{\partial}{\partial z} \mathcal{F}_{\mu_n, \delta_n}(z_{n-1}) \right| &\stackrel{\vartheta > 0}{\leq} \\ \|z_{n-1}^* - z_n^*\| \left[(1 + \vartheta^{1/2}) \sqrt{L_\nabla} \sqrt{W_{n-1}} + (1 + \vartheta^{-1/2}) \sqrt{d} \right]. \end{aligned}$$

Then, in view of Lemma 2.1, from (2.4.20) for $m = n - 1$ we obtain

$$\begin{aligned} W_n &\leq W_{n-1} + \gamma_n^2 [(1 + \vartheta) L_\nabla W_{n-1} + (1 + \vartheta^{-1}) d] + \\ &\quad 2(C_1^2 |\mu_n - \mu_{n-1}|^2 + C_2^2 |\delta_n - \delta_{n-1}|^2) - 2\gamma_n (\mu_n \lambda^- + \delta_n) W_{n-1} + \\ &\quad 2(C_1 |\mu_n - \mu_{n-1}| + C_2 |\delta_n - \delta_{n-1}|) \sqrt{W_{n-1}} + \\ &\quad 2\gamma_n (C_1 |\mu_n - \mu_{n-1}| + C_2 |\delta_n - \delta_{n-1}|) \cdot \\ &\quad \left[(1 + \vartheta^{1/2}) \sqrt{L_\nabla} \sqrt{W_{n-1}} + (1 + \vartheta^{-1/2}) \sqrt{d} \right], \end{aligned}$$

or, equivalently,

$$W_n \leq W_{n-1} (1 - \alpha_{n-1}) + \bar{\delta}_{n-1} \sqrt{W_{n-1}} + \beta_{n-1}, \quad (2.4.23)$$

where

$$\begin{aligned} \alpha_{n-1} &= 2\gamma_n (\mu_n \lambda^- + \delta_n) - \gamma_n^2 (1 + \vartheta) L_\nabla = \\ &2\gamma_n (\mu_n \lambda^- + \delta_n) \left[1 - \frac{\gamma_n (1 + \vartheta) L_\nabla}{2(\mu_n \lambda^- + \delta_n)} \right] \geq \\ &\gamma_n \delta_n 2(1 + o(1)) \left[1 - \frac{\gamma_n (1 + \vartheta) L_\nabla}{2\delta_n(o(1) + 1)} \right] \geq C_\alpha \gamma_n \delta_n, \end{aligned}$$

$$\begin{aligned} \bar{\delta}_{n-1} &= 2(C_1 |\mu_n - \mu_{n-1}| + C_2 |\delta_n - \delta_{n-1}|) [1 + \gamma_n (1 + \vartheta^{1/2}) \sqrt{L_\nabla}] \quad (2.4.24) \\ &\leq C_\delta (|\mu_n - \mu_{n-1}| + |\delta_n - \delta_{n-1}|), \end{aligned}$$

$$\begin{aligned} \beta_{n-1} &= \gamma_n^2 (1 + \vartheta^{-1}) d + (C_1^2 |\mu_n - \mu_{n-1}|^2 + C_2^2 |\delta_n - \delta_{n-1}|^2) + \\ &2\gamma_n (C_1 |\mu_n - \mu_{n-1}| + C_2 |\delta_n - \delta_{n-1}|) (1 + \vartheta^{-1/2}) \sqrt{d} \leq \gamma_n^2 C_{\beta,1} + \\ &\gamma_n (|\mu_n - \mu_{n-1}| + |\delta_n - \delta_{n-1}|) C_{\beta,2} + \\ &(|\mu_n - \mu_{n-1}|^2 + |\delta_n - \delta_{n-1}|^2) C_{\beta,3}. \end{aligned}$$

Using the inequality

$$W_n^r \leq (1 - r) \theta_n^r + \frac{r}{\theta_n^{1-r}} W_n, \quad r \in (0, 1), \quad \theta_n > 0, \quad (2.4.25)$$

for $r = 1/2$ and $\sqrt{\theta_n} = \frac{\bar{\delta}_{n-1}}{2\alpha_{n-1}(1-\rho)}$, $\rho \in (0, 1)$, the inequality (2.4.23) can be reduced to the following one

$$\begin{aligned} W_n &\leq W_{n-1} \left(1 - \alpha_{n-1} \left[1 - \frac{\bar{\delta}_{n-1}}{2\alpha_{n-1}\sqrt{\theta_n}} \right] \right) \\ &+ [\beta_{n-1} + \frac{1}{2}\bar{\delta}_{n-1}\sqrt{\theta_n}] = W_{n-1} (1 - \alpha_{n-1}\rho) + \\ &\left[\beta_{n-1} + \frac{\bar{\delta}_{n-1}^2}{4(1-\rho)\alpha_{n-1}} \right]. \end{aligned} \quad (2.4.26)$$

By Theorem 16.14 in [23], $W_n \xrightarrow{n \rightarrow \infty} 0$ if

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \quad \frac{\beta_{n-1}}{\alpha_{n-1}} + \frac{\bar{\delta}_{n-1}^2}{\alpha_{n-1}^2} \xrightarrow{n \rightarrow \infty} 0,$$

which is equivalent to (2.2.15). Theorem is proven. \square

References

1. Beltrami, E., Katehakis, M., Durinovic, S.: Multiobjective markov decisions in urban modelling. *Math. Model.* **6**(4), 333–338 (1995)
2. Benson, H.P., et al.: Matthias ehrgott, multicriteria optimization. Springer (2005) ISBN 3-540-21398-8. 323 p. *Eur. J. Oper. Res.* **176**(3), 1961–1964 (2007)
3. Clempner, J.B.: Necessary and sufficient karush-kuhn-tucker conditions for multiobjective markov chains optimality. *Automatica* **71**, 135–142 (2016)
4. Clempner, J.B.: Computing multiobjective markov chains handled by the extraproximal method. *Ann. Oper. Res.* **271**, 469–486 (2018)
5. Clempner, J.B.: A team formation method based on a markov chains games approach. *Cybern. Syst.* **50**(5), 417–443 (2019)
6. Clempner, J.B., Poznyak, A.S.: Using the manhattan distance for computing the multiobjective markov chains problem. *Int. J. Comput. Math.* **95**(11), 2269–2286 (2017)
7. Clempner, J.B., Poznyak, A.S.: A tikhonov regularization parameter approach for solving lagrange constrained optimization problems. *Eng. Optim.* **50**(11), 1996–2012 (2018)
8. Clempner, J.B., Poznyak, A.S.: A tikhonov regularized penalty function approach for solving polylinear programming problems. *J. Comput. Appl. Math.* **328**, 267–286 (2018)
9. Das, I., Dennis, J.E.: Normal-boundary intersection: an alternate approach for generating pareto-optimal points in multicriteria optimization problems. *SIAM J. Optim.* **8**, 631–657 (1998)
10. Fliege, J., Heseler, A.: Constructing approximations to the efficient set of convex quadratic multi-objective problems. University of Dortmund, Germany, Technical report (2003)
11. Garcia, C.B., Zangwill, W.I.: Pathways to Solutions, Fixed Points and Equilibria. Prentice-Hall, Englewood Cliffs (1981)
12. Garcia-Galicia, M., Carsteau, A.A., Clempner, J.: Continuous-time learning method for customer portfolio with time penalization. *Expert Syst. Appl.* **129**, 27–36 (2019)
13. Garcia-Galicia, M., Carsteau, A.A., Clempner, J.: Continuous-time mean variance portfolio with transaction costs: a proximal approach involving time penalization. *Int. J. Gen Syst* **48**(2), 91–111 (2019)
14. Germeyer, Y.: Introduction to the Theory of Operations Research. Nauka, Moscow (1971)
15. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23**(2), 462–466 (1952)
16. Markowitz, H.: Portfolio selection. *J. Finance* **7**, 77–98 (1952)
17. Markowitz, H.: The optimization of a quadratic function subject to linear constraints. *Nav. Res. Logist. Q.* **3**, 111–133 (1956)
18. Markowitz, H.M.: Mean-variance analysis. In: *Finance*, pp. 194–198. Springer (1989)
19. Miettinen, K.: Nonlinear multiobjective optimization, vol. 12. Springer Science & Business Media (2012)
20. Novák, J.: Linear programming in tector criterion markov and semi-markov decision processes. *Optim.* **20**(5), 651–670 (1989)
21. Ortiz-Cerezo, L., Carsteau, A., Clempner, J.B.: Optimal constrained portfolio analysis for incomplete information and transaction costs. *Econ. Comput. Econ. Cybern. Stud. Res.* **4**(56), 107–121 (2022)
22. Ortiz-Cerezo, L., Carsteau, A., Clempner, J.B.: Sharpe-ratio portfolio in controllable markov chains: analytic and algorithmic approach for second order cone programming. *Mathematics* **10**(18), 3221 (2022)
23. Poznyak, A.S.: Advanced Mathematical Tools for Automatic Control Engineers. Deterministic Technique, vol. 1. Elsevier, Amsterdam, Oxford (2008)
24. Poznyak, A.S., Najim, K., Gómez-Ramírez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker, Inc. (2000)
25. Sánchez, E.M., Clempner, J.B., Poznyak, A.S.: A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean-variance customer portfolio: the case of bank marketing campaigns. *Eng. Appl. Artif. Intell.* **46**, Part A, 82–92 (2015)

26. Sánchez, E.M., Clempner, J.B., Poznyak, A.S.: Solving the mean-variance customer portfolio in markov chains using iterated quadratic/lagrange programming: a credit-card customer-credit limits approach. *Expert Syst. Appl.* **42**(12), 5315–5327 (2015)
27. Schittkowski, K.: Easy-opt: an interactive optimization system with automatic differentiation - user's guide. Department of Mathematics, University of Bayreuth, Technical report (1999)
28. Sen, C.: A new approach for multi-objective rural development planning. *Indian Econ. J.* **30**(4), 91–96 (1983)
29. Steuer, R.E.: The Tchebycheff procedure of interactive multiple objective programming. In: *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, pp. 235–249. Springer, Berlin (1989)
30. Tikhonov, A., Goncharsky, A., Stepanov, V., Yagola, A.G.: *Numerical Methods for the Solution of Ill-Posed Problems*. Kluwer Academic Publishers (1995)
31. Tikhonov, A.N., Arsenin, V.Y.: *Solution of Ill-posed Problems*. Winston & Sons, Washington (1977)
32. Vazquez, E., Clempner, J.B.: Customer portfolio model driven by continuous-time markov chains: an l2 lagrangian regularization method. *Econ. Comput. Econ. Cybern. Stud. Res.* **2**, 23–40 (2020)
33. Wang, Y.M.: On lexicographic goal programming method for generating weights from inconsistent interval comparison matrices. *Appl. Math. Comput.* **173**(2), 985–991 (2006)
34. Wierzbicki, A.P.: A mathematical basis for satisficing decision making. *Math. Model.* **3**(5), 391–405 (1982)
35. Zangwill, W.I.: *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs (1969)
36. Zhang, R., Golovin, D.: Random hypervolume scalarizations for provable multi-objective black box optimization. In: *International Conference on Machine Learning*, pp. 11096–11105. PMLR (2020)

Chapter 3

Partially Observable Markov Chains



Abstract The controlled Partly Observable Markov Decision Process (POMDP) architecture has shown to be effective in a variety of fields where one is required to disclose only partial knowledge about the problem's structure and parameters. Since some state variables are difficult to track and measure correctly, it may be more beneficial to base judgments on less accurate information. This chapter focuses on the design of an observer for a class of partially observable ergodic homogeneous finite Markov chains. The major objective of the suggested approach is to derive the formulas for computing an observer and, as a consequence, the best control strategy. We create a new variable that combines the policy, the observation kernel, and the distribution vector in order to solve the issue. To retrieve the important variables, we derive the formulas. The POMDP model's parameters are being learned in a dynamic context in this work. The development of the adaptive policies is based on an identification method, in which we count the number of unobserved events to estimate the components of the utility and transition matrices. The practical applications of the theoretical concerns addressed to a portfolio optimization problem are demonstrated through the use of a numerical example.

3.1 Introduction

3.1.1 Brief Review

The inaccessibility of the observer to the dynamics of the states s poses a challenge for engineering and economics applications. Some state variables might be hard or impossible to measure precisely at times. This occurs when the accessible devices only capture a portion of the object's states. Making judgments based on imperfect knowledge about the system state, which may be derived intuitively, may be more successful in certain sorts of challenges. This understanding relates to the observation kernel $P(s|\cdot)$ in the context of *partially observable Markov chains*. The information about the states s and actions a of the unobserved Markov process is supplied in the form of a functional given. If the choice of the acceptable policies' actions d does not depend on the current or past states, a Markov chain is said to be partly

observable. We analyze an observation y belongs to a given observation state space Y in the evolution of a partly observable dynamical system that is defined by the probability measures $P(s|\cdot)$ (observation kernel) (see [3, 12, 24, 26]). We know that the estimated y appears with probability $P(s|\cdot)$ if the state s is feasible.

Cassandra et al. [3] considered that the belief space is continuous in Partially Observable Markov Decision Process (POMDP), but the existing techniques for determining optimum policies in Markov decision processes (MDP's) that function exclusively in finite state spaces. Using the minimal principle, a required condition that an optimum control must satisfy, Lai and Elliott [12] studied the optimal control of a finite state, continuous time Markov chain seen in Gaussian noise. Whiting and Pickett [24] studied the estimate of a POMDP's model order, which is the total number of independent model parameters. Based on the problem's robust geometric features and the belief vectors, Zhang [26] offered a new perspective and methodology for understanding the POMDP problem.

Different approaches to handling the unknown model parameters have been suggested in the literature for the context of dynamical processes; see, for example, [2, 9, 13, 14, 22, 25]. All of these methods put an emphasis on learning the dynamics of transition and observation for each condition. Since it is difficult to recover the probability measures of the observation kernel $P(s|\cdot)$ with unknown parameters, a significant portion of the model will be very uncertain. The process of retrieving data frequently calls for a lot of time, physical resources, and processing power, or it relies on intuition, which might be inaccurate if the expert's assumptions and expectations don't fit the actual model well. In order to handle partially observed inventory issues, Bensoussan et al. [2] suggested an approach that includes linearizing state transitions using unnormalized probabilities. When transition, observation, and reward models are unknown, Doshi et al. [9] developed an estimate based on reducing the immediate Bayes risk for making decisions. When determining a suitable statistic that reduces the issue to one of perfect state knowledge, Lesser and Oishi [13] suggested a way to maintain the multiplicative cost structure. Littman et al. [14] examined a number of straightforward problem-solving techniques and shown that they can all produce nearly optimum solutions for a number of very tiny POMDP's drawn from the learning literature. Prediction profile models are non-generative partial models for partially observable systems that only provide a certain set of predictions and are, in some situations, far simpler than generative models, according to Talvitie and Singh [22]. For time-varying uncertain discrete-time, homogeneous, first-order, finite-state, finite-alphabet hidden Markov models, Xie et al. [25] explored a robust state estimation issue.

The control objective of the earlier studies is, in general, the same as in the case of the total observability method but involves an observation kernel and takes into account that the sets of the acceptable policies rely on estimated state y . The difficulty, therefore, is in creating an observer that can keep the best control features in a dynamic environment while still being computationally tractable.

In this chapter we follow [8]:

- a technique for creating a joint observer $o \in \mathcal{O}$ with incomplete information is addressed. Here, $o = P(s|y)P(s)$ is defined as the product of the distribution vector $P(s)$ and the observation kernel $P(s|y)$, which specifies the connection between the real and the estimated state, and \mathcal{O} is the set of joint observers.
- the joint observer $o(a|y)$ and the policy $d(a|y)$ are multiplied together to create the new variable $c = P(s|y)d(a|y)P(s)$, which is used to solve the problem. The optimal policy d^* , the estimated stochastic kernel $P(s|y)$, and the distribution vector $P(s)$, which are our variables of interest, may all be recovered after the optimal c^* has been calculated. The policy d^* , the observation kernel $P^*(s|y)$, the distribution vector P^* , and the joint observer o are computed using formulae that are derived as a consequence of the suggested technique.
- a dynamic environment is taken into account during learning the POMDP model's parameters. The development of *adaptive policies* is based on an identification strategy in which we calculate the utility matrices and transition matrices by computing the total number of unobserved events. The desired optimum adaptive strategy is then computed.

3.2 Partially Observable Markov Chains

Let us associate with the finite state space S the observation set Y , which takes values in a finite space $\{1, \dots, M\}$, $M \in \mathbb{N}$. The stochastic process $\{y(n), n \in \mathbb{N}\}$ is called the *observation process*. By observing $y(n)$ at time n information regarding the true value of $s(n)$ is obtained. If an observation $y(n) = y_m$ and $a(n) = a_k$, then the real $s(n) = s_i$ will have a probability

$$q_{i|mk} := P(s(n) = s_i | y(n) = y_m, a(n) = a_k), \quad (3.2.1)$$

where $m = \overline{1, M}$, $i = \overline{1, N}$, $k = \overline{1, K}$ that denotes the relationship between the state and the observation when an action $a_k \in A(s_i)$ is chosen at time n . The observation kernel is a stochastic kernel on Y given by $Q = [q_{i|mk}]_{m=\overline{1, M}, i=\overline{1, N}, k=\overline{1, K}}$. Formally, we have that if at time n an action $a(n) = a_k$ is chosen and the state $s(n) = s_i$ is possible then with probability $q_{i|ma}$ the estimated y_m appears.

Definition 3.1 A controllable POMDP is an 8-tuple

$$POMDP = \{S, A, \Pi, Y, Q, Q_0, P_0, u\}, \quad (3.2.2)$$

where

- S is a finite state space;
- A is a finite control space;

– $\Pi = [\pi_{j|ik}]$ is a stochastic kernel representing a stationary controlled transition matrix, where $\pi_{j|ik} \equiv P(s(n+1) = s_j | s(n) = s_i, a(n) = a_k)$ is the probability associated with the transition from state s_i to state s_j under an action $a_k \in A(s_i)$, $k = 1, \dots, K$, $K \in \mathbb{N}$;

– Y is the observation set, which takes values in a finite space $\{1, \dots, M\}$, $M \in \mathbb{N}$,

– $Q = [q_{i|m}]_{m=1,M,i=1,N}$ denotes the observation kernel is a stochastic kernel on Y such that

$$\sum_m q_{i|m} = 1 \text{ for all } i = \overline{1, N}, \quad (3.2.3)$$

which means that each state i is observable with probability one,

– $Q_0 = [q_{i|m}]_{m=1,M,i=1,N}$ denotes the initial observation kernel, which is a stochastic kernel on Y given S ,

– P_0 is the (a prior) initial distribution;

– $u_{ijmk} : S \times S \times Y \times A \rightarrow \mathbb{R}$ is the one-step reward function given the real unobservable state $s_i \in S$, the next state $s_j \in S$ and the estimated (measured) state $y_m \in Y$, when the action $a_k \in A(s_i, y_m)$ is taken.

A realization of the partially observable system at time n is given by

$$(s(0), y(0), a(0), s(1), y(1), a(1), \dots) \in H := (SYA)^\infty, \quad (3.2.4)$$

where $s(0)$ has a given distribution $P(s(0) = s_i)$ and $\{a_n\}$ is a control sequence in A determined by a control policy. To define a policy, we cannot use the states $s(0), s(1), \dots$. There, we introduce the observable histories $h_0 := (y(0)) \in H_0$ and $h_n := (s(0), y(0), a(0), \dots, y(n-1), a(n-1), y(n)) \in H_n$ for all $n \geq 1$ and $H_n := H_{n-1}AY$ if $n \geq 1$. Then a *policy* for the POMDP are defined as a sequence $\{d_{k|m}(n)\}$ such that, for each n , $d_{k|m}(n)$ is a stochastic kernel on A given H_n . The set of all feasible policies is denoted by D_{adm} . A policy $d_{k|m}(n) \in D$ and an initial distribution $P(s_0)$, denoted also as P_0 , together with the stochastic kernels Π , Q and Q_0 determine on the space H all possible realizations of the POMDP.

Definition 3.2 A sequence of random stochastic matrices $D(n) = \{d_{k|m}(t)\}_{t=0,n,k=1,K,m=1,M}$ is said to be a **randomized control strategy** such that for any random action $d_{k|m}(n)$ at time n

$$\sum_{k=1}^K d_{k|m}(n) = 1, \quad d_{k|m} \geq 0, \quad m = 1, \dots, M. \quad (3.2.5)$$

For any random action $\{d_{k|m}(n)\}_{k=1,K,m=1,M}$ the conditional transition probability matrix $\Pi(\{d_{k|m}(n)\}) = [\pi_{j|i}(\{d_{k|m}(n)\})]_{i,j=\overline{1, N}}$ can be defined as follows

$$\left. \begin{aligned} \pi_{j|i}(\{d_{k|m}(n)\}) &= \sum_{k=1}^K \sum_{m=1}^M P(s(n+1)=s_j | s(n)=s_i, a(n)=a_k) \cdot \\ P(s(n) = s_i | y(n) = y_m) d_{k|m}(n) &= \sum_{k=1}^K \sum_{m=1}^M \pi_{j|ik} q_{i|m} d_{k|m}(n), \end{aligned} \right\} \quad (3.2.6)$$

which represents the probability to move from the states s_j to the state s_i under the applied random action $\{d_{k|m}(n)\}_{k=\overline{1,K}, m=\overline{1,M}}$.

3.3 Formulation of the Problem

The dynamics of the POMC are described in terms of information that cannot be observed directly. At time $n = 0$, the initial (observed) state s_0 has a given a prior distribution P_0 , and the initial observation $y(0)$ is generated according to the initial observation kernel $Q_0(y(0)|s(0))$. If at time n the state of the system is $s(n)$ and the control $a(n) \in A$ is applied, then each of strategy is allowed to randomize, with distribution $d_{k|m}(n)$, over the pure action choices $a(n) \in A(s(n))$, $m = \overline{1, M}$ and $k = \overline{1, K}$. These choices induce immediately utility u_{ijmk} . The system tries to minimize the corresponding one-step reward. Next, the system moves to the new state $s(n) = s_i$ according to the transition probabilities $\Pi(\{d_{k|m}(n)\}_{k=\overline{1,K}, m=\overline{1,M}})$. Then, the observation $y(n)$ is generated by the observation kernel $Q(y(n)|s(n))$. Based on the obtained reward, the systems adapt a mixed strategy computing $d_{k|m}(n+1)$ for the next selection of the control actions.

Within the class of all stationary strategies and considering ergodic Markov chains [6], for any fixed collection of stationary strategies $d_{k|m}(n) = d_{k|m}$ we have

$$P(s(n+1) = s_j) \rightarrow P_j, n \rightarrow \infty. \quad (3.3.1)$$

That is, in the case when the Markov chain is ergodic for any stationary strategy $d_{k|m}$ the distributions P_j exponentially fast converge to their limits $P(s_i) = P_i$ satisfying

$$P_j = \sum_{i=1}^N \left(\sum_{m=1}^M \sum_{k=1}^K \pi_{j|ik} q_{i|m} d_{k|m} \right) P_i. \quad (3.3.2)$$

Denote also by \bar{P}_m the probability to observe the estimated state m , which can be calculated as

$$\bar{P}_m = \sum_{i=1}^N \left(\sum_{k=1}^K q_{i|m} d_{k|m} \right) P_i. \quad (3.3.3)$$

The reward function is given by the values W_{imk} , so that the “average reward function” U is given by

$$U(d, o) := \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K W_{imk} d_{k|m} o_{i|m}, \quad (3.3.4)$$

where

$$W_{imk} = \sum_j u_{ijk} \pi_{j|ik}, \quad (3.3.5)$$

and the *joint observer* [5, 8] is defined as $o_{i|m} = q_{i|m} P_i$ such that

$$\sum_{m=1}^M \sum_{i=1}^N o_{i|m} = 1. \quad (3.3.6)$$

Let us denote by \mathcal{O} the set of “*feasible joint observers*”. A feasible joint observer $o^* = \{o_{k|m}^*\}$ will be referred to as *optimal*, if it realizes the optimization rule

$$(d^*, o^*) \in \operatorname{Arg} \max_{d \in D_{adm}, o \in \mathcal{O}_{adm}} U(d, o). \quad (3.3.7)$$

Without losing generality, we have assumed that the reward is a function of the action, the estimated state, and the current state. Then, taking into account the current state and action, we calculate the conditional mean of this reward with respect to the estimated state y_m (the conditional distribution of the estimated state is provided by the stochastic kernel $q_{i|m}$). The original reward may be replaced with its conditional mean in the manner previously stated without losing generality since the overall reward function includes the expectation. In fact, the objective function value (including the estimated state y_m) has the same value as the objective function involving observable states.

3.4 Description in the c-Variables

Solutions to Eq. (3.3.7) can be found using dynamic programming tools. To see this, let us introduce [6, 18, 21] the variable $c = [c_{imk}]$ as follows

$$c_{imk} = d_{k|m} q_{i|m} P_i. \quad (3.4.1)$$

In terms of the *c*-variable the optimization problem given in Eq. (3.3.7) is satisfied by the following theorem.

Theorem 3.1 *The joint observer $o \in \mathcal{O}$ and the strategy d are optimal if and only if the variable c_{imk} represents the solution of the following linear programming problem*

$$\tilde{U}(c) = \sum_{i=1}^N \sum_{m=1}^M \sum_{k=1}^K W_{imk} c_{imk} \rightarrow \max_{c \in C_{adm}}, \quad (3.4.2)$$

subject to constraints

$$c_{imk} \geq 0, \sum_{i=1}^N \sum_{m=1}^M \sum_{k=1}^K c_{imk} = 1, \sum_{i=1}^N \sum_{k=1}^K c_{ilk} = P_i > 0,$$

$$\sum_{m=1}^M \sum_{k=1}^K [\delta_{ij} - \pi_{j|ik}] c_{imk} = 0, j = 1, \dots, N,$$

$$\sum_{k=1}^K c_{imk} = q_{i|m} \sum_{l=1}^M \sum_{k=1}^K c_{ilk},$$

or

$$\sum_{l=1}^M \sum_{k=1}^K [\delta_{lm} - q_{i|m}] c_{ilk} = 0, m = 1, \dots, M.$$

Proof We have that

$$\sum_{k=1}^K d_{k|m} = 1, \sum_{m=1}^M q_{i|m} = 1, \sum_{i=1}^N P_i = 1,$$

$$P_j = \sum_{imk} \pi_{j|ik} q_{i|m} d_{k|m} P_i,$$

and for $c_{imk} = d_{k|m} o_{i|m}$ it follows

$$\sum_{k=1}^K c_{imk} = q_{i|m} P_i, \sum_{k=1}^K c_{imk} = q_{i|m} \sum_{l=1}^M \sum_{k=1}^K c_{ilk},$$

and

$$\sum_{l=1}^M \sum_{k=1}^K c_{ilk} = P_i > 0, \sum_{i=1}^N P_i = 1,$$

then,

$$\sum_{i=1}^N \sum_{l=1}^M \sum_{k=1}^K c_{ilk} = 1.$$

It follows also that

$$U(d) := \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K (u_{ijmk} \pi_{j|ik}) d_{k|m} o_{i|m} = \\ \sum_{m=1}^M \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^K W_{imk} d_{k|m} q_{i|m} P_i = \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K W_{imk} c_{imk} = \tilde{U}(c_{imk}).$$

Notice that Eq. (3.3.2) is equivalent to

$$\sum_{m=1}^M \sum_{k=1}^K c_{jmk} = \sum_{i=1}^N \pi_{j|ik} c_{imk}, \\ \text{or} \\ \sum_{m=1}^M \sum_{k=1}^K [\delta_{ij} - \pi_{j|ik}] c_{imk} = 0.$$

We define the solution of the problem (3.4.2) as c^* . Now, for obtaining q^* we have to calculate

$$d_{kim}^* = \begin{cases} \frac{c_{imk}^*}{\sum\limits_{h=1}^K c_{imh}^*} & \text{if } \sum\limits_{h=1}^K c_{imh}^* > 0, \\ 0 & \text{if } \sum\limits_{h=1}^K c_{imh}^* = 0, \end{cases} \quad (3.4.3)$$

$$P_i^* = \sum_{l=1}^M \sum_{k=1}^K c_{ilk}^*, \quad (3.4.4)$$

and

$$q_{i|m}^* = \frac{\sum\limits_{k=1}^K c_{imk}^*}{\sum\limits_{l=1}^M \sum\limits_{k=1}^K c_{ilk}^*}. \quad (3.4.5)$$

So, finally to recover the optimal joint observer we need to use the formula

$$o_{i|m}^* = q_{i|m}^* P_i^*. \quad (3.4.6)$$

Remark 3.1 The equation $\sum_{i=1}^N \sum_{k=1}^K c_{ilk} = P_i > 0$ holds because of the ergodicity property.

The expression (3.4.5) is not trivial. The resulting stochastic matrix $[q_{i|m}^*]$ represents the imprecise information concerning the system state, which is usually obtained intuitively.

Corollary 3.1 *The strategy $d_{k|m}^*$, constructed from d_{kim}^* (3.4.3), is given by*

$$d_{k|m}^* = \frac{1}{N} \sum_{i=1}^N d_{kim}^*.$$

Corollary 3.2 *The obtained distribution of unobservable states is given by*

$$\bar{P}_m^* = \sum_{i=1}^N \sum_{k=1}^K c_{imk}^*.$$

We have hereby derived the formulas of the optimal rules to recover the joint observer o^* , the optimal policy $d_{k|m}^*$, the observation kernel q^* and the distribution vector \bar{P}_m^* , which maximize Eq. (3.4.2) based on the optimal variables c_{imk}^* .

3.5 Computation of the Estimated Value by Measurable Realization: Projection Stochastic Approximation Procedure

The solution of the adaptive problem consists of a maximization optimization problem

$$U(c) = \sum_{i=1}^N \sum_{m=1}^M \sum_{k=1}^K W_{imk} c_{imk} \rightarrow \max_{c \in C_{adm}}$$

given in Theorem 3.1. For computing the estimated value of the strategy c_{imk} we employ the projected gradient method given by

$$c(t+1) = \Psi \left\{ c(t) + \gamma_t \frac{\partial U_t(c)}{\partial c} \right\} \quad (3.5.1)$$

where the step parameter $\gamma_t > 0$ is a numerical sequence, Ψ is the projection operator to the simplex S_ε^{NMK} and $c_{imk}(t) \geq \epsilon_{t_n}$ such that

$$\begin{aligned} S_\varepsilon^{NMK} := & \left\{ c \in \mathbb{R}^{NMK} : c_{imk} \geq \varepsilon \text{ where,} \right. \\ & \left. \sum_{m=1}^M \sum_{k=1}^K c_{jmk} = \sum_{i=1}^N \pi_{j|ik}(t) c_{imk}, \sum_{i=1}^N \sum_{m=1}^M \sum_{k=1}^K c_{imk} = 1 \right\}. \end{aligned}$$

The integers t_n , $n \geq 0$ characterizing the given simplex are the moments that the values $c_{imk}(t)$ are calculated by using information obtained on the time interval $[t_{n-1}, t_n]$. These values characterize the laws forming the optimal adaptive policy. The projection Ψ is defined by the condition given by

$$\|\Psi(z) - z\| = \min_{x \in S^{NMK}} \|x - z\|$$

for any vector z . The elements of the gradient are given by

$$\frac{\partial U_t(c_t)}{\partial c_{imk}} = E \left\{ \frac{\xi_U(t) \chi(y(t)=y_m, s(t)=s_i, a(t)=a_k)}{d_{k|m}(t) q_{i|m}(t) P_i(t)} \right\} = E \left\{ \frac{\xi_U(t) \chi(y(t)=y_m, s(t)=s_i, a(t)=a_k)}{c_{imk}(t)} \right\}$$

where $\xi_U(t)$ is the realization of the utility function (3.3.4) at time t .

The stochastic approximation of $\frac{\partial U_t(c)}{\partial c_{imk}}$ [18] is given by matrix

$$V(t_{n+1}) = \frac{1}{t_{n+1} - t_n} \sum_{t=t_n}^{t_{n+1}-1} \frac{\xi_U e^\top(t)}{e^\top(t) c(t)},$$

where $e^\top(t) = (\underbrace{1, \dots, 1}_{N MK \text{ elements}})$ and $c(t) = \text{col}(c_{imk}(t))$. The elements of $V_{t_{n+1}}$ are computed according to the recurrence

$$V_{mik}(t_{n+1}) = V_{mik}(t_n) - \frac{1}{t_{n+1} - t_n} \left(V_{mik}(t_n) - \frac{\xi_U(t_n) \chi(y(t)=y_m, s(t)=s_i, a(t)=a_k)}{c_{mik}(t_n)} \right),$$

where $\xi_U := u_{ijmk} + \mu_U \text{rand}[-1, 1]$, $\mu_U \leq u_{\hat{i}\hat{j}\hat{m}\hat{k}}$. If the parameters $\{\gamma_t\}, \{t_n\}$ and $\{\epsilon_t\}$ satisfies the following constraints $\gamma_t > 0$, $t_{n+1}t_n^{-1} \rightarrow 0$, $\Delta t_n = t_{n+1} - t_n \rightarrow \infty$, when $t \rightarrow \infty$ and $n \rightarrow \infty$, and

$$\begin{aligned} \frac{t_{n+1} - t_n}{\gamma_{t+1}} &\geq \frac{t_n - t_{n-1}}{\gamma_t}, \\ \sum_{n=1}^{\infty} \gamma_t \Delta t_n \epsilon_n^{-1} t_n^{-1} + \left(\frac{\Delta t_n}{t_n} \right)^2 &< 0, 0 < \epsilon_n \rightarrow 0, \\ \lim_{t \rightarrow \infty} \frac{1}{\gamma_t} \left[\Delta t_n t_n^{-1} + |\epsilon_{t+1} - \epsilon_t| + \Delta t_n \left(\sum_{l=1}^t \epsilon_l \Delta t_l \right)^{-1} \right] &= 0, \\ 0 < \epsilon_t &\rightarrow 0, \end{aligned} \tag{3.5.2}$$

then, the sequence $\{\tilde{c}(t)\}_{t=0,1,2,\dots}$, generated by

$$\tilde{c}(t+1) = \Psi \{ \tilde{c}(t) + \gamma_t V_{mik}(t+1) \} \tag{3.5.3}$$

converges in mean-square to c^* with minimum norm [17],

$$c_{imk}^*(t) := \min_{c_{imk}(t) \in C_{adm}} \sum_{n=1}^t \sum_{i=1}^N \sum_{m=1}^M \sum_{k=1}^K \|c_{imk}^*(n)\|^2,$$

(since solution of problem (3.4.2) may be not unique). Within the class of numerical sequences

$$\gamma_t = \gamma t^{-\alpha}, \quad \epsilon_t = \epsilon t^{-\beta}, \quad t_n = n^\tau$$

the conditions of the projection-gradient algorithm given in Eq.(3.5.1) are satisfied if

$$0 < \beta < \alpha < 1 - \alpha, \quad \tau > 1.$$

The maximal rate τ of convergence $E \|c_{imk}(n) - c_{imk}^*\|^2 = O(\frac{1}{n^\tau})$ is attained for $\alpha = 1/2, \beta = 1/4, \tau = 5/4$.

3.5.1 Adaptive Algorithm

In this section, we will develop an algorithm for computing the optimal adaptive rules.

The system chooses randomly an estimated state $y(t) = y_m$ from $q_{i|m}(t)$ and also selects randomly an action $a(t) = a_{\hat{k}}$ (given \hat{k}) from $\pi_{k|m}(t)$ (for a fixed \hat{m}). Next, it employs the transition matrix $P = [p_{j|\hat{i}\hat{k}}]$ to choose randomly the consecutive state $s(t+1) = s_{\hat{j}}$ from $\hat{p}_{j|\hat{i}\hat{k}}(t-1)$ (given \hat{j} for a fixed \hat{i} and \hat{k}), and choose randomly a state $y(t+1) = y_u$ from $Q_{u|j}$. Then, the environment moves to a new state $s_{\hat{j}}$ given $a_{\hat{k}}$ and $s_{\hat{i}}$. The estimated values are updated employing the learning rules for computing $\hat{p}_{j|\hat{i}\hat{k}}(t)$ and $\hat{u}_{\hat{i}\hat{m}\hat{j}\hat{k}}(t)$. The value-maximizing action at each state are taken. If the condition of estimated error e is not satisfied, then the selection of the random variables s_i, s_j and a_k is carried out again. On the other hand, the policy $\pi_{k|m}$ and the observation kernel $q_{i|m}$ are computed again applying Theorem 3.1 and the projected gradient method given in Eq.(3.5.1). The process is described in the Algorithm presented in Table 3.1.

3.6 Numerical Example: A Partially Observable Markowitz Portfolio

This example presents a Markowitz portfolio optimization problem [10, 11, 15, 16, 19, 20, 23]. Stockholders can trade in securities and observe their prices delimited to one period. We suppose that the market does not allow short selling, for all assets the selling and buying prices are the same, there are no transaction costs, and all assets are infinitely divisible. The market is arbitrage-free.

The mean-variance customer model of Markowitz [1, 4, 7] is defined as follows: the reward R given by

Table 3.1 Adaptive Algorithm**Algorithm**

-
- (1) Let $t = 0$ and so $s(0) = s_i$ be the initial state.
 - (2) Let $\pi_{k|m}(t)$ be a policy and $q_{i|m}(t)$ the observation kernel computed applying Theorem 3.1 and the projected gradient method given in Eq. (3.5.1).
 - (3) Choose randomly with probability $q_{i|m}(t)$ an estimated state y_m and randomly with probability $\pi_{k|m}(t)$ an action a_k .
 - (4) Choose randomly with probability $\hat{p}_{j|i\hat{k}}(t)$ the next state s_j and choose randomly with probability $q_{r|j}(t)$ a state y_r .
 - (5) Update the values of $\chi_{mk}(t)$ and $\chi_{muk}(t)$.
 - (6) Compute $\hat{p}_{j|i\hat{k}}(t)$ according to the update rule and $\text{Pr}: \hat{p}_{(\cdot|i\hat{k})}(t) \rightarrow \mathcal{S}^N$.
 - (7) Compute $\hat{u}_{i\hat{j}m\hat{k}}(t)$ according to the update rule.
 - (8) Compute the mean square error $e(t)$.
 - (9) If $(\|\hat{P}(t-1) - P^*\| \leq \|\hat{P}(t) - P^*\|)$ continue else if not goto step 3.
 - (10) Update the estimated values computed by the learning rules $\hat{p}_{j|i\hat{k}}(t)$ and $\hat{u}_{i\hat{j}m\hat{k}}(t)$.
 - (11) Set $s_i = s_j$ and $t \leftarrow t + 1$ and go to step 2.
-

$$\begin{aligned}\mathcal{R}(c) &= \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K \sum_{j=1}^N u_{ijmk} p_{j|ik} \pi_{k|m} q_{i|m} P(s_i) \\ &= \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K W_{imk} c_{imk} \rightarrow \max_{c \in C_{adm}},\end{aligned}$$

where $c_{imk} = d_{k|m} P_i$ and the *variance* $V(c)$ given by

$$\begin{aligned}\mathcal{V}(c) &:= \sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K [W_{imk} - \mathcal{R}(c)]^2 c_{imk} = \\ &\sum_{m=1}^M \sum_{i=1}^N \sum_{k=1}^K W_{imk}^2 c_{imk} - R^2(c) \rightarrow \min_{c \in C_{adm}}.\end{aligned}$$

The resulting Markowitz portfolio problem is

$$U(c) := \mathcal{R}(c) - \frac{\zeta}{2} \mathcal{V}(c) \rightarrow \max_{c \in C_{adm}}$$

where ζ is the *risk-aversion* parameter.

The main goal of the Stockholder is to gain a given return. A rational Stockholder makes an effort to identify the portfolio with minimal risk, which satisfies this goal. We apply a reinforcement learning algorithm for POMDP with a discrete action space to asset allocation and computing the adaptive policies of the portfolio problem. The construction of adaptive policies is based on an identification approach where we

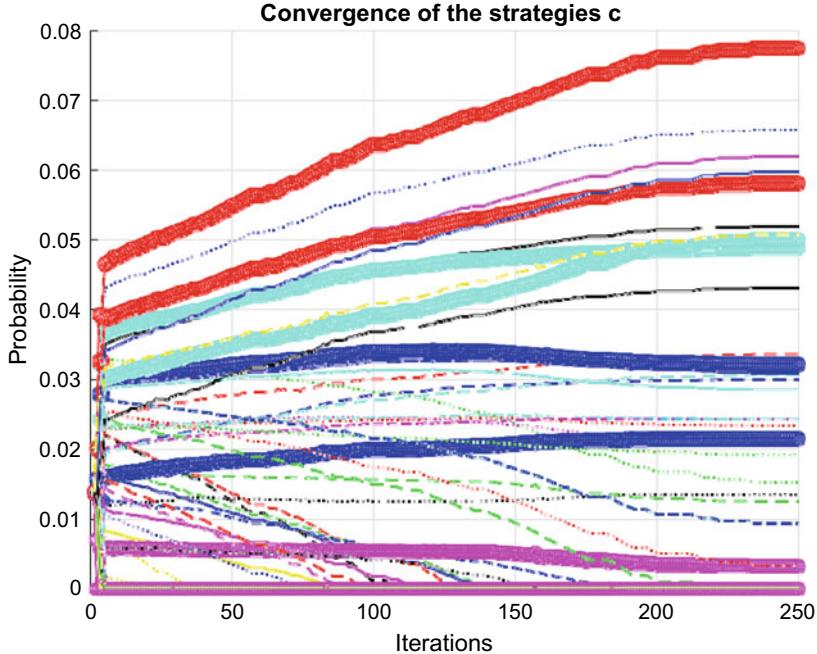


Fig. 3.1 Convergence of strategies c_{imk}

estimate the transitions matrices and the utility matrices by counting the number of unobserved experiences.

Following this purpose, we fix the values of interest $\zeta = 8.5 \times 10^{-2}$, $\gamma_0 = 2 \times 10^{-4}$, $N = M = 6$ and $K = 2$. To fulfill Stockholder's expectations, we outline the portfolio of risky assets in a diagram of the functional in Fig. 3.2. We show that the functional $U(c)$ converges. Convergence of the functional is a consequence of the convergence of the strategies in Fig. 3.1 that represents the portfolio. The convergence of the error of $\hat{p}_{j|i\hat{k}} (1 \times 10^{-6})$ is shown in Fig. 3.3 and the error of $\hat{u}_{i\hat{j}\hat{m}\hat{k}} (1 \times 10^{-1})$ is shown in Fig. 3.4. The portfolio is the result of fixing the volatility. This is exactly the portfolio, that the rational stockholder in our framework is looking for: it maximizes the expected return for a given risk. Applying the equations developed in Section 4, we obtain that the resulting values of interest are the following:

$$\pi_{k|m}^* = \begin{bmatrix} 0.5277 & 0.4723 \\ 0.8307 & 0.1693 \\ 0.5167 & 0.4833 \\ 0.8312 & 0.1688 \\ 0.4120 & 0.5880 \\ 0.2732 & 0.7268 \end{bmatrix}, \quad P_i^* = \begin{bmatrix} 0.2305 \\ 0.2408 \\ 0.1095 \\ 0.1800 \\ 0.1239 \\ 0.1153 \end{bmatrix},$$

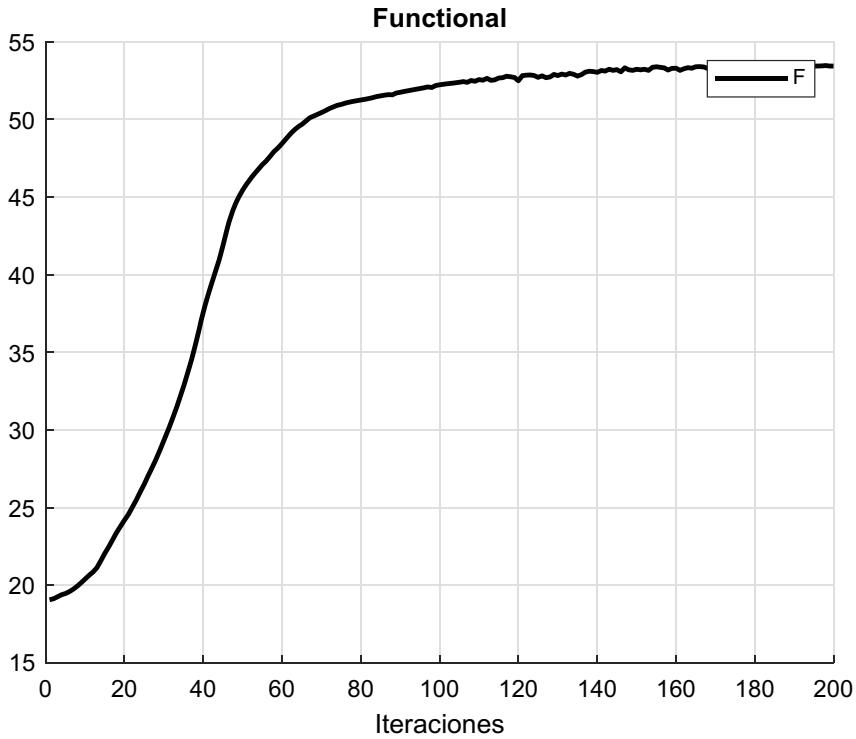


Fig. 3.2 Convergence of the functional $U(c)$

$$q_{i|m}^* = \begin{bmatrix} 0.0021 & 0.1796 & 0.1724 & 0.1670 & 0.2392 & 0.1345 \\ 0.1942 & 0.1191 & 0.1376 & 0.0543 & 0.1744 & 0.1558 \\ 0.0551 & 0.2219 & 0.2436 & 0.1757 & 0.1320 & 0.2187 \\ 0.2954 & 0.2178 & 0.0667 & 0.2232 & 0.1261 & 0.1837 \\ 0.2476 & 0.0712 & 0.2169 & 0.2214 & 0.1468 & 0.0402 \\ 0.2057 & 0.1903 & 0.1628 & 0.1585 & 0.1816 & 0.2672 \end{bmatrix},$$

$$d_{i|m}^* = \begin{bmatrix} 0.0005 & 0.0432 & 0.0189 & 0.0301 & 0.0296 & 0.0155 \\ 0.0448 & 0.0287 & 0.0151 & 0.0098 & 0.0216 & 0.0180 \\ 0.0127 & 0.0534 & 0.0267 & 0.0316 & 0.0164 & 0.0252 \\ 0.0681 & 0.0524 & 0.0073 & 0.0402 & 0.0156 & 0.0212 \\ 0.0571 & 0.0171 & 0.0238 & 0.0399 & 0.0182 & 0.0046 \\ 0.0474 & 0.0458 & 0.0178 & 0.0285 & 0.0225 & 0.0308 \end{bmatrix}.$$

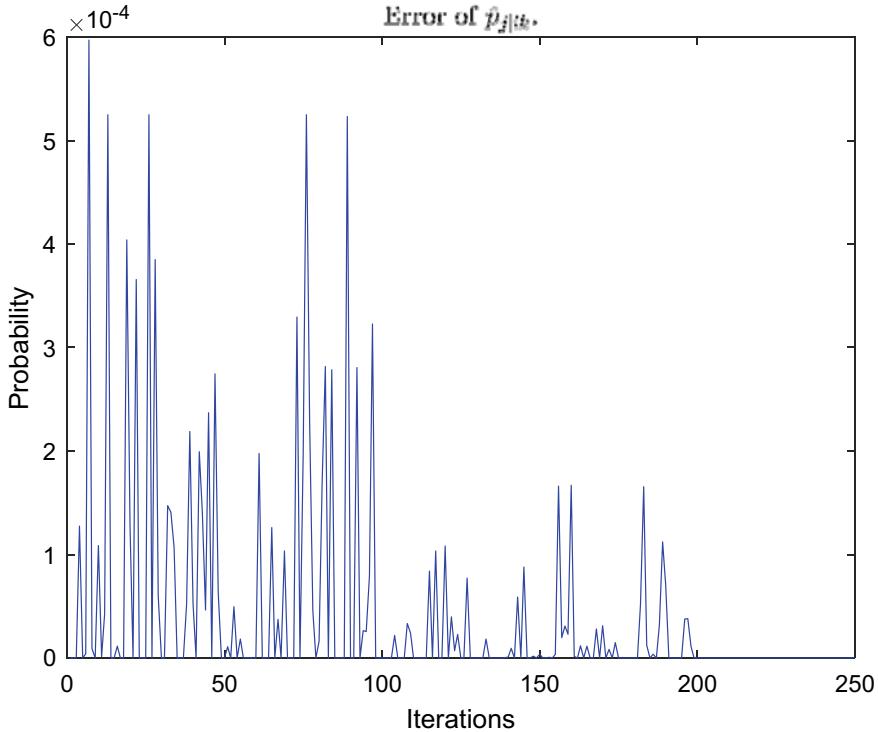


Fig. 3.3 Convergence of the error of $\hat{p}_{j|\hat{k}}$

The joint observer $d_{i|m}$ is conceptualized as the product of the distribution vector P_i and observation kernel $q_{i|m}$, which denotes the relationship between the real and the estimated state. The resulting observation kernel $q_{i|m}$ is interpreted as the intuition of the managers in their decision-making process, and it is a non-trivial result for the portfolio optimization problem.

Remark 3.2 For POMC the goal of the actions is the same as in the case of total observable Markov chains, but the sets of the admissible policies are different. It is natural that, for this class of admissible policies, and the lack of information about the states of the Markov chain, the objective function presents in a reduction of its maximum.

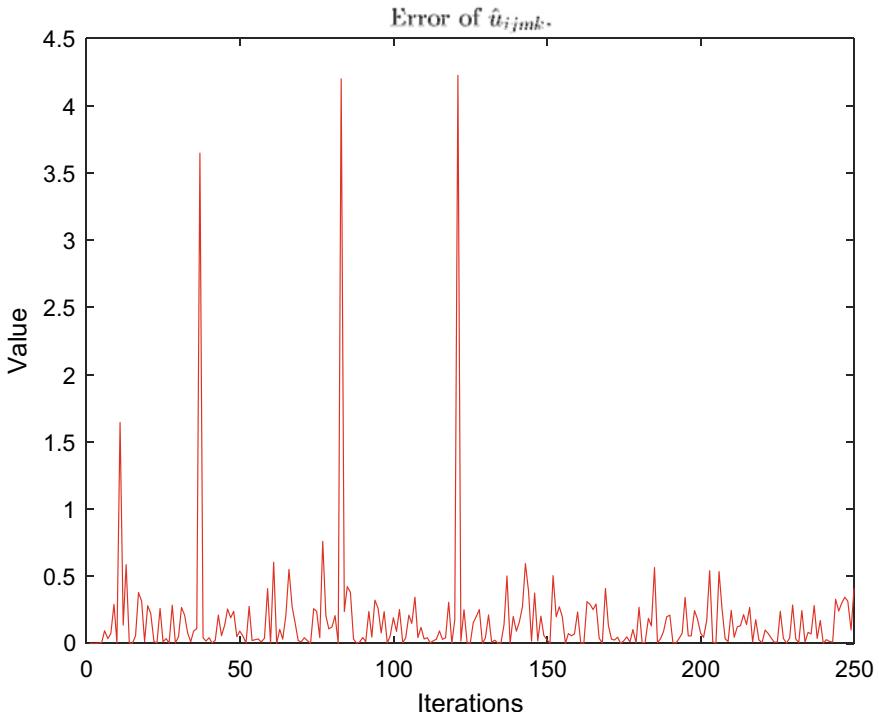


Fig. 3.4 Convergence of the error of $\hat{u}_{i\hat{j}\hat{m}\hat{k}}$

References

1. Asiain, E., Clempner, J.B., Poznyak, A.S.: A reinforcement learning approach for solving the mean variance customer portfolio for partially observable models. *Int. J. Artif. Intell. Tools* **27**(8), 1850034–1–1850034–30 (2018)
2. Bensoussan, A., Cakanyildirim, M., Sethi, S.P., Shi, R.: Computation of approximate optimal policies in a partially observed inventory model with rain checks. *Automatica* (2011)
3. Cassandra, A.R., Kaelbling, L.P., Littman, M.L.: Acting optimally in partially observable stochastic domains. In: Proceedings of Twelfth National Conference in Artificial Intelligence, vol. 2, pp. 1023–1028. Menlo Park, C.A., USA (1994)
4. Clempner, J.B.: Necessary and sufficient karush-kuhn-tucker conditions for multiobjective markov chains optimality. *Automatica* **71**, 135–142 (2016)
5. Clempner, J.B.: Revealing perceived individuals' self-interest. *J. Oper. Res. Soc.* 1–10 (2023). To be published. <https://doi.org/10.1080/01605682.2023.2195878>
6. Clempner, J.B., Poznyak, A.S.: Simple computing of the customer lifetime value: a fixed local-optimal policy approach. *J. Syst. Sci. Syst. Eng.* **23**(4), 439–459 (2014)
7. Clempner, J.B., Poznyak, A.S.: Sparse mean-variance customer markowitz portfolio optimization for markov chains: a tikhonov's regularization penalty approach. *Eng. Optim.* **19**(2), 383–417 (2018). <https://doi.org/10.1007/s11081-018-9374-9>
8. Clempner, J.B., Poznyak, A.S.: Observer and control design in partially observable finite markov chains. *Automatica* **10**, 108587 (2019)

9. Doshi, F., Pineau, J., Roy, N.: Reinforcement learning with limited reinforcement: using bayes risk for active learning in pomdps. In: Proceedings of the 25th International Conference on Machine Learning, vol. 301, pp. 256–263. Helsinki, Finland (2008)
10. Garcia-Galicia, M., Carsteanu, A.A., Clempner, J.: Continuous-time learning method for customer portfolio with time penalization. *Expert Syst. Appl.* **129**, 27–36 (2019)
11. Garcia-Galicia, M., Carsteanu, A.A., Clempner, J.: Continuous-time mean variance portfolio with transaction costs: a proximal approach involving time penalization. *Int. J. Gen. Syst.* **48**(2), 91–111 (2019)
12. Lai, Y., Elliott, R.J.: The mean squared loss control problem for a partially observed markov chain. *Int. J. Control.* (2017). To be published. <https://doi.org/10.1080/00207179.2017.1362503>
13. Lesser, K., Oishi, M.: Reachability for partially observable discrete time stochastic hybrid systems. *Automatica* **50**(8), 1989–1998 (2014)
14. Littman, M.L., Cassandra, A.R., Kaelbling, L.P.: Learning policies for partially observable environments: scaling up. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 362–370 (1995)
15. Ortiz-Cerezo, L., Carsteanu, A., Clempner, J.B.: Optimal constrained portfolio analysis for incomplete information and transaction costs. *Econ. Comput. Econ. Cybern. Stud. Res.* **4**(56), 107–121 (2022)
16. Ortiz-Cerezo, L., Carsteanu, A., Clempner, J.B.: Sharpe-ratio portfolio in controllable markov chains: Analytic and algorithmic approach for second order cone programming. *Mathematics* **10**(18), 3221 (2022)
17. Poznyak, A.S.: Advanced Mathematical tools for Automatic Control Engineers. Deterministic technique, vol. 1. Elsevier, Amsterdam, Oxford (2008)
18. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker Inc, New York (2000)
19. Sánchez, E.M., Clempner, J.B., Poznyak, A.S.: A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean-variance customer portfolio: the case of bank marketing campaigns. *Eng. Appl. Artif. Intell.* **46**, Part A, 82–92 (2015)
20. Sánchez, E.M., Clempner, J.B., Poznyak, A.S.: Solving the mean-variance customer portfolio in markov chains using iterated quadratic/lagrange programming: a credit-card customer-credit limits approach. *Expert Syst. Appl.* **42**(12), 5315–5327 (2015)
21. Sragovich, V.G.: Mathematical Theory of Adaptive Control. World Scientific Publishing Company (2006)
22. Talvitie, E., Singh, S.: Learning to make predictions in partially observable environments without a generative model. *J. Artif. Intell. Res.* **42**, 353–392 (2011)
23. Vazquez, E., Clempner, J.B.: Customer portfolio model driven by continuous-time markov chains: an l2 lagrangian regularization method. *Econ. Comput. Econ. Cybern. Stud. Res.* **2**, 23–40 (2020)
24. Whiting, R.G., Pickett, E.E.: On model order estimation for partially observed markov chains. *Automatica* **24**(4), 569–572 (1988)
25. Xie, L., Ugrinovskii, V.A., Petersen, I.R.: Finite horizon robust state estimation for uncertain finite-alphabet hidden markov models with conditional relative entropy constraints. *SIAM J. Control Optim.* **47**(1), 476–508 (2008)
26. Zhang, H.: Partially observable markov decision processes: a geometric technique and analysis. *Oper. Res.* **58**(1), 214–228 (2010)

Chapter 4

Continuous-Time Markov Chains



Abstract The c -variable approach is extended in this chapter by adding a new linear constraint for continuous time. This method's benefit is that it transforms the continuous-time Markov decision process problem into a discrete-time Markov decision process, where the linear constraints make the problem computationally tractable. Chemical reaction networks, where the concentration dynamics is described as a continuous-time Markov chain, serve as an example of the method's use. Using a state-discrete continuous-time Markov decision process, we provide a mathematical optimization method for resolving chemical processes. The first application is a single reversible reaction that produces the amidogen radical, and we were able to determine the ideal temperature that reduces a linear functional of interest. The second is a chemical reaction network that involves the proton transfer, hydration, and tautomeric reaction of anthocyanin pigments, in this case we found an optimal strategy over a set of values of pH that minimizes the corresponding linear functional.

4.1 Introduction

4.1.1 Related Work

In a *Continuous-Time Markov Decision Process* (CTMDP), which is a generalization of a discrete-time Markov chain, a decision maker or controller chooses actions (policies) to influence the behavior of a system as it evolves over time in order to make the system perform as well as possible in relation to some predetermined criterion. In order to find a policy that maximizes the average (long-run expected average reward) or discounted reward, CTMDPs are often examined across infinite horizons; for further information, see [5, 21].

The ideal policy for CTMDPs is obtained in two steps: first, the optimal value function for the relevant optimality criterion is defined, and then the optimality policies are found. For this, they employ the dynamic programming method, which entails

characterizing the CTMDP optimal value function as the solution of an optimality equation, also referred to as Bellman's equation or the Hamilton–Jacobi–Bellman equation. They then demonstrate that the deterministic stationary policy achieves the supremum in the optimality equation. The CTMDP is converted into an analogous linear programming problem [17].

The book [15] presents a CTMDP with unbounded transition rates and cost and offers a policy iteration approach to approximate the average reward optimum policy. The same CTMDP, i.e., transition rates and cost unbounded, is described in [14], where the control issue in the class of Deterministic Stationary Policies is investigated with regard to the Discounted Cost Criteria without offering a method to solve the Optimal Policy. While applying the discounted reward optimality equation and the value iteration technique, reports the best policy.

The authors [10, 25] report on the computation of the best policies for CTMDPs with finite state and action spaces in terms of the long-run anticipated average reward. There, the CTMDP is solved using either the linear programming approach or the policy iteration technique. The best course of action can be determined from a Discrete-Time Markov Decision Process (DTMDP), which is produced from the original CTMDP using uniformization (randomization) [3]. Under fairly general conditions, the best course of action is stationary, meaning it depends only on the state.

A discrete-time Markov decision process that is subordinated to a Poisson process emerges from uniformization of a CTMC. In homogeneous Markov models, algorithms based on uniformization have been used with great success (see, for example, [13, 25, 31]). This has shown to be an effective method for solving computational and theoretical issues [9, 18]; one use of this method is to speed up policy improvement methods by adding more decision epochs artificially (tijms 1986). The CTMDP is restated in [25] as an average reward, and then uniformization is used to apply discrete time approaches (policy iteration algorithm, value iteration algorithm or linear programming). Several applications have been presented in the literature [1, 6, 7, 11, 29].

When the generator matrix varies at certain time intervals in a CTMC, uniformization can be used to identify the transient solution [2]. The CTMC is not discretized in a unique way by uniformization, which is a drawback.

In particular, the linear programming formulation described in [15, 25] presents a solution for the optimal policy is based on the introduction of the n -bias optimality criteria; it is calculated as an inequality that is expressed as a primal linear program, which is then converted into a dual linear program. This transformation reduces the issue to the minimization of a linear functional, subject to linear and equality constraints, and makes it straightforward to resolve. If the ergodicity property is true, it is further demonstrated that the optimum policy's solution is distinct. This method is limited to actions and states with boundaries.

The main results of this chapter are the following:

- Presents an optimization approach for solving an average optimality criterion in CTMDP, in addition the resulting approach is applied to kinetic chemical reactions for the minimization of a average reward function of interest.

- Extends the c-variable method and introduces a new special linear constraint for continuous time. The advantage of our approach is that it reduces the CTMDP problem to a discrete-time Markov decision processes. Then, the CTMDP is easy to conceptualize and makes the problem computationally tractable.
- Applies the results to two problems of chemical kinetics. The first analyzes a reversible reaction modeled as a CTMC through a model of birth-death process, where the constants rates are function of temperature, we illustrate the method across the minimization of the cost for the average rate production of H and observed that for the minimization of the functional a high temperature is required. The second analyzes a CRN of three reversible reactions, where the pH influences the reaction constant rates, such that the actions are a set of pH values. We illustrate the method across the minimization the expected average number of molecules, and we found that for the minimization of the functional a mixed strategy is required.
- Models the kinetics of reactions as a CTMDP and solve the minimization of the long-run expected average reward that is a common optimality criterion for CTMDP in particular for finite control models [25, 27].

It is important to mention that in both examples a proposed reaction mechanism is necessary as well as the constant rates of the reaction. The solution presented for the CTMDP is recommended only for solving optimization problems where the number of molecules is small and for CRN that satisfy the ergodicity property.

4.2 Continuous-Time Markov Chains

In this section we introduce the (continuous-time, time-homogeneous) Markov process model we are interested in. As usual, \mathbb{R} and \mathbb{N} stand for the sets of real numbers and non-negative integers, respectively.

Let $S = \{x_1, \dots, x_N : N \in \mathbb{N}\}$ be a finite set called the *state space*, $\{X(t), t \geq 0\}$ a stochastic process with state space S , satisfies the *Markov property* if, letting $\mathcal{F}_{X(\tau)}$ denote all the information pertaining to the history of X up to time τ , and $\tau \leq t$ with $j, i \in S$

$$P(X(t) = j | \mathcal{F}_{X(\tau)}) = P(X(t) = j | X(\tau) = i)$$

we say that the process is time homogeneous if, $t = s + \tau$. By simple words, this property means that any distribution in the future depends only on the value $X(\tau)$ and is independent on the past values.

Throughout the remainder

$$CTMDP = (S, A, \{A(s)\}_{s \in S}, Q, r) \tag{4.2.1}$$

stands for a *continuous-time Markov decision process* (CTMDP), where the state space S is a *finite* set $\{s_{(1)}, \dots, s_{(N)}\}$, $N \in \mathbb{N}$, endowed with the discrete topology

and the action set A is the action (or control) space, a *metric space* endowed with the corresponding Borel σ -algebra $\mathcal{B}(A)$.

For each $s \in S$, $A(s) \subset A$ is the nonempty set of admissible actions at s and we shall suppose that it is *compact*. Whereas, the set $\mathbb{K} := \{(s, a) : s \in S, a \in A(s)\}$ is the class of admissible pairs, which is considered as a topological subspace of $S \times A$.

The matrix $Q = [q_{j|ik}]_{i,j=1,\overline{N},k=1,\overline{M}}$ denotes the *transition rates* which satisfy that $q_{j|ik} \geq 0$ for all $x \in S$ and $j \neq i$. The transition rates $q_{j|ik}$ are conservative, i.e.,

$$\boxed{\sum_{j=1}^N q_{j|ik} = 0} \quad (4.2.2)$$

and stable, which means that

$$q_i^* := \sup_{a \in A(i)} q_i(a) < \infty \forall i \in S$$

where

$$\boxed{q_i := -q_{i,i} \geq 0 \text{ for all } x \in S.} \quad (4.2.3)$$

Finally, $r \in \mathcal{B}(\mathbb{K})$ is the (measurable) one-stage cost function.

We denote the probability transition matrix by

$$\Pi(t) = [\pi_{(s,i,\tau,j,k)}]_{i,j=1,\overline{N},k=1,\overline{M}}, \tau \geq s$$

such that,

$$\pi_{(s,i,\tau,j,k)} = \pi_{(0,i,t,j,k)}, t = \tau - s \forall i, j \in S, \quad (4.2.4)$$

and where $\sum_{j=1}^N \pi_{j|ik} = 1, \forall i \in S$.

The *Kolmogorov forward equations*, can be written as the matrix differential equation:

$$\boxed{\Pi'(t) = \Pi(t)Q; \Pi(0) = I} \quad (4.2.5)$$

$\Pi(t) \in \mathbb{R}^{N \times N}$, $I \in \mathbb{R}^{N \times N}$ is the identity matrix. This system can be solved by

$$\Pi(t) = \Pi(0)e^{Qt} = e^{Qt} := \sum_{k=0}^{\infty} \frac{t^k Q^k}{k!}$$

and at the stationary state, the probability transition matrix is defined as

$$\Pi^* = \lim_{t \rightarrow \infty} \Pi(t). \quad (4.2.6)$$

We also point out that given a state space S , the infinitesimal generator Q completely determines the CTMC. Thus, it is sufficient to characterize a chain by simply providing a state space S , and the generator Q .

Definition The vector $p \in \mathbb{R}^N$ is called *stationary distribution vector* if

$$\Pi^{\top *} p = p,$$

where $\sum_{i=1}^N p_i = 1$ and

$$p_i = P(X(t) = i).$$

This vector can be seen as the long run proportion of time that the process is in state $i \in S$.

Theorem Let $X(t)$ be an irreducible and recurrent CTMC, then the following statements are equivalent:

- $Q^\top p = 0$,
- $\Pi^{\top *} p = p$.

Remark The proof of this fact is easy in the case of a finite state space, recalling the Kolmogorov backward equation.

A *strategy* is then defined as a sequence $d = \{d(t), t \geq 0\}$ of stochastic kernels $d(t)$ such that:

- (a) for each time $t \geq 0$ $d_{(k|i)}(t)$ is a probability measure on A such that $d_{(A(i)|i)}(t) = 1$ and,
- (b) for every $E \in \mathcal{B}(A)$ $d_{(E|i)}(t)$ is a Borel measurable function in $t \geq 0$.

We denoted by D the family of all strategies. From now on, we will consider only stationary strategies $d_{(k|i)}(t) = d_{(k|i)}$.

For each action the matrix $Q(a) := [q_{j|i,k}]$, $a \in A$ denotes the transition rates matrix for the action a such that

$$[q_{ij}(a)] := [q_{j|i,k}] = \begin{cases} -\sum_i \lambda_{ij}(a), & \text{if } i = j \\ \lambda_{ij}(a), & \text{if } i \neq j \end{cases}$$

while, for each strategy d the associated transition rate matrix is defined as:

$$Q(d) := [q_{ij}(d)] = \sum_{k=1}^M q_{j|i,k} d_{k|i},$$

(4.2.7)

such that on a stationary state distribution for all $d_{k|i}$ and $t \geq 0$ from (4.2.6) we have that

$$\Pi^*(d) = \lim_{T \rightarrow \infty} e^{Q(d)T},$$

where $\Pi^*(d)$ is a stationary transition controlled matrix.

$$\Pi^*(d) := [\pi_{ij}(d)] = \sum_{k=1}^M \pi_{j|ik} d_{k|i}. \quad (4.2.8)$$

Moreover, let $r(a)$ be the vector whose components $r_i(a) := r_{ik}$ are the reward of the state $i \in S$ given the action $a_k \in A$, we can express the reward vector in terms of the strategy d as:

$$r(d) := [r_i(d)] = \sum_{k=1}^M r_{ik} d_{k|i}.$$

Definition Let us denote by $J(d)$ the long-run expected average reward, also referred to as the gain of a policy d , as the vector

$$\mathbf{J}(d) := \liminf_{T \rightarrow \infty} \frac{\frac{1}{T} \int_0^T \Pi(t, d) r(d) dt}{T}$$

with i th component $\mathbf{J}(i, d)$. At the steady state

$$J^*(d) = \Pi^*(d) r(d).$$

Pondering the long-run expected average reward over the states at steady state the following linear functional $J(d)$ under the fixed strategy $d_{k|i}(t) = d_{k|i}$ can be defined as follows

$$J(d) := \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M r_{ik} \pi_{j|ik} d_{k|i} p_i = \sum_{i=1}^N \sum_{k=1}^M r_{ik} d_{k|i} p_i. \quad (4.2.9)$$

Formulation of the problem. In an ergodic chain, the steady-state probabilities are independent of the initial state of the system. In the long run, the average cost per stage should be a constant regardless of the initial state. Because the controllable ergodic Markov chains reach a stationary situation exponentially quickly, we consider a stationary distribution and by the property in (4.2.7), the value function (4.2.9), under a given optimal policy is:

$$J(d) = \sum_{i=1}^N \sum_{k=1}^M r_{ik} d_{k|i} p_i(d) \longrightarrow \min_d \quad (4.2.10)$$

4.3 Programming Solver for CTMDP

In this section we present the extended c-variable method and the linear programming solver for CTMDP.

4.3.1 The c-Variable Method

The model presented in Eq. (4.2.10) is not linear. Then, if we define new decision variables the problem presented above can be reformulated as a linear programming as follows. Let introduce the *joint strategy variable* c_{ik} as

$$c_{ik} = p_i d_{k|i}.$$

The variable c_{ik} belongs to the set of matrices $c \in C_{adm}$ and is restricted by the following constraints:

1. each vector from the matrix $c := [c_{ik}]$ represents a stationary mixed-strategy that belongs to the simplex

$$\mathcal{S}^{N \times M} := \begin{cases} c \in \mathbb{R}^{N \times M} \text{ for } c_{ik} \geq 0, \\ \text{where } \sum_{i=1}^N \sum_{k=1}^M c_{ik} = 1; \end{cases} \quad (4.3.1)$$

2. the variable c_{ik} satisfies the ergodicity constraints i.e.:

$$p_j(d) = \sum_{i=1}^N \sum_{k=1}^M \pi_{j|ik} p_i d_{k|i},$$

that in terms of c_{ik} takes the form:

$$\sum_{k=1}^M c_{jk} = \sum_{i=1}^N \sum_{k=1}^M \pi_{j|ik} c_{ik}. \quad (4.3.2)$$

3. From Eq.(4.2.7):

$$\sum_{i=1}^N q_{ij}(d) p_i(q) = \sum_{i=1}^N \sum_{k=1}^M q_{j|ik} d_{k|i} p_i = 0.$$

this expression in terms of c_{ik} takes the form:

$$\boxed{\sum_{i=1}^N \sum_{k=1}^M q_{j|ik} c_{ik} = 0.} \quad (4.3.3)$$

Once the model (ergodic Markov decision process) is solved in order to recover the quantities of interest we have that:

- **Stationary Distribution**

$$\boxed{p_i(d) = \sum_{k=1}^M c_{ik}.} \quad (4.3.4)$$

- **Policy**

$$\boxed{d_{ik} = \frac{c_{ik}}{\sum_{k=1}^M c_{ik}}.} \quad (4.3.5)$$

Then, in terms of c -variables the reward function $J(d)$ becomes:

$$\boxed{J(c) = \sum_{i=1}^N \sum_{k=1}^M r_{ik} d_{k|i} p_i(d) = \sum_{i=1}^N \sum_{k=1}^M r_{ik} c_{ik}.} \quad (4.3.6)$$

4.3.2 Linear Programming Solver

Consider the problem:

$$C^T X \rightarrow \min_X,$$

such that

$$A_{eq} X = b_{eq},$$

where

$$0 \leq X \leq 1, A_{eq} \in \mathbb{R}^{(2N+1) \times (NM)} \text{ and } b_{eq} \in \mathbb{R}^{(2N+1)}.$$

The vector $C^T := C(N|M)$ is defined as:

$$\begin{aligned} C(1|1) &= r_{11}, \dots, C(1|M) = r_{1M} \\ C(2|1) &= r_{21}, \dots, C(2|M) = r_{2M}, \dots, \\ C(N|1) &= r_{N1}, \dots, C(N|M) = r_{NM}. \end{aligned}$$

(1) We will construct the matrix $A \in \mathbb{R}^{N \times (NM)}$ using the ergodicity constraints defined in (4.3.1) as

$$0 = \sum_{k=1}^M \left(\sum_{i=1}^N \pi_{j|ik} c_{ik} - c_{jk} \right).$$

Then, we have that

$$\begin{aligned} j = 1 \quad & \sum_{k=1}^M \left(\sum_{i=1}^N \pi_{j|ik} c_{ik} - c_{1k} \right) = 0, \\ & \vdots \qquad \qquad \qquad \vdots \\ j = N \quad & \sum_{k=1}^M \left(\sum_{i=1}^N \pi_{j|ik} c_{ik} - c_{Nk} \right) = 0. \end{aligned} \tag{4.3.7}$$

Developing the formulas of (4.3.7) we have that

$$A = [\pi_{j|ik} - \delta_{j,i}]_{j=\overline{1,N} i=\overline{1,N}},$$

where $\delta_{(j,i)}$ is the Kronecker's delta.

(2) For satisfying the time constraints defined in Eq.(4.3.2) we will construct the matrix $B \in \mathbb{R}^{N \times (NM)}$ as follows

$$B = [q_{j|ik}]_{j=\overline{1,N} i=\overline{1,N}}.$$

(3) Finally, for satisfying the ergodicity constraints defined in (4.3.1) as $c_{ik} \in \mathcal{S}^{(N \times M)}$, we have that

$$A_{eq} = \begin{bmatrix} A \\ B \\ e^\top \end{bmatrix},$$

where $e = (1, \dots, 1)^\top \in \mathbb{R}^{(NM)}$ and the vector b_{eq} is defined as

$$b_{eq} = (\underbrace{0, \dots, 0}_{2N \text{ times}}, 1)^T \in \mathbb{R}^{2N+1}.$$

4.4 Chemical Reaction Markov Models

In this section we follow [4].

A *Chemical Reaction Network* (CRN) comprises a set of reactants, a set of products, and a set of reactions. Dynamical properties of CRNs are studied by means of differential equations (deterministic models) [16, 23], or by a kinetic scheme, which describes the dynamics of variables by a network of states and connections between them (stochastic models). In our case the time-dependent concentrations of the chemical species.

The stochastic models for CRNs are represented in terms of Poisson processes. This random time-change representation produces a stochastic equation that can be formulated as continuous-time Markov chains (CTMC). The states of the chemical reaction network are a set with the number of molecules for every chemical specie and the reactions are the possible transitions of the chain. In the CRN, the molecules have collisions randomly and they may undertake chemical reactions, which modify the state of the system. The rate at which the transitions occur is given by the law of mass action.

The graph of the CTMC is as follows: the nodes are elements where the i -th component represents a set with the number of molecules of every chemical specie. An edge $A \rightarrow B$ represents a positive rate of transition between nodes A and B , which in turn exists if there is a reaction in the CRN that allows this transition. Then, the graph of the stochastic model is associated with the CTMC and it inherits the structural elements from the graph of the CRN. Different results have shown that the discreteness and randomness of the chemical reactions need to be considered for certain applications [24, 26, 30, 32]. Then, CTMC models have increased importance for describing the dynamics in chemical reaction networks.

4.4.1 Example 1. Formation of the Amidogen Radical

In this example we consider the reaction



performed by the flash photolysis-shock tube technique, using atomic resonance absorption to monitor the concentration of H over the time, namely $[H(t)]$. A set of experiments at different temperatures, over the range of $900\text{--}1620\text{ K}$ were performed in [28]. The initial concentration of $[H]$ and $[NH_2]$ was the same i.e., $[H]_0 = [NH_2]_0$ and both the $[NH_3]$ and $[H_2]$ were maintained in large excess such that the kinetics is simplified to a system of two opposing first-order reactions described in equation (4.4.1). The Table 4.1 shows the experimental results reported in (Sutherland and Michael [28]) for the reaction rate at different temperatures:

Table 4.1 Kinetic data of the reaction: $H + NH_3 \rightleftharpoons NH_2 + H_2$, T(K), $k_1(10^{-13}) s^{-1}$, $k_{-1}(10^{-13}) s^{-1}$

T	k_1	k_{-1}
1260	5.01	3.58
1482	13.06	5.94
1620	20.77	9.23

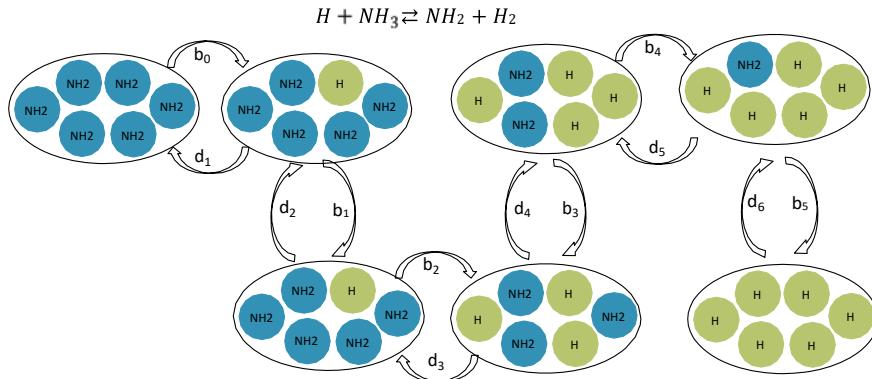


Fig. 4.1 Schematic illustration of Continuous time Markov Chain for the reaction $H + NH_3 \rightleftharpoons H_2 + NH_2$, $N = 7$, $[H(t)] = \{0, 1, 2, \dots, 6\}$

To describe the dynamics of $[H(t)]$, we propose a model of CTMC based on a *birth-death process*. Where the state $S = [H(t)] = \{0, 1, 2, \dots, N - 1\}$, is the feasible number of molecules of H that are converted into new molecules of H (birth), or consumed (death), with rates given by the mass action law, $b_n(T)$ and $d_n(T)$ respectively:

$$b_n(T) = k_{-1}(T)(N - n), \quad d_n(T) = k_1(T)n, \quad \forall n \in S.$$

The Fig. 4.1, shows the graph representation for the CTMC with $N = 7$, where the i -th state corresponds to i number of molecules of H under certain action a , and b_i, d_j are the birth and death rates respectively, for all $i, j \in S$.

For a birth-death process the generator matrix [22] for a fix N takes the form:

$$Q(T) = \begin{bmatrix} -b_0(T) & b_0(T) & 0 & \dots & 0 & 0 \\ d_1(T) & -(b_1(T) + d_1(T)) & b_1(T) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & d_{N-1}(T) & -d_{N-1}(T) \end{bmatrix}, \quad (4.4.2)$$

where $N = [H]_0 + [NH_2]_0$. The constants $k_{-1}(T)$ and $k_1(T)$ are function of the temperature, this variable is used as a control for the direction of the reaction, i.e., $A = \{T_1, T_2, \dots, T_k\}$ where A is a set of admissible temperatures. For the

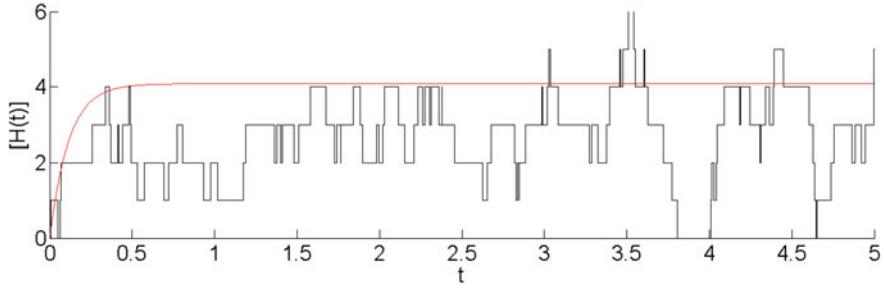


Fig. 4.2 Comparison between CTMC for the embedded chain for $[H(t)]$, under a pure strategy $T = 1260$ and $Q(1260)$ (line black), and the ODE model (line red)

sake of simplicity we consider a chain with 7 states, $N = 7$, $S = \{0, 1, 2, \dots, 6\}$ and a set of 3 actions, $k = 3$, such that the set of admissible temperatures is $A = \{1260, 1482, 1620\}$. Using the values for $k_1(T)$ and $k_{-1}(T)$ reported in Table 4.1, and according to the expression in (4.4.2), the following matrix generators are obtained:

$$Q(T_1) = \begin{bmatrix} -21.48 & 21.48 & 0 & 0 & 0 & 0 & 0 \\ 5.01 & -22.91 & 17.90 & 0 & 0 & 0 & 0 \\ 0 & 10.02 & -24.34 & 14.32 & 0 & 0 & 0 \\ 0 & 0 & 15.03 & -25.77 & 10.74 & 0 & 0 \\ 0 & 0 & 0 & 20.04 & -27.20 & 7.16 & 0 \\ 0 & 0 & 0 & 0 & 25.05 & -28.63 & 3.58 \\ 0 & 0 & 0 & 0 & 0 & 30.06 & -30.06 \end{bmatrix},$$

$$Q(T_2) = \begin{bmatrix} -35.64 & 35.64 & 0 & 0 & 0 & 0 & 0 \\ 13.06 & -42.76 & 29.70 & 0 & 0 & 0 & 0 \\ 0 & 26.12 & -49.88 & 23.76 & 0 & 0 & 0 \\ 0 & 0 & 39.18 & -57.00 & 17.82 & 0 & 0 \\ 0 & 0 & 0 & 52.24 & -64.12 & 11.88 & 0 \\ 0 & 0 & 0 & 0 & 65.30 & -71.24 & 5.94 \\ 0 & 0 & 0 & 0 & 0 & 78.36 & -78.36 \end{bmatrix},$$

$$Q(T_3) = \begin{bmatrix} -55.38 & 55.38 & 0 & 0 & 0 & 0 & 0 \\ 20.77 & -66.92 & 46.15 & 0 & 0 & 0 & 0 \\ 0 & 41.54 & -78.46 & 36.92 & 0 & 0 & 0 \\ 0 & 0 & 62.31 & -90.00 & 27.69 & 0 & 0 \\ 0 & 0 & 0 & 83.08 & -101.54 & 18.46 & 0 \\ 0 & 0 & 0 & 0 & 103.85 & -113.08 & 9.23 \\ 0 & 0 & 0 & 0 & 0 & 124.62 & -12.62 \end{bmatrix}.$$

In the following, we consider $[H]_0 = 0$ and $[NH_2]_0 = 6$, the initial distribution of the CTMC is $p_0 = [1, 0, 0, 0, 0, 0, 0]$, using the embedded chain, generated by the matrix $Q(T)$ is possible to develop the construction of the CTMC as shown in [20]. The Fig. 4.2 shows the comparison between the model for the CTMC and the ODE of the deterministic kinetic model $dH(t)/dt = k_{-1}(N - H(t)) - k_1 H(t)$.

The probability transition matrix $\Pi^*(T)$ is calculated by the Chapman–Kolmogorov equation in (4.2.6). In the biology literature this system of equations is termed *Chemical Master Equation* [8, 19].

$$\Pi^*(T_1) = \begin{bmatrix} 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \\ 0.0394 & 0.1688 & 0.3015 & 0.2872 & 0.1539 & 0.0440 & 0.0052 \end{bmatrix},$$

$$\Pi^*(T_2) = \begin{bmatrix} 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \\ 0.1015 & 0.2878 & 0.3273 & 0.1985 & 0.0677 & 0.0123 & 0.0009 \end{bmatrix},$$

$$\Pi^*(T_3) = \begin{bmatrix} 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \\ 0.1101 & 0.2936 & 0.3262 & 0.1933 & 0.0644 & 0.0115 & 0.0008 \end{bmatrix}.$$

To illustrate the optimization problem, suppose that a negative reward (cost) is obtained according to the velocity of the individual reaction in the direction $NH_2 + H_2 \longrightarrow H + NH_3$, this idea is reflected in the following cost vector:

$$r_n = -k_1(T)n \quad \forall n \in S,$$

then the functional (4.2.9) to minimize, represents the mathematical expectation of the rate at which consumes $[H]$, once the steady state is reach.

The rewards $r(T)$ for every temperature are:

$$r(T_1) = \begin{bmatrix} 0 \\ -5.01 \\ -10.02 \\ -15.03 \\ -20.04 \\ -25.05 \\ -30.06 \end{bmatrix}, r(T_2) = \begin{bmatrix} 0 \\ -13.06 \\ -26.12 \\ -39.18 \\ -52.24 \\ -65.30 \\ -78.36 \end{bmatrix}, r(T_3) = \begin{bmatrix} 0 \\ -20.77 \\ -41.54 \\ -62.31 \\ -83.08 \\ -103.85 \\ -124.62 \end{bmatrix}.$$

The results of the c-variable method with the constraints described in expressions (4.3.1), (4.3.2), (4.3.3), are presented below:

$$c = \begin{bmatrix} 2.5 \times 10^{-11} & 4.4 \times 10^{-11} & 1.1 \times 10^{-1} \\ 9.4 \times 10^{-11} & 1.0 \times 10^{-10} & 2.9 \times 10^{-1} \\ 1.3 \times 10^{-10} & 1.4 \times 10^{-10} & 3.2 \times 10^{-1} \\ 1.0 \times 10^{-10} & 1.2 \times 10^{-10} & 1.9 \times 10^{-1} \\ 3.6 \times 10^{-11} & 8.7 \times 10^{-11} & 6.4 \times 10^{-2} \\ 1.1 \times 10^{-12} & 4.1 \times 10^{-11} & 1.1 \times 10^{-2} \\ 2.3 \times 10^{-12} & 3.1 \times 10^{13} & 8.4 \times 10^{-4} \end{bmatrix}, d = \begin{bmatrix} 2.2 \times 10^{-10} & 4.0 \times 10^{-10} & 9.9 \times 10^{-1} \\ 3.2 \times 10^{-10} & 3.6 \times 10^{-10} & 9.9 \times 10^{-1} \\ 4.0 \times 10^{-10} & 4.5 \times 10^{-10} & 9.9 \times 10^{-1} \\ 5.3 \times 10^{-10} & 6.5 \times 10^{-10} & 9.9 \times 10^{-1} \\ 5.6 \times 10^{-10} & 1.3 \times 10^{-9} & 9.9 \times 10^{-1} \\ 9.8 \times 10^{-11} & 3.6 \times 10^{-9} & 9.9 \times 10^{-1} \\ 2.7 \times 10^{-9} & 3.7 \times 10^{-10} & 9.9 \times 10^{-1} \end{bmatrix}.$$

From the results obtained in d observe that the optimal temperature that minimize the functional bellow is $T_3 = 1620^\circ\text{C}$ (Fig. 4.3).

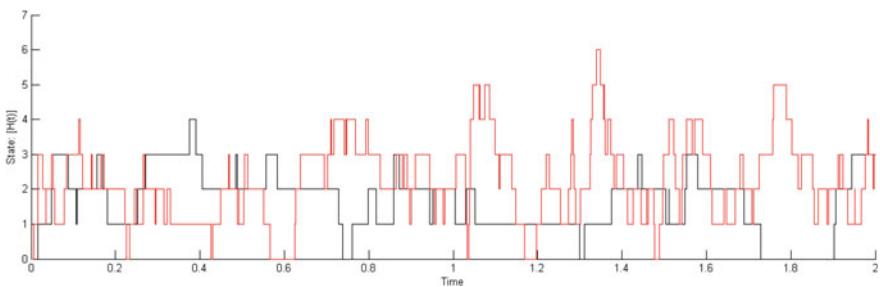
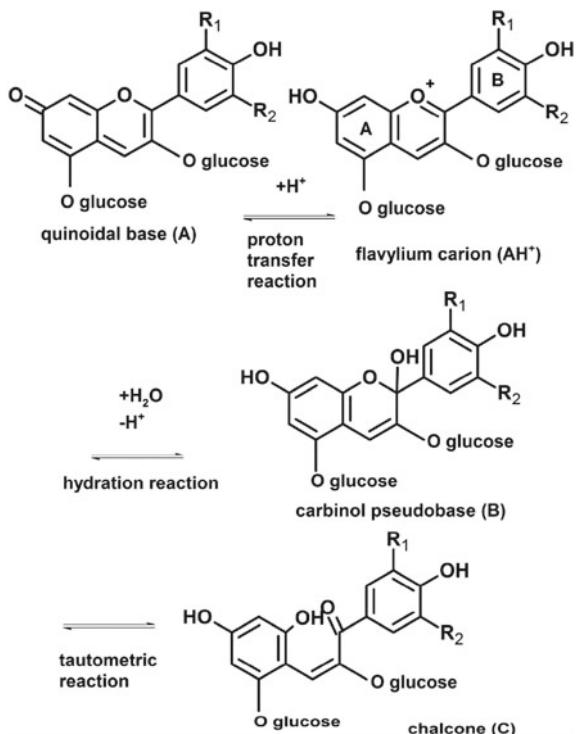


Fig. 4.3 Comparison between CTMC for the embedded chain for the state, $[H(t)]$, under a pure strategy $T = 1260$ and $Q(1260)$ (line black), and the CTMC for the embedded chain under a mixed strategy d , generated by $Q(d)$

Fig. 4.4 Proton transfer, hydration and tautomeric reaction of anthocyanin pigments



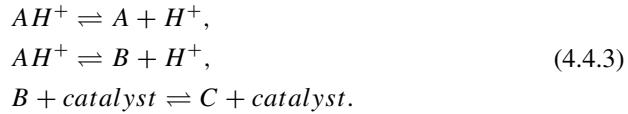
4.4.2 Example 2. Proton Transfer, Hydration and Tautomeric Reaction of Anthocyanin Pigments

In this example we study a CRN of three reversible reactions: proton transfer, hydration and tautomeric reaction of anthocyanin pigments.

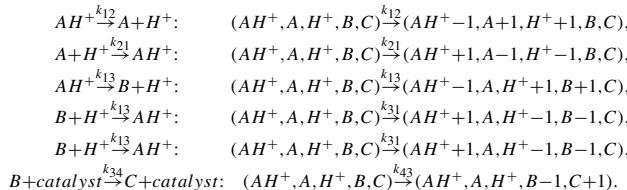
The CRN of Fig. 4.4 is represented symbolically by:

Table 4.2 Rate constants of the structural transformations of Malvidin 3-Glucoside in 0.2 M Ionic Aqueous Acid Medium at 25°C

$k'_{12} = 4.7(\pm 0.4) \times 10^4 s^{-1}$	$k'_{21} = 6.7(\pm 0.5) \times 10^8 M^{-1}s^{-1}$
$k'_{13} = 8.5(\pm 1) \times 10^{-2} s^{-1}$	$k'_{31} = 34(\pm 3) M^{-1}s^{-1}$
$k'_{43} = 3.8(\pm 0.2) \times 10^{-4} s^{-1}$	$k'_{34} = 4.5(\pm 0.3) \times 10^{-5} s^{-1}$



We model the dynamics of the CRN as a CTMC where every state represent a possible path for the reactions in (4.4.3). The construction of the chain states is as follows:



In particular for the initial condition $(AH_0^+, A_0, H_0^+, B_0, C_0), (AH_0^+, 0, H_0^+, 0, 0) = (2, 0, 5, 0, 0)$, the transitions between the states are schematized in Fig. 4.5.

The experimental rate constants for the CRN are reported in Table 4.2.

When the order of the chemical reaction is greater than one, the relationship of the stochastic reaction rate with the deterministic reaction rate depends on both, the volume of the system and the numbers of each reactant required for the reaction [12], therefore the kinetic constants for CTMDP used are $k_{12} = 4.7 \times 10^4 s^{-1}$, $k_{21} = 1.1 \times 10^2 s^{-1}$, $k_{13} = 8.5 \times 10^{-2} s^{-1}$, $k_{31} = 5.5 \times 10^{-5} s^{-1}$, $k_{34} = 4.5 \times 10^{-5} s^{-1}$, $k_{43} = 43.8 \times 10^{-4} s^{-1}$.

The control action are the pH of the medium, i.e., the number of molecules of H^+ such that, $A = \{H_1^+, H_2^+, H_3^+\} = \{5, 120, 530\}$.

The set of actions A , results in the following generator matrices:

$$Q(H_1^+) = \begin{bmatrix} -7.7 \times 10^2 & 3.8 \times 10^{-4} & 0 & 0 & 0 & 0 & 0 & 0 & 7.7 \times 10^2 \\ 4.5 \times 10^{-5} & -7.7 \times 10^2 & 0 & 3.08 \times 10^{-3} & 0 & 7.7 \times 10^2 & 0 & 0 & 0 \\ 0 & 0 & -1.54 \times 10^3 & 1.54 \times 10^3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8.5 \times 10^{-2} & 4.7 \times 10^4 & -4.76 \times 10^4 & 6.6 \times 10^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9.4 \times 10^4 & -9.4 \times 10^4 & 1.7 \times 10^{-1} & 0 & 0 & 0 \\ 0 & 4.7X10^4 & 0 & 0 & 2.31X10^{-3} & -4.7X10^4 & 8.5X10^{-2} & 0 & 4.5X10^{-5} \\ 0 & 0 & 0 & 0 & 0 & 6.93X10^{-3} & -6.975X10^{-3} & 4.5X10^{-5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.8X10^{-4} & -3.505X10^{-3} & 4.5X10^{-5} & 3.08X10^{-3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.8X10^{-4} & -3.8X10^{-4} & 0 \\ 4.7X10^4 & 0 & 0 & 0 & 0 & 0 & 0 & 8.5X10^{-2} & 0 & -4.7X10^4 \end{bmatrix},$$

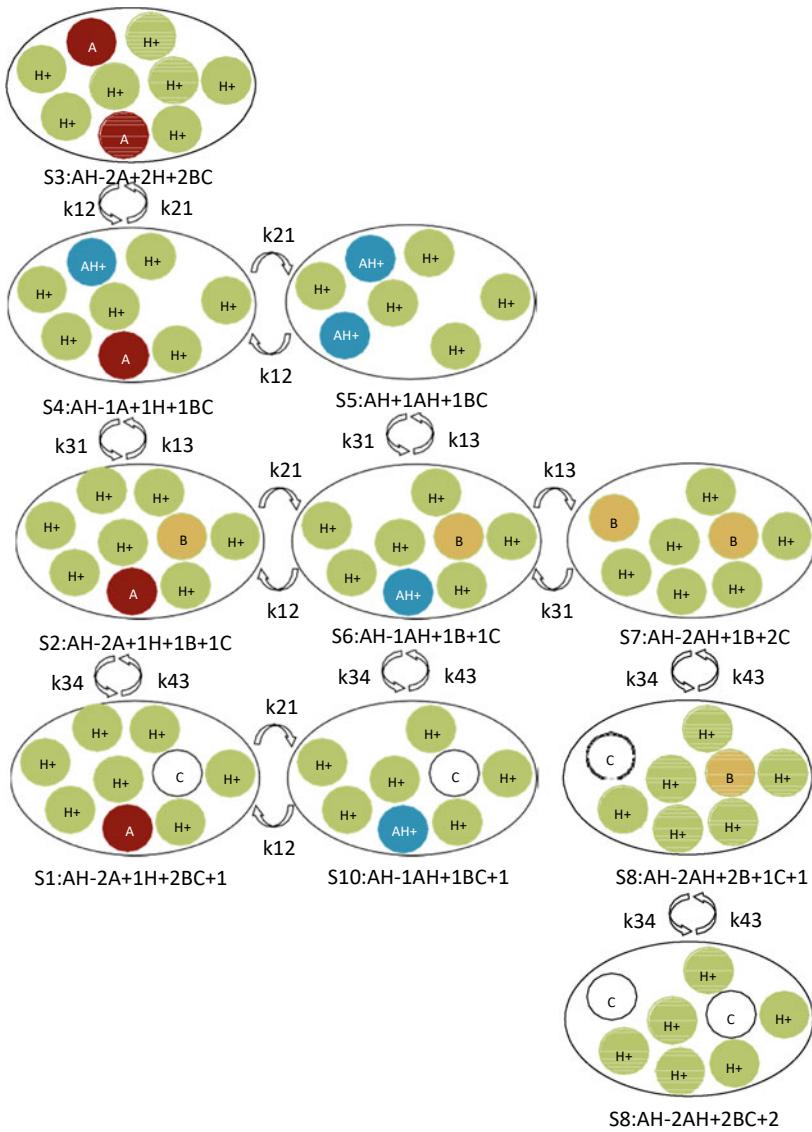


Fig. 4.5 Schematic representation of reaction as a continuous Markov Chain under action $H_1^+ = 5$

$$Q(H_2^+) = \begin{bmatrix} -1.342 \times 10^4 & 3.8 \times 10^{-4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.342 \times 10^4 \\ 4.5 \times 10^{-5} & -1.3421 \times 10^4 & 0 & 8.2533 \times 10^{-2} & 0 & 1.342 \times 10^4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2.684 \times 10^4 & 2.684 \times 10^4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8.5 \times 10^{-2} & 4.7 \times 10^4 & -6.031 \times 10^4 & 1.331 \times 10^4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9.4 \times 10^4 & -9.4 \times 10^4 & 1.7 \times 10^{-1} & 0 & 0 & 0 & 0 \\ 0 & 4.7 \times 10^4 & 0 & 0 & 8.1191 \times 10^{-1} & -4.7 \times 10^4 & 8.5 \times 10^{-2} & 0 & 0 & 4.5 \times 10^{-5} \\ 0 & 0 & 0 & 0 & 0 & 1.6671 & -1.6641 & 4.5 \times 10^{-5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.8 \times 10^{-4} & -8.2576 \times 10^{-1} & 4.5 \times 10^{-5} & 8.151 \times 10^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.8 \times 10^{-4} & -3.8 \times 10^{-4} & 0 \\ 4.7 \times 10^4 & 0 & 0 & 0 & 0 & 0 & 0 & 8.5 \times 10^{-2} & 0 & -4.7 \times 10^4 \end{bmatrix}.$$

$$Q(H_3^+) = \begin{bmatrix} -1.419 \times 10^5 & 3.8 \times 10^{-4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.419 \times 10^5 \\ 4.5 \times 10^{-5} & -1.419 \times 10^5 & 0 & 9.2235 \times 10^{-1} & 0 & 1.419 \times 10^5 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2.838 \times 10^5 & 2.838 \times 10^5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8.5 \times 10^{-2} & 4.7 \times 10^4 & -1.878 \times 10^5 & 1.408 \times 10^5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9.4 \times 10^4 & -9.4 \times 10^4 & 1.7 \times 10^{-1} & 0 & 0 & 0 & 0 \\ 0 & 4.7 \times 10^4 & 0 & 0 & 9.081 \times 10^{-1} & -4.7001 \times 10^4 & 8.5 \times 10^{-2} & 0 & 0 & 4.5 \times 10^{-5} \\ 0 & 0 & 0 & 0 & 0 & 1.8589 & -1.8589 & 4.5 \times 10^{-5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.8 \times 10^{-4} & -8.1935 \times 10^{-2} & 4.5 \times 10^{-5} & 8.151 \times 10^{-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.8 \times 10^{-4} & -3.8 \times 10^{-4} & 0 \\ 4.7 \times 10^4 & 0 & 0 & 0 & 0 & 0 & 0 & 8.5 \times 10^{-2} & 0 & -4.7 \times 10^4 \end{bmatrix}.$$

The probability transition matrix Π^* is calculated by the Chapman–Kolmogorov equation in Eq. (4.2.6).

$$\Pi^*(H_1^+) = \begin{bmatrix} 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \\ 0.0402 & 0.3984 & 0.4399 & 0.0014 & 0.0001 & 0.0065 & 0.0805 & 0.0172 & 0.0020 & 0.0007 \end{bmatrix},$$

$$\Pi^*(H_2^+) = \begin{bmatrix} 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \\ 0.0051 & 0.0339 & 0.5747 & 0.3282 & 0.0465 & 0.0097 & 0.0005 & 0.0001 & 0.0001 & 0.0014 \end{bmatrix},$$

$$\Pi^*(H_3^+) = \begin{bmatrix} 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \\ 0.0007 & 0.0027 & 0.1970 & 0.4906 & 0.3048 & 0.0033 & 9.0 \times 10^{-6} & 4.8 \times 10^{-6} & 5.6 \times 10^{-7} & 0.0009 \end{bmatrix}.$$

For every action $a \in A$ and state $s \in S$ the reward r_{ik} is given by the total number of molecules, i.e.,

$$r(a) = \begin{bmatrix} (A+1)(H^++2)+(C+1) \\ (B+1)+((B+1)(H^++2))+((A+1)(H^++2)) \\ (A+2)(H^++2) \\ (AH^+-1)+(AH^+-1)+(A+1)(H^++1) \\ (AH^+)+(AH^+) \\ (B+1)(H^++1)+(AH^+-1)+(AH^+-1)+(B+1) \\ (B+2)(H^++2)+(B+1) \\ (C+1)+(B+1)+((H^++2)(B+1)) \\ C+1 \\ (AH^+-1)+(AH^+-1) \end{bmatrix}, \forall a \in A$$

$$r(H_1^+) = \begin{bmatrix} 8 \\ 15 \\ 14 \\ 8 \\ 4 \\ 9 \\ 15 \\ 9 \\ 1 \\ 2 \end{bmatrix}, r(H_2^+) = \begin{bmatrix} 123 \\ 15219 \\ 244 \\ 123 \\ 4 \\ 14765 \\ 30257 \\ 15008 \\ 2 \end{bmatrix}, r(H_3^+) = \begin{bmatrix} 533 \\ 1065 \\ 1064 \\ 533 \\ 4 \\ 534 \\ 1065 \\ 534 \\ 1 \\ 2 \end{bmatrix}.$$

The functional to minimize is the expected average total number of molecules, given a set of admissible values for the pH. Solving the linear programming problem given the constraints described in Eqs.(4.3.1), (4.3.2) and (4.3.3), we obtain the following results:

$$c = \begin{bmatrix} 3.21 \times 10^{-3} & 4.65 \times 10^{-3} & 1.12 \times 10^{-14} \\ 3.20 \times 10^{-2} & 3.08 \times 10^{-2} & 7.42 \times 10^{-5} \\ 3.46 \times 10^{-2} & 5.29 \times 10^{-1} & 1.12 \times 10^{-14} \\ 4.09 \times 10^{-14} & 3.03 \times 10^{-1} & 6.14 \times 10^{-13} \\ 4.02 \times 10^{-14} & 4.29 \times 10^{-2} & 7.79 \times 10^{-12} \\ 1.54 \times 10^{-3} & 7.87 \times 10^{-3} & 2.23 \times 10^{-13} \\ 6.85 \times 10^{-3} & 4.06 \times 10^{-14} & 2.40 \times 10^{-5} \\ 1.50 \times 10^{-3} & 3.12 \times 10^{-12} & 7.22 \times 10^{-6} \\ 4.05 \times 10^{-14} & 3.49 \times 10^{-12} & 1.79 \times 10^{-4} \\ 4.01 \times 10^{-14} & 1.00 \times 10^{-3} & 3.73 \times 10^{-4} \end{bmatrix}, d = \begin{bmatrix} 0.4081 & 0.5919 & 0 \\ 0.5095 & 0.4893 & 0.0012 \\ 0.0615 & 0.9385 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0.1637 & 0.8363 & 0 \\ 0.9965 & 0 & 0.0035 \\ 0.9952 & 0 & 0.0048 \\ 0 & 0 & 1 \\ 0 & 0.7302 & 0.2698 \end{bmatrix}.$$

The Fig. 4.6, shows the CTMC generated by $Q(d)$ using a mixed strategy d , and using a pure strategy ($Q(H_1^+)$). Where the states 9 and 10 have a small number of total molecules when the CTMC is generated by $Q(d)$.

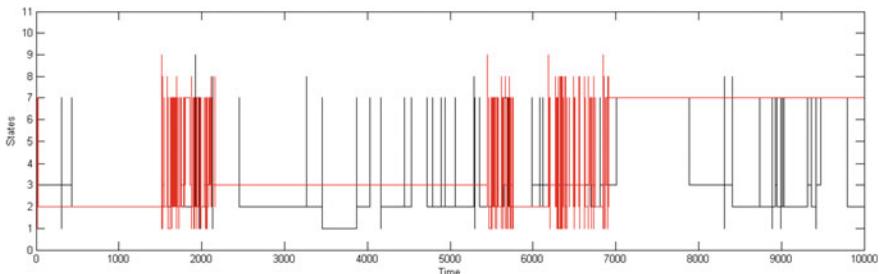


Fig. 4.6 Comparison between CTMC for the embedded chain for the state, under a pure deterministic action H_1^+ and $Q(H_1^+)$ (line black), and the CTMC for the embedded chain under the strategy d , mixed action, $Q(d)$

References

1. Aragon-Gómez, R., Clempner, J.B.: Traffic-signal control reinforcement learning approach for continuous-time markov games. *Eng. Appl. Artif. Intell.* **89**, 103415 (2020)
2. Borkar, V.S.: Topics in Controlled Markov Chains. Longman Sc & Tech (1991)
3. Buchholz, P., Schulz, I.: Numerical analysis of continuous time markov decision processes over finite horizons. *Comput. Oper. Res.* **38**(3), 651–659 (2011)
4. Carrillo, L., Escobar, J.A., Clempner, J.B., Poznyak, A.S.: Optimization problems in chemical reactions using continuous-time markov chains. *J. Math. Chem.* **54**, 1233–1254 (2016)
5. Cavazos-Cadena, R.: Recent results on conditions for the existence of average optimal stationary policies. *Ann. Oper. Res.* **28**(1–4), 3 (1991)
6. Clempner, J.B.: A continuous-time markov stackelberg security game approach for reasoning about real patrol strategies. *Int. J. Control.* **91**(11), 2494–2510 (2018)
7. Clempner, J.B.: Learning attack-defense response in continuous-time discrete-states stackelberg security markov games. *J. Exp. Theor. Artif. Intell.* (2022). <https://doi.org/10.1080/0952813X.2022.2135615>
8. Didier, F., Henzinger, T.A., Mateescu, M., Wolf, V.: Fast adaptive uniformization of the chemical master equation. In: 2009 International Workshop on High Performance Computational Systems Biology, pp. 118–127. IEEE, Trento, Italy (2009)
9. van Dijk, N.M.: Approximate uniformization for continuous-time markov chains with an application to performanceability analysis. *Stoch. Process. Appl.* **40**(2), 339–357 (1992)
10. Feinberg, E.A., Mandava, M., Shiryaev, A.N.: Sufficiency of markov policies for continuous-time markov decision processes and solutions to kolmogorov's forward equation for jump markov processes. In: 2013 IEEE 52nd Annual Conference on Decision and Control (CDC), pp. 5728–5732. IEEE (2013)
11. González, R.C., Clempner, J.B., Poznyak, A.S.: Solving traffic queues at controlled-signalized intersections in continuous-time markov games. *Math. Comput. Simul.* **166**, 283–297 (2019)
12. Goss, P.J.E., Peccoud, J.: Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. *Proc. Natl. Acad. Sci.* **95**(12), 6750–6755 (1998)
13. Grassmann, W.K.: Finding transient solutions in markovian event systems through randomization. California State University, San Bernardino, Technical report (1991)
14. Guo, X., Hernández-Lerma, O.: Continuous-time controlled markov chains. *Ann. Appl. Probab.* **13**(1), 363–388 (2003)
15. Guo, X., Hernández-Lerma, O.: Continuous-time Markov Decision Processes. Springer (2009)
16. Hill, C.G., Root, T.W.: Introduction to Chemical Engineering Kinetics and Reactor Design. Wiley (2014)
17. Janssen, J.: Book review: discrete-time markov control processes: basic optimality criteria. o. hernandez-lerma and j.-b. lasserre. Springer, Berlin (1996). xiv+216pp. dm84 (hardcover/softcover) ISBN 0-387-94579-2. *Appl. Stoch.* **12**(4), 281–282 (1996)
18. Jensen, A.: Markoff chains as an aid in the study of markoff processes. *Scand. Actuar. J.* 87–91 (1953)
19. Koeppl, H., Densmore, D., Setti, G., di Bernardo, M.: Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology. Springer Science & Business Media (2011)
20. Miguel, M.d.G.M., Formosinho, S.J.: Markov chains for plotting the course of complex reactions. *J. Chem. Educ.* **56**(9), 582 (1979)
21. Miller, B.L.: Finite state continuous time markov decision processes with an infinite planning horizon. *J. Math. Anal.* **22**(3), 552–569 (1968)
22. Nazarathy, Y., Weiss, G.: The asymptotic variance rate of the output process of finite capacity birth-death queues. *Queueing Syst.* **59**(2), 135–156 (2008)
23. Octave, L.: Chemical Reaction Engineering. Wiley (1999)
24. Paulsson, J.: Summing up the noise in gene networks. *Nature* **427**(6973), 415–418 (2004)
25. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley (2014)

26. Rao, C.V., Wolf, D.M., Arkin, A.P.: Control, exploitation and tolerance of intracellular noise. *Nature* **420**(6912), 231–237 (2002)
27. Ross, S.M.: Introduction to Stochastic Dynamic Programming. Academic (2014)
28. Sutherland, J.W., Michael, J.V.: The kinetics and thermodynamics of the reaction $\text{h} + \text{nh}_3 \text{ nh}_2 + \text{h}_2$ by the flash photolysis shock tube technique: Determination of the equilibrium constant, the rate constant for the back reaction, and the enthalpy of formation of the amidogen radical. *J. Chem. Phys.* **88**(2), 830–834 (1988)
29. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the bargaining approach for equalizing the ratios of maximal gains in continuous-time markov chains games. *Comput. Econ.* **54**, 933–955 (2019)
30. Turner, T.E., Schnell, S., Burrage, K.: Stochastic approaches for modelling in vivo reactions. *Comput. Biol. Chem.* **28**(3), 165–178 (2004)
31. Van Moorsel, A.P.A., Sanders, W.H.: Transient solution of markov models by combining adaptive and standard uniformization. *IEEE Trans. Reliab.* **46**(3), 430–440 (1997)
32. Wilkinson, D.J.: Stochastic Modelling for Systems Biology. CRC Press (2011)

Chapter 5

Nash and Stackelberg Equilibrium



Abstract We provide an approach to locating the Nash equilibrium in this chapter. The technique depends on identifying a scalar λ^* and the associated strategies $d^*(\lambda^*)$ fixing particular boundaries (*min* and *max*) that belong to the Pareto front. Bounds refer to limits placed by the player over the Pareto front that form a specific decision region where the strategies can be selected. We first use a nonlinear programming issue to illustrate the Pareto front of the game, introducing a set of linear constraints for the Markov chain game based on the *c*-variable technique. We suggest using the Euler method and a penalty function with regularization to solve the strong Nash equilibrium issue. The convergence to a single (strong) equilibrium point is ensured using Tikhonov's regularization method. The subsequent single-objective restricted problems that result from using the regularized functional of the game were then solved using a nonlinear programming technique. We use the gradient approach to resolve the first-order optimality requirements in order to accomplish the aim. The approach solves an optimization issue by adding linear constraints necessary to identify the best strong strategy, d (lambda d), starting from a *utopia point* (Pareto optimum point) given an initial *lambda* of the individual objectives. We demonstrate that the game's functional in the regularized issue decreases and ultimately converges, demonstrating the presence and exclusivity of strong Nash equilibrium (Pareto-optimal Nash equilibrium). We also provide a method for calculating the Markov chain games' strong Stackelberg/Nash equilibrium. The minimization of the L_p -norm, which shortens the distance to the utopian point in Euclidian space, is taken into consideration while solving the cooperative n -leaders and m -followers Markov game. Next, we formulate a Pareto-optimal solution to the optimization issue. For finding the strong L_p -Stackelberg/Nash equilibrium, we use a bi-level programming technique that is carried out through extraproximal optimization.

5.1 Optimization and Equilibrium

The *Nash equilibrium* [16, 18], named after the mathematician John Nash, is the most typical technique to characterize the solution of a non-cooperative game involving two or more players in game theory [24]. Each player in a Nash equilibrium is

considered to be aware of the equilibrium tactics of the other players, and altering one's own strategy will not benefit anybody.¹

The current set of strategy options represents a Nash equilibrium if each player has selected a strategy—an action plan based on what has occurred so far in the game—and no one can improve their personal expected payoff by altering their strategy while the other players remain theirs the same.

Remark 5.1 In optimization problems (with one player), the scenario is the same: if the cost function is strongly convex and we are in the minimal point, then any movement away from this point will result in the greatest number of payout values. From this viewpoint, the Nash equilibrium generalizes the conventional optimum approach to one participant situations.

Remark 5.2 The Nash equilibrium formulation assume that in the realization of the the game participants play the strategies *simultaneously*.

Several algorithms have been reported in the literature elaborated for the computation of one solution of the Nash equilibrium problems. For instance, Nabetani et al. [17] provided two different types of parametrized variational inequalities related to the generalized Nash equilibrium problem, one price-directed and the other resource-directed, showing that under mild constraints their solutions yield all the generalized Nash equilibria. Facchinei and Lampariello [14] suggested a new approach for the computation of non-variational solutions of jointly convex problems. Dreves et al. [12] presented a constrained optimization reformulation of the player convex generalized Nash equilibrium problem using a regularized Nikaido-Isoda function. Clempner and Poznyak [4] presented a natural existence of the Nash equilibrium point is ensured by definition using Lyapunov games. Gabriel et al. [15] provided a methodology to solve Nash-Cournot energy production games allowing some variables to be discrete. Facchinei et al. [13] proposed to solve a general quasi-variational inequality by using its Karush-Kuhn-Tucker conditions employing a globally convergent algorithm. Clempner [2] suggested that the stability conditions and the equilibrium point properties of Cournot and Lyapunov meet in potential games. Clempner and Poznyak [6] proposed a method for computing the strong Nash equilibrium for Markov chains games. Trejo et al. [21, 22] extended the method of Antipin [1] for computing the strong equilibrium for Stackelberg-Markov chains games. Dreves [11] considering linear generalized Nash equilibrium problems designed an algorithm that is able to compute the entire solution set of Nash equilibrium problems. Clempner and Poznyak [5, 7] showed that for Markov games the best-reply strategies lead necessarily to a Lyapunov/Nash equilibrium point. Trejo et al. [23] proposed a method for computing the Lp-strong Nash equilibrium for Markov chains games. The same authors [10] presented a procedure to construct the Pareto frontier and efficiently compute the strong Nash equilibrium for a class of discrete-time ergodic controllable Markov-chain games. Clempner [3] proposed an algorithm for computing the

¹ The concept of Nash equilibrium dates back to Cournot, who used it to explain how rival enterprises choose their outputs in 1838 [19].

Nash equilibrium based on an iterative approach of both the proximal and the gradient method.

The main results of this chapter are as follows:

- Presents the extraproximal method for computing the Nash and Stackelberg equilibria.
- Transforms the game theory problem into a system of equations, in which each equation itself is an independent optimization problem.
- Each equation in this system is an optimization problem for which the necessary condition of a minimum is solved using the projectional gradient method.
- Conceptualizes the problem as a poly-linear programming problem. This general framework captures most of the basic Markov game models.
- Restricts the solution to a class of homogeneous, finite, ergodic and controllable Markov chains.
- Regularizes the poly-linear functional employing a Lagrange regularization method for ensuring the method to converge to some of the Nash and Stackelberg equilibria.
- Provides the convergence analysis of the method and compute the rate of convergence of the step-size parameter.
- Shows that the iterated approach provides a quick rate of convergence in the numerical simulation.

5.2 ε -Nash Equilibrium and Tanaka's Function

5.2.1 Individual Cost Function

The dynamic of the game for Markov chains is described as follows. The game consists of a set of $\mathcal{N} = \{1, \dots, n\}$ players (indexed by $l = \overline{1, n}$). Below we will consider only stationary strategies $d_{k_l|i_l}^l(t) = d_{k_l|i_l}^l$. In the ergodic case we have (see Chap. 1)

$$P^l(s_{j_l}) = \sum_{i_l=1}^{N_l} \left(\sum_{k_l=1}^{M_l} \pi_{j_l|i_l k_l}^l d_{k_l|i_l}^l \right) P^l(s_{(i_l)}). \quad (5.2.1)$$

The cost function of each player, depending on the states and actions of all participants, are given by the values $v_{(i_1, k_1, j_1; \dots; i_n, k_n, j_n)}^l$, which describes the cost for the player l when the each participant $q = 1, \dots, n$ of the this game realizes the transfer from the state i_q to the state j_q after the application of the action k_q . So, the “average cost function” \mathbf{J}^l for each player l in the stationary regime can be expressed as

$$\begin{aligned}
\mathbf{J}^l(c^1, \dots, c^n) := & \sum_{i_1, k_1, j_1} \cdots \sum_{i_n, k_n, j_n} v_{(i_1, k_1, j_1; \dots; i_n, k_n, j_n)}^l \underbrace{\prod_{q=1}^n \pi_{j_q | i_q, k_q}^q d_{k_q | i_q}^q P^q(s_{i_q})}_{c_{i_q k_q}^q} = \\
& \underbrace{\sum_{i_1, k_1, j_1} \cdots \sum_{i_n, k_n, j_n} v_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l \prod_{q=1}^n \pi_{j_q | i_q, k_q}^q c_{i_q k_q}^q}_{w_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l} = \\
& \sum_{i_1, k_1, j_1} \cdots \sum_{i_n, k_n, j_n} w_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l \prod_{q=1}^n c_{i_q k_q}^q = \\
& \sum_{i_1, k_1} \cdots \sum_{i_n, k_n} \underbrace{\left(\sum_{j_1, \dots, j_n} w_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l \right)}_{\tilde{w}_{(i_1, k_1; \dots; i_n, k_n)}^l} \prod_{q=1}^n c_{i_q k_q}^q,
\end{aligned}$$

or equivalently,

$$\boxed{\mathbf{J}^l(c^1, \dots, c^n) := \sum_{i_1, k_1} \cdots \sum_{i_n, k_n} \tilde{w}_{i_1, k_1; \dots; i_n, k_n}^l \prod_{q=1}^n c_{i_q k_q}^q,} \quad (5.2.2)$$

where $c^q := \|c_{i_l k_l}^q\|_{i_l=\overline{1, N_l}; k_l=\overline{1, M_l}}$ is a matrix with elements

$$\boxed{c_{i_l k_l}^q = d_{k_l | i_l}^q P^q(s_{i_l}),} \quad (5.2.3)$$

satisfying (see Chap. 1)

$$\boxed{c^l \in C_{adm}^l = \begin{cases} c^l : \sum_{i_l, k_l} c_{i_l k_l}^l = 1, c_{i_l k_l}^l \geq 0, \\ \sum_{k_l} c_{j_l k_l}^l = \sum_{i_l, k_l} \pi_{j_l | i_l k_l}^l c_{i_l k_l}^l, \end{cases}} \quad (5.2.4)$$

and

$$\begin{aligned}
\tilde{w}_{(i_1, k_1; \dots; i_n, k_n)}^l &= \sum_{j_1, \dots, j_n} w_{(i_1, k_1, j_1; \dots; i_n, k_n, j_n)}^l, \\
w_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l &= v_{i_1, k_1, j_1; \dots; i_n, k_n, j_n}^l \prod_{q=1}^n \pi_{j_q | i_q, k_q}^q.
\end{aligned} \quad (5.2.5)$$

Remark 5.3 Notice that for a single participant case ($n = 1$) the expression (5.2.4) becomes

$$\begin{aligned}\mathbf{J}^{l=1}(c^1) &:= \sum_{i_1, k_1} \tilde{w}_{i_1, k_1}^1 c_{i_1 k_1}^1, \\ \tilde{w}_{i_1, k_1}^1 &= \sum_{j_1} w_{i_1, k_1, j_1}^1 = \sum_{j_1} v_{(i_1, k_1, j_1)}^1 \pi_{j_1 | i_1 k_1}^1,\end{aligned}$$

which coincides with the definitions in Chap. 1.

The *individual aim* of each participant is

$$\boxed{\mathbf{J}^l(c^1, \dots, c^n) \rightarrow \min_{c^l \in C_{adm}^l}.} \quad (5.2.6)$$

Obviously, here we have the *conflict situation* which can be resolved by the Nash equilibrium concept discussed above.

Remark 5.4 We have the following Nash equilibrium definitions:

A *Nash equilibrium* is a strategy $c^* = (c^{0*}, \dots, c^{n*})$ such that

$$\mathbf{J}(c^{1*}, \dots, c^{n*}) \leq \mathbf{J}(c^{1*}, \dots, c^l, \dots, c^{n*}) \text{ for } c^l \in C_{adm}.$$

A *strong Nash equilibrium* (global unique) is a strategy $c^{**} = (c^{1**}, \dots, c^{n**})$ such that there does not exist any $c^l \in C_{adm}$

$$\mathbf{J}(c^{1**}, \dots, c^l, \dots, c^{n**}) < \mathbf{J}(c^{1**}, \dots, c^{n**}).$$

5.2.2 Regularized Lagrange Function

The local optimal condition (5.2.6) can be equivalently represented, using the regularized *Lagrange function* method [8, 9], as

$$\begin{aligned}\mathcal{L}_\delta^l(c^1, \dots, c^n | \lambda^{0,l}, \lambda^{l,j_l}, l = \overline{1, n}, j_l = \overline{1, N_l}) &= \\ \mathbf{J}_\delta^l(c^1, \dots, c^n) + \sum_{l=1}^n \lambda^{0,l} \left(\sum_{i_l, k_l} c_{i_l k_l}^l - 1 \right) + \\ \sum_{l=1}^n \sum_{j_l} \lambda^{l, j_l} \left(\sum_{k_l} c_{j_l k_l}^l - \sum_{i_l, k_l} \pi_{j_l | i_l k_l}^l c_{i_l k_l}^l \right) - \frac{\delta}{2} \sum_{l=1}^n \sum_{j_l} (\lambda^{l, j_l})^2 &\rightarrow \min_{c_{i_l k_l}^l \geq 0, \lambda^{0,l}, \lambda^{l, j_l}},\end{aligned} \quad (5.2.7)$$

where

$$\mathbf{J}_\delta^l(c^1, \dots, c^n) = \mathbf{J}^l(c^1, \dots, c^n) + \frac{\delta}{2} \sum_{l=1}^n \left(\sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} (c_{i_l k_l}^l)^2 \right), \delta > 0 \quad (5.2.8)$$

and the regularized Lagrange function \mathcal{L}_δ^l includes only the equality constraints from C_{adm}^l .

5.2.3 Tanaka's Function

To simplify the description let us introduce below the new variables

$$\left. \begin{aligned} x^l &:= \text{col } c^{(l)} = \left(c_{11}^{(l)}, \dots, c_{1M_l}^{(l)}; \dots; c_{N_l 1}^{(l)}, \dots, c_{N_l M_l}^{(l)} \right)^\top, \\ X^l &:= X_+^l = \{x_1^l \geq 0, \dots, x_{N_l M_l}^l \geq 0\}. \end{aligned} \right\} \quad (5.2.9)$$

Consider a *multi-players game* with the joint strategy

$$x = (x^1, \dots, x^n) \in X := \bigotimes_{l=1}^n X^l.$$

The players are trying to reach one of ε -Nash equilibria, that is, find a joint strategy $x \in X$ satisfying for any admissible $x^l \in X^l$ and any $l = \overline{1, n}$ the system of inequalities (*the ε -Nash equilibrium condition*)

$$\boxed{\begin{aligned} g_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l}) &\leq \varepsilon \text{ for any } y^l \in X^l \text{ and all } l = \overline{1, n}, j_l = \overline{1, N_l}, \varepsilon \geq 0, \\ g_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l}) &:= \mathcal{L}_\delta^l(x | \lambda^{0,l}, \lambda^{l,j_l}) - \mathcal{L}_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l}), \end{aligned}} \quad (5.2.10)$$

where $x^{\hat{l}}$ is a strategy of the rest of the followers adjoint to x^l , namely,

$$x^{\hat{l}} := (x^1, \dots, x^{l-1}, x^{l+1}, \dots, x^n) \in X^{\hat{l}} := \bigotimes_{m=1, m \neq l}^n X^m,$$

Here $\mathcal{L}_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l})$ is the Lagrange function of the player l which plays the strategy $y^l \in X^l$ and the rest of the players the strategy $x^{\hat{l}} \in X^{\hat{l}}$.

Lemma 5.1 ([20]) *The joint strategy $x \in X$ (5.2.10) is an ε -Nash equilibrium if and only if*

$$\boxed{\begin{aligned} G_\delta(x, y | \lambda) &:= \sum_{l=1}^n g_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l}) = \\ &\sum_{l=1}^n [\mathcal{L}_\delta^l(x | \lambda^{0,l}, \lambda^{l,j_l}) - \mathcal{L}_\delta^l(y^l, x^{\hat{l}} | \lambda^{0,l}, \lambda^{l,j_l})] \leq \varepsilon \end{aligned}} \quad (5.2.11)$$

for all $y \in X$, where $\lambda = [\lambda^{0,l}, \lambda^{l,j_l}]_{l=1,n, j_l=1,N_l}$. The function $G_\delta(x, y|\lambda)$ is referred to as a **regularized Tanaka's function**, which for $\delta = 0$ and $\lambda = 0$ is the original Tanaka's function [20].

Proof Summing (5.2.10) implies (5.2.11). And inverse, taking $y^m = x^m$ for all $m \neq l$ in (5.2.11), which is valid for any admissible y^l , we obtain (5.2.10). This means that x belongs to the ε -Nash equilibrium set. \square

Notice that the condition $G_\delta(x, y) \leq \varepsilon$ (5.2.11), is equivalent to

$$\boxed{\max_{\lambda^{0,l}, \lambda^{l,j_l}} \max_{y \in X} G_\delta(x, y|\lambda^{0,l}, \lambda^{l,j_l}) \leq \varepsilon.} \quad (5.2.12)$$

So, any strategy $x \in X$ is a Nash equilibrium if it satisfies (5.2.12).

Remark 5.5 The considered multiparicipant game has at least one ε -Nash equilibrium if

$$\boxed{\min_{x \in X} \left(\max_{\lambda^{0,l}, \lambda^{l,j_l}} \max_{y \in X} G_\delta(x, y|\lambda^{0,l}, \lambda^{l,j_l}) \right) \leq \varepsilon.} \quad (5.2.13)$$

5.2.4 ASG Continuous-Time Algorithm

Let us consider the following continuous-time procedure

$$\boxed{\begin{aligned} \dot{x}(t) &= (t + \theta)^{-1} [\zeta_x(t) - x(t) - \eta_x C], \\ \hat{x}(t) &:= x(t_0) + \int_{\tau=t_0}^t \dot{x}(\tau) d\tau, \\ x(t) &= [\hat{x}(t)]_+, x(t_0) \in X, \frac{d}{dt} \hat{x}(t) = \dot{x}(t), \\ \dot{\zeta}_x(t) &= -\partial_x G_\delta(x, y|\lambda^{0,l}, \lambda^{l,j_l}), \\ \partial_x G_\delta(x, y|\lambda^{0,l}, \lambda^{l,j_l}) &\in D_x(x, y|\lambda^{0,l}, \lambda^{l,j_l}), \quad \zeta_x(t_0) = 0, \end{aligned}} \quad (5.2.14)$$

and

$$\left. \begin{aligned} \dot{\hat{y}}(t) &= (t + \theta)^{-1} [\zeta_y(t) + y(t) + \eta_y], \\ \hat{y}(t) &:= y(t_0) + \int_{\tau=t_0}^t \dot{\hat{y}}(\tau) d\tau, X = Y, \\ y(t) &= [\hat{y}(t)]_+, y(t_0) \in Y, \frac{d}{dt} \hat{y}(t) = \dot{\hat{y}}(t) \\ \dot{\zeta}_y(t) &= \partial_y G_\delta(x, y | \lambda^{0,l}, \lambda^{l,j_i}), \\ \partial_y G_\delta(x, y | \lambda^{0,l}, \lambda^{l,j_i}) &\in D_y(x, y | \lambda^{0,l}, \lambda^{l,j_i}), \quad \zeta_y(t_0) = 0, \end{aligned} \right\} \quad (5.2.15)$$

where the operator $[\cdot]_+$ makes each component equal to zero if the corresponding argument is negative, $D_x(x, y | \lambda^{0,l}, \lambda^{l,j_i})$ and $D_y(x, y | \lambda^{0,l}, \lambda^{l,j_i})$ are the sets of the subgradients in the corresponding points, and the Lagrange multipliers $\lambda^{0,l}$ and λ^{l,j_i} evolve according to the following differential rule:

$$\left. \begin{aligned} \frac{d}{dt} \lambda^{0,l} &= -\gamma^{0,l} \partial_{\lambda^{0,l}} G_\delta(x, y | \lambda^{0,l}, \lambda^{l,j_i}), \gamma^{0,l} > 0, \\ \frac{d}{dt} \lambda^{l,j_i} &= -\gamma^{l,j_i} \partial_{\lambda^{l,j_i}} G_\delta(x, y | \lambda^{0,l}, \lambda^{l,j_i}), \gamma^{l,j_i} > 0 \end{aligned} \right\} \quad (5.2.16)$$

with any fixed initial conditions.

Theorem 5.1 (On the functional convergence of the AGS algorithm) *Under the conditions of the Tanaka's function convexity, which results from the polylinearity property (5.2.2) of $\mathbf{J}^l(c^1, \dots, c^n)$ on C_{adm}^l , and in view of the Lemma 8.2, for any $\varepsilon > 0$ exists $\theta \geq \frac{c}{\varepsilon}$ and any constant vectors $\eta_x, \eta_y \in \mathbb{R}^N$, the ASG strategies (5.2.14), (5.2.15) guarantee the following property for any $t \geq t_0$*

$$\max_{y \in X} G_\delta(\hat{x}(t), y) \leq \varepsilon \quad (5.2.17)$$

where

$$c := \frac{1}{2} \|x^* + \eta_x\|^2 + \frac{1}{2} \|y^* + \eta_y\|^2 + (t_0 + \theta) |G_\delta(x, y)|_{t_0} \quad (5.2.18)$$

and x^*, y^* defined by

$$\left. \begin{aligned} x^* &= \arg \min_{x \in X} \max_{y \in X} G_\delta(x, y), \\ y^* &= \arg \max_{y \in X} \min_{x \in X} G_\delta(x, y). \end{aligned} \right\} \quad (5.2.19)$$

Remark 5.6 Notice that the points x^* and y^* satisfy the **saddle point** condition

$$G_\delta(x^*, y) \leq G_\delta(x^*, y^*) \leq G_\delta(x, y^*) \quad (5.2.20)$$

for all $x, y \in X$.

As it follows from the inequality (5.2.17), the trajectories $\hat{x}(t)$, $t \geq t_0$, belong to the set of ε -Nash equilibrium points. But we are interesting in the trajectories $x(t)$, $t \geq t_0$. What can we say about this?

The next Lemma responds this questions.

5.3 Extraproximal Method

The *Extraproximal Method* for the conditional optimization problems was suggested in [1]. We describe below this method for the Stackelberg-Nash game in a general format.

5.3.1 Proximal Format

The problem

$$\min_{x \in X} \left(\max_{\lambda \geq 0} \max_{y \in X} G_\delta(x, y|\lambda) \right) \quad (5.3.1)$$

can be expressed in proximal format as (see, [1, 21])

$$\boxed{\begin{aligned} \lambda_\delta^* &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_\delta^*\|^2 + \gamma G_\delta(x_\delta^*, y_\delta^*|\lambda) \right\}, \\ x_\delta^* &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_\delta^*\|^2 + \gamma G_\delta(x, y_\delta^*|\lambda_\delta^*) \right\}, \\ y_\delta^* &= \arg \max_{y \in X} \left\{ -\frac{1}{2} \|y - y_\delta^*\|^2 + \gamma G_\delta(x_\delta^*, y|\lambda_\delta^*) \right\}, \end{aligned}} \quad (5.3.2)$$

where the solutions x_δ^* , y_δ^* and λ_δ^* depend on the small parameters $\delta, \gamma > 0$.

The general format iterative version ($n = 0, 1, \dots$) of the extraproximal method for computing the Nash equilibrium with some fixed admissible initial values ($x_0, y_0 \in X$ and $\lambda_0 \geq 0$) is as follows:

1. The *first half-step* (prediction):

$$\begin{aligned}\bar{\lambda}_n &= \arg \max_{\lambda \geq 0} \left\{ \frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma G_\delta(x_n, y_n | \lambda) \right\}, \\ \bar{x}_n &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_n\|^2 + \gamma G_\delta(x, y_n | \lambda_n) \right\}, \\ \bar{y}_n &= \arg \max_{y \in X} \left\{ -\frac{1}{2} \|y - y_n\|^2 + \gamma G_\delta(x_n, y | \lambda_n) \right\}.\end{aligned}\quad (5.3.3)$$

2. The *second (basic) half-step*

$$\begin{aligned}\lambda_{n+1} &= \arg \max_{\lambda \geq 0} \left\{ \frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma G_\delta(\bar{x}_n, \bar{y}_n | \lambda) \right\}, \\ x_{n+1} &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_n\|^2 + \gamma G_\delta(x, \bar{y}_n | \bar{\lambda}_n) \right\}, \\ y_{n+1} &= \arg \max_{y \in X} \left\{ -\frac{1}{2} \|y - y_n\|^2 + \gamma G_\delta(\bar{x}_n, y | \bar{\lambda}_n) \right\}.\end{aligned}\quad (5.3.4)$$

Remark 5.7 The direct one-step procedure, when $\lambda_{n+1} = \bar{\lambda}_n$, does not work (see the counter-example in [1]).

5.4 Numerical Example: Strong Nash Equilibrium in Pareto Front

5.4.1 Euler Approach

Let us define the regularized penalty function as follows

$$\left. \begin{aligned}\Phi_\delta(\lambda) &:= \Phi(\lambda) + \frac{\delta}{2} \|\lambda\|^2, 0 < \delta \ll 1, \\ \Phi(\lambda) &:= \sum_l^n \Phi^l(\lambda), \\ \Phi_l(\lambda) &:= \frac{1}{2} \left[J^l(v^*(\lambda)) - J^{l+} \right]_+^2 + \frac{1}{2} \left[J^{l-} - J^l(v^*(\lambda)) \right]_+^2,\end{aligned}\right\} \quad (5.4.1)$$

where

$$[\zeta]_+ := \begin{cases} \zeta & \text{if } \zeta \geq 0, \\ 0 & \text{if } \zeta < 0. \end{cases} \quad (5.4.2)$$

The minimizing solution λ^* can be expressed mathematically as follows:

$$\lambda^* \in \arg \min_{\lambda \in [0,1]} \Phi(\lambda) = \arg \min_{\lambda \in [0,1]} \Phi_{\delta=0}(\lambda). \quad (5.4.3)$$

As it has been mentioned above λ^* , satisfying (5.4.3), may be not unique that provokes several problems for the numerical procedure implementation. Taking $\delta > 0$ we can guarantee the uniqueness of the solution so that we will try to find

$$\lambda_\delta^{**} = \arg \min_{\lambda \in \mathcal{S}^N} \Phi_\delta(\lambda), \delta > 0. \quad (5.4.4)$$

This solution corresponds to one of the solutions λ^* (5.4.3) which has the minimal norm of the vector $\|\mathbf{v}\|^2$.

To find λ_δ^{**} let us apply the following numerical procedure

$$\lambda_{n+1} = \text{Pr}_{\mathcal{S}^N} \left[\lambda_n - \gamma_n \frac{\Phi'_\delta(\lambda_n)}{|\Phi'_\delta(\lambda_n)| + \varepsilon} \right],$$

$$\lambda_0 = (1/n, \dots, 1/n), n = 0, 1, \dots$$

$$\gamma_n > 0, \sum_{n=0}^{\infty} \gamma_n = \infty,$$

where Pr is the projection operator into the simplex \mathcal{S}^N . Notice that the derivative $\Phi'_\delta(\lambda_n)$ is

$$\begin{aligned} \Phi'_\delta(\lambda_n) &= \frac{d}{d\lambda} \Phi_\delta(\lambda_n) = \delta(2\lambda_n - 1) + \\ &\sum_{l=1}^n \frac{d}{d\lambda} J^l(v^*(\lambda_n)) \left([J^l(v^*(\lambda_n)) - J^{l+}]_+ + [J^{l-} - J^l(v^*(\lambda_n))]_+ \right), \end{aligned}$$

where the terms $\frac{d}{d\lambda} J^l(v^*(\lambda_n))$ may be approximated by the Euler method as

$$\frac{d}{d\lambda} J^l(v^*(\lambda_n)) \simeq \varepsilon^{-1} [J^l(v^*(\lambda_n + \varepsilon)) - J^l(v^*(\lambda_n))], 0 < \varepsilon \ll 1.$$

Finally, the suggested numerical procedure with $\gamma_n = \gamma$ for finding λ_δ^{**} looks as follows

$$\left. \begin{aligned}
\check{\gamma}_{n+1} &= \Pr_{S^N} \left\{ \check{\gamma}_n - \gamma_n \nabla \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n) \right\}, \\
\check{\gamma}_0 &= (1/n, \dots, 1/n), n = 0, 1, \dots, \\
\nabla \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n) &:= \left(\frac{\partial}{\partial \lambda_1} \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n), \frac{\partial}{\partial \lambda_2} \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n), \dots, \frac{\partial}{\partial \lambda_N} \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n) \right)^T, \\
\frac{\partial}{\partial \lambda_i} \tilde{\Phi}_{\delta, \varepsilon} (\check{\gamma}_n) &:= \text{where} \\
\varepsilon^{-1} \sum_{l=1}^n [\Delta_l J^l (\check{\gamma}_n, \varepsilon)] &\left([J^l (v^*(\check{\gamma}_n)) - J^{l+}]_+ - [J^{l-} - J^l (v^*(\check{\gamma}_n))]_+ \right) + \frac{\delta}{2} \frac{\partial}{\partial \lambda_i} \|\check{\gamma}\|^2, \\
\text{with } \Delta_l J^l (\check{\gamma}_n, \varepsilon) &:= [J^l (v^*(\check{\gamma}_n + \varepsilon e_i)) - J^l (v^*(\check{\gamma}_n))], \\
\text{and } e_i &:= \left(\underbrace{0, \dots, 0}_i, 1, \dots, 0 \right)^T.
\end{aligned} \right\} \quad (5.4.5)$$

Theorem 5.2 (Uniqueness of the SNE) *Let $\Phi_\delta(\lambda)$ be a continuous and strictly concave function over the Pareto front \mathcal{P} . Then, the Markov chains game has one strong Nash equilibrium.*

Proof The strategy v_δ^* is a strong Pareto policy because every point on the Pareto front $\mathbf{J}(\mathcal{P})$ is regularized and satisfies that $\mathbf{J}(v_\delta) < \mathbf{J}(v)$, specifically $\mathbf{J}(v_\delta^*) < \mathbf{J}(v)$. The continuity of function $\Phi(\lambda)$ is given by its definition and the construction of the Pareto front \mathcal{P} established by

$$\{v_\delta^*, \lambda_\delta^*, \mu_\delta^*, \theta_\delta^*\} = \arg \min_{\lambda \in \mathcal{S}^N, v \geq 0, \mu} \max_{\theta \geq 0} \{\mathcal{L}_{\delta, \varrho}(v, \lambda, \mu, \theta)\}.$$

The functional $\Phi(\lambda_n)$ is monotonically non-increasing and bounded from below then, by the Weierstrass theorem it converges, i.e. $\lim_{n \rightarrow \infty} \Phi(\lambda_n) \rightarrow \Phi(\lambda)$. The constraints (5.2.4) are linear and therefore they are convex. Since the intersection of convex sets is convex, $V = C_{adm}$ is convex. The strict convexity of the Pareto front $\mathbf{J}(\mathcal{P})$ is ensured by the introduction of the Tikhonov's regularizer $\frac{\delta}{2} \|\lambda\|^2$ over $\Phi(\lambda)$. Then, by convexity $\Phi_\delta(\lambda)$ converges to a unique equilibrium point. \square

5.4.2 Numerical Data

Let us consider a game with two players ($l = 1, 2$) trying to conform a coalition. Let $N_1 = N_2 = 6$ and $M_1 = M_2 = 2$. The transition matrices of the example are as follows:

$$\pi_{j|i1}^1 = \begin{bmatrix} 0.1927 & 0.0659 & 0.2264 & 0.1874 & 0.1606 & 0.1670 \\ 0.0332 & 0.1716 & 0.1523 & 0.3011 & 0.2377 & 0.1041 \\ 0.0357 & 0.2689 & 0.2248 & 0.1842 & 0.2087 & 0.0778 \\ 0.3662 & 0.3868 & 0.0569 & 0.0143 & 0.1572 & 0.0185 \\ 0.2248 & 0.0560 & 0.1499 & 0.3018 & 0.2330 & 0.0345 \\ 0.7468 & 0.0732 & 0.0712 & 0.0719 & 0.0064 & 0.0306 \end{bmatrix},$$

$$\pi_{j|i2}^1 = \begin{bmatrix} 0.1937 & 0.2134 & 0.1978 & 0.0332 & 0.2094 & 0.1525 \\ 0.1469 & 0.3683 & 0.0717 & 0.2308 & 0.1182 & 0.0642 \\ 0.3138 & 0.0617 & 0.0911 & 0.3169 & 0.1671 & 0.0493 \\ 0.0138 & 0.1958 & 0.2718 & 0.1361 & 0.2795 & 0.1030 \\ 0.1138 & 0.1156 & 0.1699 & 0.1518 & 0.2310 & 0.2180 \\ 0.4672 & 0.0377 & 0.0037 & 0.0051 & 0.0220 & 0.4643 \end{bmatrix},$$

$$\pi_{j|i1}^2 = \begin{bmatrix} 0.3819 & 0.4364 & 0.0046 & 0.0205 & 0.1536 & 0.0030 \\ 0.0824 & 0.2956 & 0.0967 & 0.0183 & 0.3281 & 0.1789 \\ 0.3265 & 0.1587 & 0.2660 & 0.1865 & 0.0042 & 0.0582 \\ 0.0958 & 0.2274 & 0.2063 & 0.2133 & 0.0923 & 0.1648 \\ 0.0780 & 0.2321 & 0.1509 & 0.3704 & 0.0643 & 0.1043 \\ 0.0468 & 0.0022 & 0.0030 & 0.0058 & 0.0354 & 0.9068 \end{bmatrix},$$

$$\pi_{j|i2}^2 = \begin{bmatrix} 0.0478 & 0.0817 & 0.0572 & 0.4207 & 0.0975 & 0.2949 \\ 0.1998 & 0.2205 & 0.2569 & 0.0801 & 0.2040 & 0.0387 \\ 0.1916 & 0.2289 & 0.0445 & 0.1105 & 0.0619 & 0.3627 \\ 0.0248 & 0.2845 & 0.2294 & 0.2369 & 0.0403 & 0.1842 \\ 0.0825 & 0.0282 & 0.2944 & 0.1554 & 0.3131 & 0.1264 \\ 0.8611 & 0.0381 & 0.0059 & 0.0009 & 0.0499 & 0.0441 \end{bmatrix},$$

The individual cost matrices are as follows

$$J_{ij,1}^1 = \begin{bmatrix} 2 & 2 & 10 & 2 & 8 & 6 \\ 2 & 6 & 10 & 0 & 6 & 8 \\ 4 & 16 & 10 & 14 & 0 & 10 \\ 4 & 2 & 8 & 8 & 12 & 6 \\ 14 & 80 & 12 & 18 & 14 & 18 \\ 4 & 16 & 6 & 12 & 6 & 40 \end{bmatrix}, J_{ij,2}^1 = \begin{bmatrix} 4 & 0 & 6 & 6 & 0 & 10 \\ 4 & 2 & 10 & 4 & 6 & 2 \\ 6 & 2 & 12 & 6 & 6 & 6 \\ 10 & 0 & 6 & 8 & 2 & 10 \\ 8 & 10 & 14 & 10 & 20 & 60 \\ 80 & 4 & 16 & 12 & 16 & 12 \end{bmatrix},$$

$$J_{ij,1}^2 = \begin{bmatrix} 91 & 68 & 59 & 31 & 19 & 20 \\ 3 & 5 & 2 & 4 & 3 & 7 \\ 2 & 7 & 9 & 3 & 1 & 2 \\ 2 & 2 & 3 & 5 & 0 & 16 \\ 1 & 0 & 8 & 29 & 6 & 4 \\ 4 & 7 & 4 & 83 & 58 & 83 \end{bmatrix}, J_{ij,2}^2 = \begin{bmatrix} 142 & 188 & 60 & 20 & 180 & 10 \\ 2 & 6 & 16 & 14 & 8 & 2 \\ 8 & 10 & 8 & 4 & 10 & 0 \\ 2 & 10 & 20 & 6 & 14 & 8 \\ 6 & 2 & 0 & 8 & 10 & 4 \\ 116 & 140 & 12 & 16 & 32 & 6 \end{bmatrix}.$$

Fig. 5.1 Pareto min and max bounds

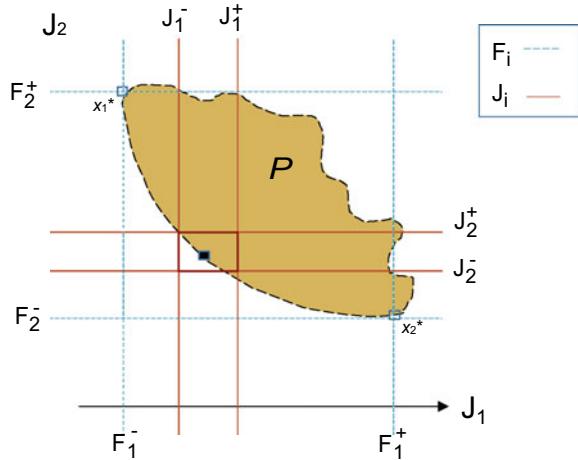
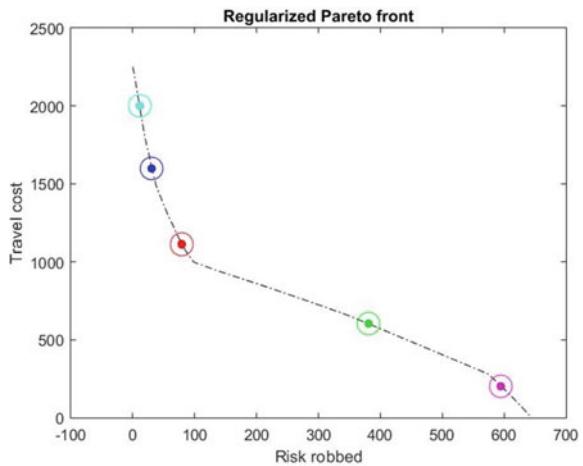


Fig. 5.2 Pareto front



Fixing $\delta = 0.0001$, $\gamma = 0.6$, $\varrho = 0.03$, $n_0 = 100$ and the bounds $J1^- = 183$, $J1^+ = 260$, $J2 = 3300$, $J2 = 3870$ (see Fig. 5.1). Then, Fig. 5.2 shows the 2000 points of the Pareto front, Fig. 5.3 shows the Strong Nash equilibrium, Fig. 5.4 shows the convergence of the parameter $\lambda_{final} = 0.4994$ and finally, Fig. 5.5 shows the convergence of the gradient.

The corresponding strong Pareto policies are as follows

Fig. 5.3 Strong Nash equilibrium

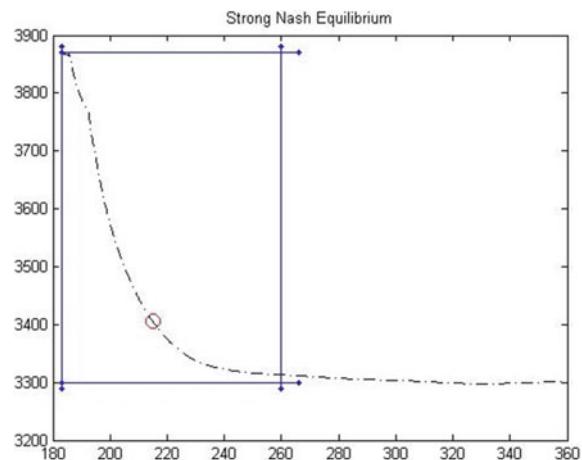
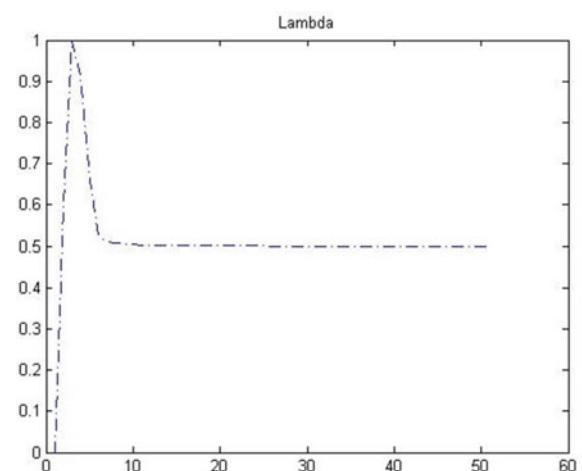
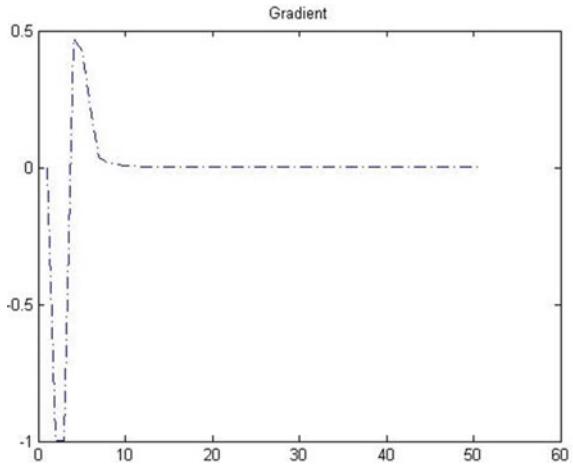


Fig. 5.4 Convergence of Lambda



$$c_{ik}^{1*} = \begin{bmatrix} 0.1092 & 0.1088 \\ 0.0889 & 0.0869 \\ 0.0815 & 0.0744 \\ 0.0814 & 0.0895 \\ 0.0915 & 0.0904 \\ 0.0946 & 0.0029 \end{bmatrix}, \quad d_{ik}^{1*} = \begin{bmatrix} 0.5008 & 0.4992 \\ 0.5057 & 0.4943 \\ 0.5226 & 0.4774 \\ 0.4764 & 0.5236 \\ 0.5029 & 0.4971 \\ 0.9705 & 0.0295 \end{bmatrix},$$

Fig. 5.5 Convergence of the Gradient



$$c_{ik}^{2*} = \begin{bmatrix} 0.0006 & 0.2328 \\ 0.0998 & 0.0549 \\ 0.0007 & 0.1137 \\ 0.0911 & 0.0968 \\ 0.0435 & 0.0796 \\ 0.0011 & 0.1854 \end{bmatrix}, \quad d_{ik}^{2*} = \begin{bmatrix} 0.0028 & 0.9972 \\ 0.6453 & 0.3547 \\ 0.0061 & 0.9939 \\ 0.4847 & 0.5153 \\ 0.3533 & 0.6467 \\ 0.0059 & 0.9941 \end{bmatrix}.$$

and the utility $J(d^*) = (214.9335, 3.4054e + 03)$.

5.5 The Stackelberg-Nash Equilibrium Concept

5.5.1 Specific Features

The *Stackelberg model* is a game in which the leader moves first and then the follower moves *sequentially* (opposite to Remark 5.2 where all players play simultaneously).

Remark 5.8 In our approach we consider one leader and several followers that play themselves a Nash game, and it is because we called this type of games Stackelberg-Nash (sequential) games.

To simplify the descriptions below let us introduce the new variables

$$\left. \begin{aligned} x &:= \text{col } c^{(0)}, X := C_{adm}^{(0)}, \\ v^l &:= \text{col } c^{(l)}, V^l := C_{adm}^{(l)} (l = \overline{1, n}) \end{aligned} \right\} \quad (5.5.1)$$

Consider a *non-zero sum game with a leader* whose strategies are denoted by $x \in X$ and n followers with strategies $v^l \in V^l$ ($l = \overline{1, n}$). Denote by $v = (v^1, \dots, v^n) \in V := \bigotimes_{l=1}^n V^l$ the joint strategy of the followers.

Let us introduce $v^{\hat{l}}$ is a strategy of the rest of the followers adjoint to v^l , namely,

$$v^{\hat{l}} := (v^1, \dots, v^{l-1}, v^{l+1}, \dots, v^n) \in V^{\hat{l}} := \bigotimes_{m=1, m \neq l}^n V^m,$$

so that $v = (v^l, v^{\hat{l}})$ for any l .

The leader is assumed to anticipate the reactions of the followers. They are trying to reach one of Nash equilibria for any fixed strategy x of the leader.

5.5.2 Individual Aims and Tanaka's Representation

Denote by $J^0(x|v)$ and $J^l(v|x(v))$, $l = 1, \dots, n$ the cost functions for all the participants.

- **Individual aim for the leader:**

$$J^0(x|v) \rightarrow \min_{x \in X} \quad (5.5.2)$$

with $x(v) \in \arg \min_{x' \in X} J^0(x'|v)$ as a solution.

- **Individual aim for the followers** is to find $v^{l*} \in V^l$ such that

$$g^l(v^{l*}, v|x(v^{l*}, v^{\hat{l}})) := J^l(v^{l*}, v^{\hat{l}}|x(v^{l*}, v^{\hat{l}})) - J^l(v^l, v^{\hat{l}}|x(v^l, v^{\hat{l}})) \leq \varepsilon, \varepsilon \geq 0, \quad (5.5.3)$$

valid for all $v^l \in V^l$, $v^{\hat{l}} \in V^{\hat{l}}$ and all $l = \overline{1, n}$. Below v^{l*} will be associated with the ε -Nash equilibrium in Stackelberg-Nash game.

The definition of individual aim for the followers can be represented in joint format, using the regularized Tanaka's function description, as follows:

$$G_\delta(v^*, v|\lambda, x(v^*, v)) := \left\{ \sum_{l=1}^n g^l_\delta(v^{l*}, v|\lambda^{0,l}, \lambda^{l,j_l}, x(v^{l*}, v^{\hat{l}})) \leq \varepsilon \right\} \quad (5.5.4)$$

for all $v \in V$, $\lambda = [\lambda^{0,l}, \lambda^{l,j_l}]_{l=\overline{1,n}, j_l=\overline{1, N_l}}$ and

$$g^l_\delta(v^{l*}, v|\lambda^{0,l}, \lambda^{l,j_l}, x(v^{l*}, v^{\hat{l}})) = \mathcal{L}_\delta^l(v^{l*}|\lambda^{0,l}, \lambda^{l,j_l}, x(v^{l*}, v^{\hat{l}})) - \mathcal{L}_\delta^l(v|\lambda^{0,l}, \lambda^{l,j_l}, x(v^l, v^{\hat{l}})) \quad (5.5.5)$$

where $\mathcal{L}_\delta^l(v|\lambda^{0,l}, \lambda^{l,j_l}, x(v^l, v^{\hat{l}}))$ corresponds with the Lagrange function in (5.2.7).

The function $G_\delta(v^*, v|\lambda, x(v^*, v))$ is referred to as a **regularized Tanaka's function** for Stackelberg-Nash game, which for $\delta = 0$ and $\lambda = 0$ is the original Tanaka's function [20].

Definition 5.1 A strategy $x^* \in X$ of the leader together with the collection $v^* \in V$ of the followers is said to be a **Stackelberg-Nash equilibrium** ($\varepsilon = 0$) if

$$(x^*, v^*) \in \arg \max_{\lambda \geq 0} \min_{v \in V} \min_{x' \in X} \{J^0(x'|v) | G_\delta(v, \hat{v}|\lambda, x(v, \hat{v})) \leq 0\}. \quad (5.5.6)$$

Let us use the Lagrange multipliers approach we can represent (5.5.6) as follows:

$$\mathbb{L}_\delta(x', v, \hat{v}, \mu, \lambda) \rightarrow \max_{\mu, \lambda \geq 0} \min_{v \in V} \min_{x' \in X}, \quad (5.5.7)$$

where

$$\mathbb{L}_\delta(x', v, \hat{v}, \mu, \lambda) = J^0(x'|v) + \mu G_\delta(v, \hat{v}|\lambda, x(v, \hat{v})). \quad (5.5.8)$$

With $\delta > 0$ the considered functions becomes to be strictly convex providing the uniqueness of the considered conditional optimization problem (5.5.7). Notice also that the Lagrange function in (5.5.8) satisfies the saddle-point condition, namely, for all $x' \in X$ and $\lambda \geq 0$ we have

$$\mathbb{L}_\delta(x'^*, v^*, \hat{v}^*, \mu, \lambda) \leq \mathbb{L}_\delta(x'^*, v^*, \hat{v}^*, \mu^*, \lambda^*) \leq \mathbb{L}_\delta(x', v, \hat{v}, \mu^*, \lambda^*). \quad (5.5.9)$$

5.5.3 Extraproximal Procedure

The problem (5.5.7) can be represented in the following proximal format as (see, [1, 21])

$$\left. \begin{aligned} \lambda^* &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda^*\|^2 + \gamma \mathbb{L}_\delta(x'^*, v^*, \hat{v}^*, \mu^*, \lambda) \right\}, \\ \mu^* &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu^*\|^2 + \gamma \mathbb{L}_\delta(x'^*, v^*, \hat{v}^*, \mu, \lambda^*) \right\}, \\ x'^* &= \arg \min_{x' \in X} \left\{ \frac{1}{2} \|x' - x'^*\|^2 + \gamma \mathbb{L}_\delta(x', v^*, \hat{v}^*, \mu^*, \lambda^*) \right\}, \\ v^* &= \arg \min_{v \in V} \left\{ \frac{1}{2} \|v - v^*\|^2 + \gamma \mathbb{L}_\delta(x'^*, v, \hat{v}^*, \mu^*, \lambda^*) \right\}, \\ \hat{v}^* &= \arg \min_{\hat{v}^* \in V} \left\{ \frac{1}{2} \|\hat{v} - \hat{v}^*\|^2 + \gamma \mathbb{L}_\delta(x'^*, v^*, \hat{v}, \mu^*, \lambda^*) \right\}, \end{aligned} \right\} \quad (5.5.10)$$

where the solutions x'^* , v^* , \hat{v}^* , μ^* , and λ^* depend on the small parameter $\gamma > 0$.

The general format iterative version ($n = 0, 1, \dots$) of the extraproximal method for computing the Stackelberg-Nash equilibrium with some fixed admissible initial values ($x'_0 \in X$, $v_0, \bar{v}_0 \in V$ and $\mu_0, \lambda_0 \geq 0$) is as follows:

1. The *first half-step* (prediction):

$$\left. \begin{aligned} \bar{\lambda}_n &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathbb{L}_\delta(x'_n, v_n, \hat{v}_n, \mu_n, \lambda) \right\}, \\ \bar{\mu}_n &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 + \gamma \mathbb{L}_\delta(x'_n, v_n, \hat{v}_n, \mu, \lambda_n) \right\}, \\ \bar{x}'_n &= \arg \min_{x' \in X} \left\{ \frac{1}{2} \|x' - x'_n\|^2 + \gamma \mathbb{L}_\delta(x', v_n, \hat{v}_n, \mu_n, \lambda_n) \right\}, \\ \bar{v}_n &= \arg \min_{v \in V} \left\{ \frac{1}{2} \|v - v_n\|^2 + \gamma \mathbb{L}_\delta(x'_n, v, \hat{v}_n, \mu_n, \lambda_n) \right\}, \\ \bar{\hat{v}}_n &= \arg \min_{\hat{v}^* \in V} \left\{ \frac{1}{2} \|\hat{v} - \hat{v}_n\|^2 + \gamma \mathbb{L}_\delta(x'_n, v_n, \hat{v}, \mu_n, \lambda_n) \right\}. \end{aligned} \right\} \quad (5.5.11)$$

2. The *second (basic) half-step*

$$\left. \begin{aligned} \lambda_{n+1} &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathbb{L}_\delta(\bar{x}'_n, \bar{v}_n, \bar{\hat{v}}_n, \bar{\mu}_n, \lambda) \right\}, \\ \mu_{n+1} &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 + \gamma \mathbb{L}_\delta(\bar{x}'_n, \bar{v}_n, \bar{\hat{v}}_n, \mu, \bar{\lambda}_n) \right\}, \\ x'_{n+1} &= \arg \min_{x' \in X} \left\{ \frac{1}{2} \|x' - x'_n\|^2 + \gamma \mathbb{L}_\delta(x', \bar{v}_n, \bar{\hat{v}}_n, \bar{\mu}_n, \bar{\lambda}_n) \right\}, \\ v_{n+1} &= \arg \min_{v \in V} \left\{ \frac{1}{2} \|v - v_n\|^2 + \gamma \mathbb{L}_\delta(\bar{x}'_n, v, \bar{\hat{v}}_n, \bar{\mu}_n, \bar{\lambda}_n) \right\}, \\ \hat{v}_{n+1} &= \arg \min_{\hat{v}^* \in V} \left\{ \frac{1}{2} \|\hat{v} - \hat{v}_n\|^2 + \gamma \mathbb{L}_\delta(\bar{x}'_n, \bar{v}_n, \hat{v}, \bar{\mu}_n, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (5.5.12)$$

Define the extended variables

$$\tilde{x} := \begin{pmatrix} x' \\ v \\ \hat{v} \end{pmatrix} \in \tilde{X} := X \times V \times V, \quad \tilde{y} := \begin{pmatrix} \mu \\ \lambda \end{pmatrix} \in \tilde{Y} := \mathbb{R}^+ \times \mathbb{R}^+,$$

$$\tilde{w} = \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \end{pmatrix} \in \tilde{X} \times \tilde{Y}, \quad \tilde{v} = \begin{pmatrix} \tilde{v}_1 \\ \tilde{v}_2 \end{pmatrix} \in \tilde{X} \times \tilde{Y},$$

and the functions

$$\begin{aligned}\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{y}) &:= \mathbb{L}_\delta(x', v, \hat{v}, \mu, \lambda), \\ \Psi_\delta(\tilde{w}, \tilde{v}) &:= \tilde{\mathcal{L}}_\delta(\tilde{w}_1, \tilde{v}_2) - \tilde{\mathcal{L}}_\delta(\tilde{v}_1, \tilde{w}_2).\end{aligned}\tag{5.5.13}$$

For $\tilde{w}_1 = \tilde{x}$, $\tilde{w}_2 = \tilde{y}$, $\tilde{v}_1 = \tilde{v}_1^* = \tilde{x}^*$ and $\tilde{v}_2 = \tilde{v}_2^* = \tilde{y}^*$ we have

$$\Psi_\delta(\tilde{w}, \tilde{v}) := \tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{y}^*) - \tilde{\mathcal{L}}_\delta(\tilde{x}^*, \tilde{y}).\tag{5.5.14}$$

In these variables the relation (5.5.10) can be represented in a “short format” as

$$\tilde{v}^* = \arg \min_{\tilde{v} \in \tilde{X} \times \tilde{Y}} \left\{ \frac{1}{2} \|\tilde{w} - \tilde{v}^*\|^2 + \gamma \Psi_\delta(\tilde{w}, \tilde{v}^*) \right\}\tag{5.5.15}$$

with any positive γ .

5.6 Convergence Analysis

5.6.1 Auxiliary Results

Let us prove the following auxiliary results.

Lemma 5.2 *Let $f(z)$ be a convex function defined on a convex set Z . If z^* is a minimizer of function*

$$\varphi(z) = \frac{1}{2} \|z - x\|^2 + \alpha f(z)\tag{5.6.1}$$

on Z with fixed x , then $f(z)$ satisfies the inequality

$$\frac{1}{2} \|z^* - x\|^2 + \alpha f(z^*) \leq \frac{1}{2} \|z - x\|^2 + \alpha f(z) - \frac{1}{2} \|z - z^*\|^2.\tag{5.6.2}$$

Proof By the necessary condition for a minimum at z^*

$$\langle z^* - x + \alpha \nabla f(z^*), z - z^* \rangle \geq 0,$$

and the convexity property of $f(z)$, expressed as

$$f(z) \geq f(z^*) + \langle \nabla f(z^*), z - z^* \rangle,$$

we derive

$$0 \leq \langle z^* - x + \alpha \nabla f(z^*), z - z^* \rangle = \langle z^* - x, z - z^* \rangle + \alpha [f(z) - f(z^*)].$$

Using this inequality in the identity

$$\frac{1}{2} \|z - x\|^2 = \frac{1}{2} \|z - z^*\|^2 + \langle z^* - x, z - z^* \rangle + \frac{1}{2} \|z^* - x\|^2,$$

we get (5.6.2). The Lemma is proven. \square

Lemma 5.3 *If all partial derivative of $\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{y})$ (5.5.13 and 5.5.14) satisfy the Lipschitz condition with positive constant C_0 , then the following Lipschitz-type condition holds:*

$$\|[\Psi_\delta(\tilde{w} + h, \tilde{v} + g) - \Psi_\delta(\tilde{w}, \tilde{v} + g)] - [\Psi_\delta(\tilde{w} + h, \tilde{v}) - \Psi_\delta(\tilde{w}, \tilde{v})]\| \leq C_0 \|h\| \|g\|, \quad (5.6.3)$$

valid for any $\tilde{w}, h, \tilde{v}, g \in \tilde{X} \times \tilde{Y}$.

Proof By the Lagrange's formula

$$f(x + h) - f(x) = \int_0^1 \langle \nabla f(x + th), h \rangle dt,$$

we have

$$\begin{aligned} & \|[\Psi_\delta(\tilde{w} + h, \tilde{v} + g) - \Psi_\delta(\tilde{w}, \tilde{v} + g)] - [\Psi_\delta(\tilde{w} + h, \tilde{v}) - \Psi_\delta(\tilde{w}, \tilde{v})]\| = \\ & \left\| \int_0^1 \langle \nabla \Psi_\delta(\tilde{w} + th, \tilde{v} + g) - \nabla \Psi_\delta(\tilde{w} + th, \tilde{v}), h \rangle dt \right\| \leq \\ & \int_0^1 | \langle \nabla \Psi_\delta(\tilde{w} + th, \tilde{v} + g) - \nabla \Psi_\delta(\tilde{w} + th, \tilde{v}), h \rangle | dt \leq \\ & \leq \int_0^1 C_0 \|h\| \|g\| dt \leq C_0 \|h\| \|g\|, \end{aligned}$$

which proves the Lemma. \square

5.6.2 Main Convergence Theorem

The following theorem presents the convergence conditions of (9.7.4)–(5.5.12) and gives the estimate of its rate of convergence.

Theorem 5.3 *Assume that problem (5.5.15) has a solution. Let $\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{y})$ be differentiable in \tilde{x} and \tilde{y} , whose partial derivative with respect to \tilde{y} satisfies the Lipschitz condition with positive constant C . Then, for any $\delta \in (0, 1)$, there exists a small-enough*

$$\gamma_0 = \gamma_0(\delta) < C := \min \left\{ \frac{1}{\sqrt{2C_0}}, \frac{1 + \sqrt{1 + 2C_0^2}}{2C_0^2} \right\}$$

such that, for any $0 < \gamma \leq \gamma_0$, sequence $\{\tilde{v}_n\}$, which generated by the equivalent extraproximal procedure (9.7.4)–(5.5.12), monotonically converges in norm with geometric progression rate $q \in (0, 1)$ to one of the equilibrium points \tilde{v}^* , i.e.,

$$\|\tilde{v}_n - \tilde{v}^*\|^2 \leq q^n \|\tilde{v}_0 - \tilde{v}^*\|^2, \quad (5.6.4)$$

where

$$q = 1 + \frac{4(\delta\gamma)^2}{1 + 2\delta\gamma - 2\gamma^2 C^2} - 2\delta\gamma. \quad (5.6.5)$$

Proof (1) Taking in (5.6.2) $\alpha = \gamma$ and

$$z = \tilde{w}, x = \tilde{v}_n, z^* = \hat{v}_n,$$

$$f(z) = \Psi_\delta(\tilde{w}, \tilde{v}_n), f(z^*) = \Psi_\delta(\hat{v}_n, \tilde{v}_n),$$

we obtain

$$\frac{1}{2} \|\hat{v}_n - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\hat{v}_n, \tilde{v}_n) \leq \frac{1}{2} \|\tilde{w} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{w}, \tilde{v}_n) - \frac{1}{2} \|\tilde{w} - \hat{v}_n\|^2. \quad (5.6.6)$$

Again putting in (5.6.2) $\alpha = \gamma$ and

$$z = \tilde{w}, x = \tilde{v}_n, z^* = \tilde{v}_{n+1},$$

$$f(z) = \Psi_\delta(\tilde{w}, \hat{v}_n), f(z^*) = \Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n),$$

we get

$$\frac{1}{2} \|\tilde{v}_{n+1} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) \leq \frac{1}{2} \|\tilde{w} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{w}, \hat{v}_n) - \frac{1}{2} \|\tilde{w} - \tilde{v}_{n+1}\|^2. \quad (5.6.7)$$

Selecting $\tilde{w} = \tilde{v}_{n+1}$ in (5.6.6) and $\tilde{w} = \hat{v}_n$ in (5.6.7) we obtain

$$\frac{1}{2} \|\hat{v}_n - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\hat{v}_n, \tilde{v}_n) \leq \frac{1}{2} \|\tilde{v}_{n+1} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_{n+1}, \tilde{v}_n) - \frac{1}{2} \|\tilde{v}_{n+1} - \hat{v}_n\|^2, \quad (5.6.8)$$

$$\frac{1}{2} \|\tilde{v}_{n+1} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) \leq \frac{1}{2} \|\hat{v}_n - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\hat{v}_n, \hat{v}_n) - \frac{1}{2} \|\hat{v}_n - \tilde{v}_{n+1}\|^2. \quad (5.6.9)$$

Adding (5.6.8) with (5.6.9) and using (5.6.3) for

$$\tilde{w} + h = \tilde{v}_{n+1}, \tilde{w} = \hat{v}_n, \tilde{v} + g = \tilde{v}_n, \tilde{v} = \hat{v}_n,$$

$$h = \tilde{v}_{n+1} - \hat{v}_n, g = \tilde{v}_n - \hat{v}_n,$$

we finally conclude

$$\|\tilde{v}_{n+1} - \hat{v}_n\|^2 \leq \gamma [\Psi_\delta(\tilde{v}_{n+1}, \tilde{v}_n) - \Psi_\delta(\hat{v}_n, \tilde{v}_n)] - \gamma [\Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) - \Psi_\delta(\hat{v}_n, \hat{v}_n)] \leq$$

$$\gamma C \|\tilde{v}_{n+1} - \hat{v}_n\| \|\tilde{v}_n - \hat{v}_n\|,$$

which implies

$$\|\tilde{v}_{n+1} - \hat{v}_n\| \leq \gamma C \|\tilde{v}_n - \hat{v}_n\|. \quad (5.6.10)$$

(2) Now, taking $\tilde{w} = \tilde{v}_{n+1}$ in (5.6.6) and $\tilde{w} = \tilde{v}_\delta^*$ in (5.6.7) we get

$$\frac{1}{2} \|\hat{v}_n - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\hat{v}_n, \tilde{v}_n) \leq \frac{1}{2} \|\tilde{v}_{n+1} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_{n+1}, \tilde{v}_n) - \frac{1}{2} \|\tilde{v}_{n+1} - \hat{v}_n\|^2,$$

$$\frac{1}{2} \|\tilde{v}_{n+1} - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) \leq \frac{1}{2} \|\tilde{v}_\delta^* - \tilde{v}_n\|^2 + \gamma \Psi_\delta(\tilde{v}_\delta^*, \hat{v}_n) - \frac{1}{2} \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2.$$

Adding these two inequalities and multiplying by two yields

$$\begin{aligned} & \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + \|\hat{v}_n - \tilde{v}_n\|^2 - 2\gamma \Psi_\delta(\tilde{v}_\delta^*, \hat{v}_n) + 2\gamma [\Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) \\ & + \Psi_\delta(\hat{v}_n, \tilde{v}_n) - \Psi_\delta(\tilde{v}_{n+1}, \tilde{v}_n)] \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2. \end{aligned}$$

Adding and subtracting the term $\Psi_\delta(\hat{v}_n, \hat{v}_n)$ we have

$$\begin{aligned} & \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + \|\hat{v}_n - \tilde{v}_n\|^2 + 2\gamma [\Psi_\delta(\hat{v}_n, \hat{v}_n) - \Psi_\delta(\tilde{v}_\delta^*, \hat{v}_n)] + 2\gamma [\Psi_\delta(\tilde{v}_{n+1}, \hat{v}_n) - \Psi_\delta(\hat{v}_n, \tilde{v}_n) + \Psi_\delta(\hat{v}_n, \tilde{v}_n) - \Psi_\delta(\tilde{v}_{n+1}, \tilde{v}_n)] \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2. \end{aligned}$$

Using (5.6.3) with $\tilde{w} + h = \tilde{v}_{n+1}$, $\tilde{w} = \hat{v}_n$, $\tilde{v} + k = \tilde{v}_n$ and $\tilde{v} = \hat{v}_n$ we have $h = \tilde{v}_{n+1} - \hat{v}_n$ and $k = \tilde{v}_n - \hat{v}_n$, and the inequality above becomes

$$\begin{aligned} & \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + \|\hat{v}_n - \tilde{v}_n\|^2 + 2\gamma [\Psi_\delta(\hat{v}_n, \hat{v}_n) - \Psi_\delta(\tilde{v}_\delta^*, \hat{v}_n)] - \\ & 2\gamma C \|\tilde{v}_{n+1} - \hat{v}_n\| \|\tilde{v}_n - \hat{v}_n\| \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2. \end{aligned}$$

Applying (5.6.10) to the last term in the left-hand side and in view of the strict convexity property of Ψ_δ given by

$$\Psi_\delta(\hat{v}_n, \hat{v}_n) - \Psi_\delta(\tilde{v}_\delta^*, \hat{v}_n) \geq \delta \|\hat{v}_n - \tilde{v}_\delta^*\|^2,$$

we get

$$\|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + 2\gamma \delta \|\hat{v}_n - \tilde{v}_\delta^*\|^2 + (1 - 2\gamma^2 C^2) \|\tilde{v}_n - \hat{v}_n\|^2 \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2.$$

Applying the identity $2\langle a - c, c - b \rangle = \|a - b\|^2 - \|a - c\|^2 - \|c - b\|^2$ with $a = \hat{v}_n$, $b = \tilde{v}_\delta^*$ and $c = \tilde{v}_n$, to the left-hand side of the last inequality we have

$$\begin{aligned} & \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + (1 - 2\gamma^2 C^2) \|\tilde{v}_n - \hat{v}_n\|^2 + 2\gamma \delta [2\langle \hat{v}_n - \tilde{v}_n, \tilde{v}_n - \tilde{v}_\delta^* \rangle + \|\tilde{v}_n - \hat{v}_n\|^2] + \\ & \|\tilde{v}_n - \tilde{v}_\delta^*\|^2 = \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + \|\hat{v}_n - \tilde{v}_n\|^2 + (1 + 2\gamma \delta - 2\gamma^2 C^2) \|\tilde{v}_n - \hat{v}_n\|^2 + 4\gamma \delta \langle \hat{v}_n - \tilde{v}_n, \tilde{v}_n - \tilde{v}_\delta^* \rangle \\ & + 2\gamma \delta \|\tilde{v}_n - \tilde{v}_\delta^*\|^2 \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2. \end{aligned}$$

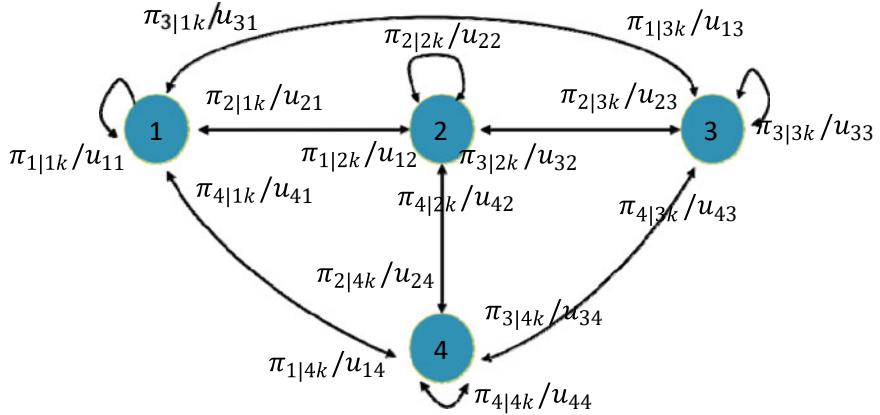


Fig. 5.6 Supermarket Markov Chain

Defining $d = 1 + 2\gamma\delta - 2\gamma^2C^2$ and completing the square form of the third and fourth terms yields

$$\|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + d\|\tilde{v}_n - \hat{v}_n\|^2 + 4\gamma\delta\langle \hat{v}_n - \tilde{v}_n, \tilde{v}_n - \tilde{v}_\delta^* \rangle + \frac{(2\gamma\delta)^2}{d}\|\tilde{v}_n - \tilde{v}_\delta^*\|^2 - \frac{(2\gamma\delta)^2}{d}\|\tilde{v}_n - \tilde{v}_\delta^*\|^2 + 2\gamma\delta\|\tilde{v}_n - \tilde{v}_\delta^*\|^2 \leq \|\tilde{v}_\delta^* - \tilde{v}_n\|^2,$$

and

$$\begin{aligned} & \|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 + \|\tilde{v}_{n+1} - \hat{v}_n\|^2 + \left\| \sqrt{d}(\tilde{v}_n - \hat{v}_n) + \frac{2\gamma\delta}{\sqrt{d}}(\tilde{v}_n - \tilde{v}_\delta^*) \right\|^2 \\ & \leq \left(1 - 2\gamma\delta + \frac{(2\gamma\delta)^2}{d}\right) \|\tilde{v}_\delta^* - \tilde{v}_n\|^2, \end{aligned}$$

finally implying

$$\|\tilde{v}_\delta^* - \tilde{v}_{n+1}\|^2 \leq q \|\tilde{v}_\delta^* - \tilde{v}_n\|^2 \leq q^{n+1} \|\tilde{v}_\delta^* - \tilde{v}_0\|^2 \xrightarrow{n \rightarrow \infty} 0,$$

with $q = 1 - 2\gamma\delta + \frac{(2\gamma\delta)^2}{d} \in (0, 1)$. The Theorem is proven. \square

5.7 Application Example: Four Supermarkets Chain

This example analyzes the effectiveness of relationship marketing strategies within the department store sector of the retail industry considering two supermarket leaders with $l = 1, 2$ and two supermarkets followers with $m = 3, 4$. The three supermarkets

are branching out into non-food items and they are also department stores in their own right, selling items as clothes, entertainment products for example toys, books, cosmetics, non-prescription drugs and many other household goods. All the supermarkets offer loyalty cards having their own system with the purpose to attract customers, encourage customer loyalty and build strong customer relationships. As well, loyalty cards create an advantage for supermarkets developing profiles of individuals' personal shopping habits. When linked with the personal details that customers disclosed when signing up for the scheme, the store is in a position to target promotions that are tailored around specific customers shopping habits. Based on the available data, supermarkets discretize the client space in four sub-segments according to the regularly of purchasing, using frequency of the loyalty card and the revenue. Figure 5.6 describes the segments and promotions corresponding to the Markov chain of the marketing problem. Here a customer is said to be in state s_1 if he/she become a Potential customer. A Low-Frequent customer corresponds with the state s_2 and a Regular customer is a frequent customer of the loyalty card that is said to be in state s_3 . A Loyal Customer corresponds with the state s_4 and he/she is a high-frequency user of the loyal card. The promotions (actions) offered by the supermarkets include two different benefits: 1) points and 2) discounts. We are interested in contrasting the strategies applied by the supermarkets defined over all possible combinations of states (i, j) and actions (k) given a fixed utility $J_{ij,k}$.

Our goal is to analyze a four-player Stackelberg game for the norm $p = 1$ in a class of ergodic controllable finite Markov chains. Let $N_1 = N_2 = N_3 = N_4 = 4$, $M_1 = M_2 = M_3 = M_4 = 2$. The individual utility for each player are defined by

$$\begin{aligned}
 J_{ij,1}^{(1)} &= \begin{bmatrix} 567 & 822 & 733 & 830 \\ 261 & 896 & 85 & 568 \\ 30 & 996 & 634 & 261 \\ 288 & 90 & 806 & 785 \end{bmatrix}, \quad J_{ij,2}^{(1)} = \begin{bmatrix} 170 & 27 & 57 & 699 \\ 275 & 855 & 224 & 919 \\ 50 & 205 & 46 & 909 \\ 398 & 861 & 751 & 806 \end{bmatrix}, \\
 J_{ij,1}^{(2)} &= \begin{bmatrix} 810 & 36 & 27 & 9 \\ 63 & 90 & 567 & 72 \\ 81 & 0 & 9 & 45 \\ 855 & 594 & 441 & 9 \end{bmatrix}, \quad J_{ij,2}^{(2)} = \begin{bmatrix} 5 & 370 & 30 & 0 \\ 40 & 40 & 195 & 10 \\ 165 & 20 & 75 & 45 \\ 250 & 35 & 25 & 125 \end{bmatrix}, \\
 J_{ij,1}^{(3)} &= \begin{bmatrix} 22 & 7 & 11 & 6 \\ 10 & 0 & 9 & 8 \\ 23 & 28 & 23 & 9 \\ 90 & 5 & 12 & 1 \end{bmatrix}, \quad J_{ij,2}^{(3)} = \begin{bmatrix} 11 & 0 & 21 & 7 \\ 3 & 13 & 40 & 1 \\ 6 & 3 & 10 & 26 \\ 11 & 17 & 30 & 8 \end{bmatrix}, \\
 J_{ij,1}^{(4)} &= \begin{bmatrix} 0 & 6 & 2 & 6 \\ 10 & 26 & 36 & 48 \\ 14 & 56 & 28 & 24 \\ 8 & 12 & 16 & 38 \end{bmatrix}, \quad J_{ij,2}^{(4)} = \begin{bmatrix} 6 & 4 & 54 & 12 \\ 0 & 4 & 2 & 16 \\ 6 & 8 & 50 & 2 \\ 12 & 30 & 48 & 14 \end{bmatrix}.
 \end{aligned}$$

The transition matrices for each player are defined as follows

$$\pi_{j|i1}^{(1)} = \begin{bmatrix} 0.2759 & 0.4886 & 0.0366 & 0.1989 \\ 0.1752 & 0.0953 & 0.3825 & 0.3470 \\ 0.1695 & 0.2629 & 0.4103 & 0.1574 \\ 0.2612 & 0.1665 & 0.4124 & 0.1600 \end{bmatrix}, \quad \pi_{j|i2}^{(1)} = \begin{bmatrix} 0.0863 & 0.3672 & 0.3201 & 0.2264 \\ 0.4339 & 0.1684 & 0.1919 & 0.2058 \\ 0.3856 & 0.2349 & 0.1324 & 0.2471 \\ 0.1475 & 0.3500 & 0.1903 & 0.3122 \end{bmatrix},$$

$$\pi_{j|i1}^{(2)} = \begin{bmatrix} 0.1761 & 0.1204 & 0.3883 & 0.3151 \\ 0.2207 & 0.1632 & 0.2354 & 0.3807 \\ 0.0708 & 0.3708 & 0.1364 & 0.4219 \\ 0.0132 & 0.5169 & 0.4127 & 0.0572 \end{bmatrix}, \quad \pi_{j|i2}^{(2)} = \begin{bmatrix} 0.2033 & 0.2456 & 0.2667 & 0.2844 \\ 0.2732 & 0.1032 & 0.3046 & 0.3190 \\ 0.1207 & 0.0930 & 0.3997 & 0.3866 \\ 0.1032 & 0.6976 & 0.1609 & 0.0383 \end{bmatrix},$$

$$\pi_{j|i1}^{(3)} = \begin{bmatrix} 0.4109 & 0.1654 & 0.0918 & 0.3319 \\ 0.3015 & 0.2201 & 0.1029 & 0.3756 \\ 0.1709 & 0.5673 & 0.0292 & 0.2326 \\ 0.1885 & 0.1491 & 0.3317 & 0.3307 \end{bmatrix}, \quad \pi_{j|i2}^{(3)} = \begin{bmatrix} 0.3046 & 0.2883 & 0.2573 & 0.1498 \\ 0.2470 & 0.0978 & 0.3060 & 0.3492 \\ 0.3006 & 0.0439 & 0.4387 & 0.2169 \\ 0.1141 & 0.3397 & 0.1855 & 0.3607 \end{bmatrix},$$

$$\pi_{j|i1}^{(4)} = \begin{bmatrix} 0.2610 & 0.3145 & 0.2088 & 0.2158 \\ 0.3777 & 0.1968 & 0.1574 & 0.2681 \\ 0.2593 & 0.0308 & 0.5113 & 0.1986 \\ 0.3401 & 0.4638 & 0.1200 & 0.0761 \end{bmatrix}, \quad \pi_{j|i2}^{(4)} = \begin{bmatrix} 0.0316 & 0.4652 & 0.2221 & 0.2811 \\ 0.1624 & 0.3245 & 0.3691 & 0.1440 \\ 0.1448 & 0.5777 & 0.2087 & 0.0688 \\ 0.2536 & 0.1996 & 0.3231 & 0.2237 \end{bmatrix}.$$

Given $\delta = 0.05$ and $\gamma = 0.0001$ and applying the extraproximal method we obtain the convergence of the strategies in terms of the variable $c_{i|k}^l$ for the leaders (see Fig. 5.7) and the convergence of the strategies in terms of the variable $c_{i|k}^m$ for the followers (see Fig. 5.8). In addition, the Fig. 5.9 show the convergence of the parameters J_i and Ω .

With final values $\lambda^{(1)*} = 0.6096$ and $\lambda^{(2)*} = 0.3904$ for the leaders, and $\theta^{(1)*} = 0.4952$ and $\theta^{(2)*} = 0.5048$ for the followers (see Fig. 5.10), the mixed strategies

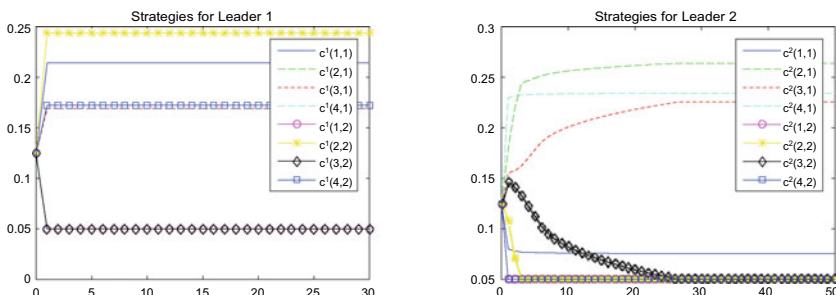
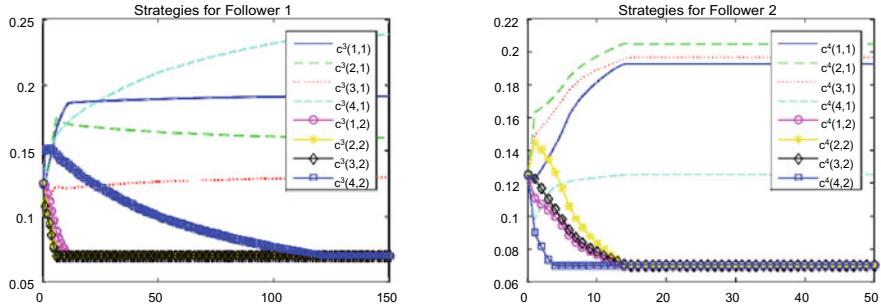
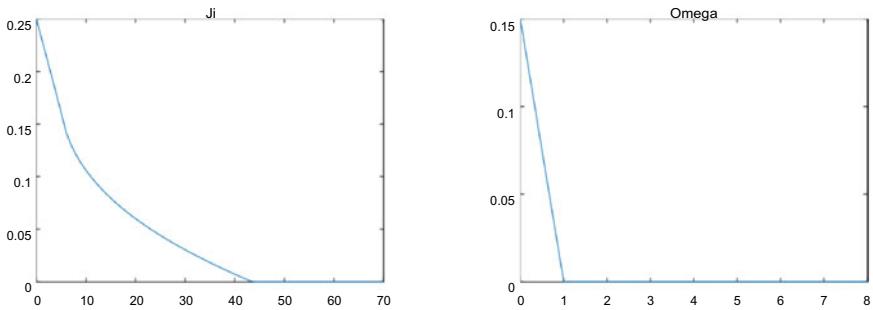
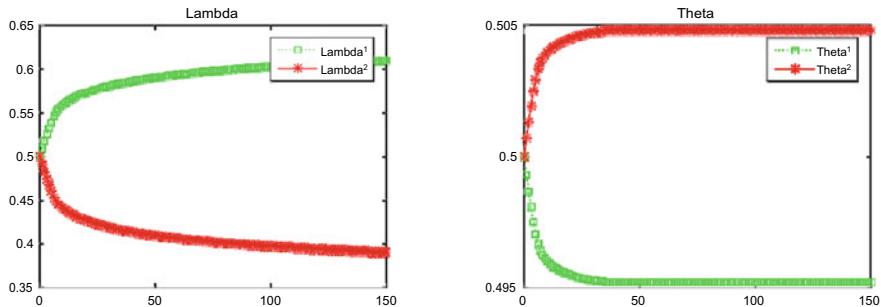


Fig. 5.7 Convergence of the strategies of the leader 1 (left) leader 2 (right)

**Fig. 5.8** Convergence of the strategies of the follower 1 (left) and follower 2 (right)**Fig. 5.9** Convergence of the parameters J_i and Ω **Fig. 5.10** Convergence of λ and θ

obtained for determining the strong Stackelberg/Nash equilibrium for all the players are as follows

$$d^{(1)*} = \begin{bmatrix} 0.8110 & 0.1890 \\ 0.1701 & 0.8299 \\ 0.7720 & 0.2280 \\ 0.2249 & 0.7751 \end{bmatrix}, \quad d^{(2)*} = \begin{bmatrix} 0.6023 & 0.3977 \\ 0.8408 & 0.1592 \\ 0.8187 & 0.1813 \\ 0.8242 & 0.1758 \end{bmatrix}, \quad (5.7.1)$$

$$d^{(3)*} = \begin{bmatrix} 0.7326 & 0.2674 \\ 0.6958 & 0.3042 \\ 0.6500 & 0.3500 \\ 0.7728 & 0.2272 \end{bmatrix}, \quad d^{(4)*} = \begin{bmatrix} 0.7337 & 0.2663 \\ 0.7454 & 0.2546 \\ 0.7376 & 0.2624 \\ 0.6418 & 0.3582 \end{bmatrix}.$$

The resulting utilities by segment are as follows:

$$J^{(1)}(s_i) = \begin{bmatrix} 129, 130 \\ 92, 800 \\ 84, 590 \\ 121, 520 \end{bmatrix}, \quad J^{(2)}(s_i) = \begin{bmatrix} 10, 463 \\ 21, 684 \\ 618 \\ 64, 189 \end{bmatrix}, \quad (5.7.2)$$

$$J^{(3)}(s_i) = \begin{bmatrix} 55.1402 \\ 50.1599 \\ 92.6922 \\ 396.9852 \end{bmatrix}, \quad J^{(4)}(s_i) = \begin{bmatrix} 321.4290 \\ 241.9381 \\ 171.7218 \\ 98.8860 \end{bmatrix}. \quad (5.7.3)$$

The resulting utilities by promotion are as follows:

$$\left. \begin{array}{l} J^{(1)}(k_i) = [226, 830 201, 200] \\ J^{(2)}(k_i) = [93, 934 3, 019] \end{array} \right\} \quad (5.7.4)$$

$$\left. \begin{array}{l} J^{(3)}(k_i) = [508.5371 86.4404] \\ J^{(4)}(k_i) = [608.5561 225.4188] \end{array} \right\} \quad (5.7.5)$$

Relationship marketing recognizes that the focus of marketing is to build a relationship with existing customers. The main purpose of the game is to discover the extent to which customers use and are influenced by relationship marketing strategies. In addition, it is to analyze the impact that these strategies have on customer loyalty and the development of customer-department store relationship. The supermarket leaders (players 1 and 2) fix their strategies (5.7.1) to ensure high degrees of customer loyalty and retention as well utility by segment (5.7.2) and promotion. For segment 1, the leader 1 made a strong emphasis on offering points (0.8110) for attracting Potential customers. Instead, the leader 2 made emphasis on offering points (0.6023) and discounts (0.3977) for the same segment. Looking at the utilities of the leaders (5.7.2), the follower1 resolved for competing highlighting points (0.7326). Instead, the follower 2 decided for offering points (0.7337) and discounts (0.2663). For segment 2 corresponding to Low-Frequent customers the leader 1 promoted points (0.1701) and discounts (0.8299) and, the leader 2 chose offering points (0.8408) and discounts (0.1592). However, for competing with the leaders, follower 1 and follower 2 made emphasis on points (0.6958 and 0.7454 respectively). For Regular customers the leader 1 focused on points (0.7720) and discounts (0.2280) and, the leader 2 made emphasis on points (0.8187). The follower 1 preferred offering

points (0.6500) and discounts (0.3500). Instead, follower 2 made emphasis on points (0.7376) and discounts (0.2624). For Loyal customers the leader 1 made emphasis on points (0.2249) and discounts (0.7751), leader 2 focus on points (0.8242) and discounts (0.1758) as well, follower 1 chose the same strategies—points (0.7728) and discounts (0.2272)—. The follower 2 made emphasis on points (0.6418) and discounts (0.3582). For the leaders the most profitable segments are the Potential customers and the Loyal customers (see 5.7.2 vs. 5.7.3). An insight into the mind of the consumer is obvious from the findings the importance that is placed on a given policy: the utilities obtained by action for the leaders and followers are shown in Eqs. (5.7.4) and (5.7.5) respectively.

References

1. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. *Comput. Math. Math. Phys.* **45**(11), 1893–1914 (2005)
2. Clemplner, J.B.: Setting cournot versus lyapunov games stability conditions and equilibrium point properties. *Int. Game Theory Rev.* **17**, 1–10 (2015)
3. Clemplner, J.B.: A proximal/gradient approach for computing the Nash equilibrium in controllable Markov games. *J. Optim. Theory Appl.* **188**(3), 847–862 (2021)
4. Clemplner, J.B., Poznyak, A.S.: Convergence method, properties and computational complexity for Lyapunov games. *Int. J. Appl. Math. Comput. Sci.* **21**(2), 349–361 (2011)
5. Clemplner, J.B., Poznyak, A.S.: Analysis of best-reply strategies in repeated finite Markov chains games. In: IEEE Conference on Decision and Control (2013)
6. Clemplner, J.B., Poznyak, A.S.: Computing the strong Nash equilibrium for Markov chains games. *Appl. Math. Comput.* **265**, 911–927 (2015)
7. Clemplner, J.B., Poznyak, A.S.: Convergence analysis for pure and stationary strategies in repeated potential games: Nash, Lyapunov and correlated equilibria. *Expert Syst. Appl.* **46**, 474–484 (2016)
8. Clemplner, J.B., Poznyak, A.S.: A Tikhonov regularization parameter approach for solving Lagrange constrained optimization problems. *Eng. Optim.* **50**(11), 1996–2012 (2018)
9. Clemplner, J.B., Poznyak, A.S.: A Tikhonov regularized penalty function approach for solving polylinear programming problems. *J. Comput. Appl. Math.* **328**, 267–286 (2018)
10. Clemplner, J.B., Poznyak, A.S.: Finding the strong Nash equilibrium: computation, existence and characterization for Markov games. *J. Optim. Theory Appl.* **186**, 1029–1052 (2020)
11. Dreves, A.: Computing all solutions of linear generalized Nash equilibrium problems. *Math. Methods Oper. Res.* (2016). <https://doi.org/10.1007/s00186-016-0562-0>
12. Dreves, A., Kanzow, C., Stein, O.: Nonsmooth optimization reformulations of player convex generalized Nash equilibrium problems. *J. Glob. Optim.* **53**(4), 587–614 (2012)
13. Facchinei, F., Kanzow, C., Sagratella, S.: Solving quasi-variational inequalities via their KKT conditions. *Math. Program.* **144**(1–2), 369–412 (2014)
14. Facchinei, F., Sagratella, S.: On the computation of all solutions of jointly convex generalized Nash equilibrium problems. *Optim. Lett.* **5**(3), 531–547 (2011)
15. Gabriel, S.A., Siddiqui, S., Conejo, A.J., Ruiz, C.: Solving discretely-constrained Nash-Cournot games with an application to power markets. *Netw. Spat. Econ.* **13**(3), 307–326 (2013)
16. Kreps, D.M.: Nash equilibrium. In: *Game Theory*, pp. 167–177. Springer (1989)
17. Nabetani, K., Tseng, P., Fukushima, M.: Parametrized variational inequality approaches to generalized Nash equilibrium problems with shared constraints. *Comput. Optim. Appl.* **8**(3), 423–452 (2011)
18. Nash, J.F.: Non-cooperative games. *Ann. Math.* **54**, 286–295 (1951)

19. Osborne, M.J., Rubinstein, A.: *A Course in Game Theory*. MIT Press (1994)
20. Tanaka, K., Yokoyama, K.: On ϵ -equilibrium point in a noncooperative n-person game. *J. Math. Anal.* **160**, 413–423 (1991)
21. Trejo, K.K., Clemmpner, J.B., Poznyak, A.S.: Computing the Stackelberg/Nash equilibria using the extraproximal method: convergence analysis and implementation details for Markov chains games. *Int. J. Appl. Math. Comput. Sci.* **25**(2), 337–351 (2015)
22. Trejo, K.K., Clemmpner, J.B., Poznyak, A.S.: An optimal strong equilibrium solution for cooperative multi-leader-follower Stackelberg Markov chains games. *Kibernetika* **52**(2), 258–279 (2016)
23. Trejo, K.K., Clemmpner, J.B., Poznyak, A.S.: Computing the L^p -strong Nash equilibrium for Markov chains games. *Appl. Math. Modell.* **41**, 399–418 (2017)
24. von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*, 2nd rev. Princeton University Press (1947)

Chapter 6

Best-Reply Strategies in Repeated Games



Abstract The actions that players naturally and frequently choose to perform throughout a repeated game are the “*best-reply strategy*”. Making the determination that such strategies lead to an equilibrium point is a difficult and sensitive operation since, often, the behavior of a single cost-function when such best-reply techniques are used turns out to be non-monotonic. The convergence to a stable equilibrium is not always guaranteed, even in repeated games. In this chapter, we demonstrate that the best-reply actions surely lead to an equilibrium point for a type of finite controlled Markov Chains dynamic games. The *Lyapunov Games concept*, which is based on the design of a unique Lyapunov function (associated with a unique cost function) that monotonically reduces (non-increases) throughout the game, achieves this result. We provide a technique for creating a Lyapunov-like function that describes how players behave in a recurrent Markov chain game. The Lyapunov-like function replaces the components of the ergodic system, which simulate players’ anticipated behavior in one-shot games, for the recursive process. We first provide a non-converging state-value function that varies (increases and decreases) between states of the repeating Markov game in order to demonstrate our claim. Then, we demonstrate that using a one-step-ahead fixed-local-optimal approach, it is possible to describe that function in a recursive format. Thus, we show that the previous recursive expression for the repeated game can be used to construct a Lyapunov-like function; the resulting Lyapunov-like function is a monotonic function that can only decrease (or remain the same) over time, regardless of the initial distribution of probabilities. The best-reply methods examined in this study are connected to what are known as pure and stationary fixed-local-optimal actions, or, to put it another way, with one step ahead optimization algorithms that are extensively employed in the *Artificial Intelligence* theory. The Bank Marketing Campaigns Game and the Duel Game with Best-Reply Strategies Application serve as examples of the proposed strategy.

6.1 Introduction

The behaviors that players naturally and frequently employ when playing a game repeatedly are known as the “*best-reply* ”strategy. A local (one-step) projected optimization is actually realized by this operation, supposing that the prior history (states

and actions) cannot be modified going forward. It turns out that the behavior of a particular cost-function when such strategies are used is non-monotonic; as a result, it is difficult and requires extra study to conclude that such tactics lead to an equilibrium point, often the Nash equilibrium (see [13]). The convergence to a stationary equilibrium is not always guaranteed, even in repeated games (see [4, 12]).

Iterative solution approaches for strategic games postulate that players utilize an internal process of reasoning to weed out irrational tactics as they search for an equilibrium point [14, 17, 20, 22]. The definitions of unreasonable techniques used by various iterative solution approaches in the literature vary. On the one hand, there are so-called rationalizability notions [2, 18–20], for which a strategy is considered irrational if it is not the optimum response to a certain belief. However, there are notions known as dominance solutions for which a strategy is considered irrational if it is outperformed by another approach [3, 19].

In this chapter, we demonstrate that the best-reply actions inevitably lead to one of the Nash equilibrium points for a class of ergodic finite controllable Markov Chains dynamic games. The Lyapunov Games concept, which is based on the creation of a unique Lyapunov function (associated with a unique cost function) that monotonically reduces (non-increases) throughout the game, achieves this result. According to [7], the equilibrium point is guaranteed to occur naturally in Lyapunov games by definition. It is also certain that convergence to an equilibrium point will occur. A Lyapunov-like function monotonically decrements and reaches a Lyapunov equilibrium point while moving ahead through the state space. The behavior of a Lyapunov-like function is implemented naturally by the best-reply dynamics. A Lyapunov game also has the advantage that all players are aware that just the best reply is chosen. Additionally, a Lyapunov equilibrium point has stability characteristics that aren't always present in a Nash equilibrium point [5, 6, 9–11].

Without taking into account the starting strategies the players start with, a game is said to be stable with regard to a set of strategies if the iterated process of strategy selection (in our case, the best-reply dynamics) converges to an equilibrium point. Every player chooses their tactics in order to reach an equilibrium point by maximizing his own cost function while taking into account the available strategies from other players, according to [16]. Any departure from this equilibrium point would bring the system back to it. This is because the iterated process of selecting strategies in natural evolution attempts to follow the best ones and corrects the trajectory to arrive at a stable equilibrium point (this is the case when the equilibrium point is unique). In this sense, we may say that a Lyapunov equilibrium point is stable because once a player's choices have reached a stable state, it is not in their best advantage to modify their strategy on their own. The fact that every ergodic system may be represented by a Lyapunov-like function is a significant benefit of the Lyapunov games. A recursive method is used for a repeating (ergodic) game to support an equilibrium play [7, 8]. We have found an equilibrium position, and furthermore, a very reasonable one, if the stochastic game's ergodic process converges [21].

We outline a technique for creating a Lyapunov-like function that is one-to-one with a specified cost function and has monotonic behavior. A declining Lyapunov-like function that is constrained from below enables the convergence of the cost

function to a minimal value while also ensuring the presence of an equilibrium point for the used pure and stationary local-optimal techniques, according quote [7]. The vector Lyapunov-like function that results is monotonic, meaning that its components can only get smaller with time. As a consequence, a one-shot game might be used to symbolize a repeating game. In our case, repeated games are converted into one-shot games, substituting the recursive process with a Lyapunov-like function. This complicates the problem's justification. The Duel game with best-reply actions application and repeated asynchronous Bank marketing campaigns serve as examples of the offered methodology.

The main contribution of this chapter is as follows:

- Demonstrates that the cost sequence associated with the local-optimal (best-reply) strategy exhibits non-monotonic behavior, making it impossible to prove with exactness the existence of a limit point;
- Proposes a one-to-one mapping between the current cost function and a new energy function (Lyapunov-like function) that is monotonically non-increasing on the trajectories of the system.
- Shows that a Lyapunov equilibrium point is a Nash equilibrium point (the converse is false), but it also offers a number of benefits: A Lyapunov equilibrium point is guaranteed to exist by definition, a Lyapunov-like function can be built to adhere to the Markov game's constraints, a Lyapunov-like function invariably converges to a Lyapunov equilibrium point, and a Lyapunov equilibrium point exhibits properties of stability.
- Provides for a class of ergodic controllable finite Markov chains the convergence of the pure and stationary local-optimal (best-reply) strategy.
- Presents an analytical formula for the numerical implementation of the local-optimal (best-reply) strategy .

6.2 Preliminaries

6.2.1 Controllable Markov Decision Process

As usual let the set of real numbers be denoted by \mathbb{R} and let the set of non-negative integers be denoted by \mathbb{N} . The inner product for two vectors u, v in \mathbb{R}^n is denoted by $\langle u, v \rangle = v^T u$. Let S be a finite set, called the *state space*, consisting of all positive integers $N \in \mathbb{N}$ of states $\{s_1, \dots, s_N\}$. A *Stationary Markov chain* [15] is a sequence of S -valued random variables $s(t)$, $t \in \mathbb{N}$, satisfying the *Markov condition*:

$$\begin{aligned} P(s(t+1) = s_j | s(t) = s_i, s(t-1) = s_{i_{t-1}}, \dots, s(0) = s_{i_0}) = \\ P(s(t+1) = s_j | s(t) = s_i) := \pi_{j|i}(t). \end{aligned} \tag{6.2.1}$$

The Markov chain can be represented by a complete graph whose nodes are the states, where each edge $(s_i, s_j) \in S^2$ is labeled by the transition probability (6.2.1). The matrix $\Pi = (\pi_{j|i})_{(s_i, s_j) \in S} \in [0, 1]^{S \times S}$ determines the evolution of the chain: for each $k \in \mathbb{N}$, the power Π^k has in each entry (s_i, s_j) the probability of going from state s_i to state s_j in exactly k steps.

A *Controllable Markov Decision Process* is a 5-tuple

$$MDP = \{S, A, \Upsilon, \Pi, V\}, \quad (6.2.2)$$

where:

- S is a finite set of states, $S \subset \mathbb{N}$, endowed with discrete topology;
- A is the set of actions, which is a metric space. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \bigcup_{s \in S} A(s)$;
- $\Upsilon = \{(s, a) | s \in S, a \in A(s)\}$ is the set of admissible state-action pairs, which is a measurable subset of $S \times A$;
- $\Pi(k) = [\pi_{j|ik}]$ is a stationary transition controlled matrix, where

$$\pi_{j|ik}(t) := P(s(t+1) = s_j | s(t) = s_i, a(t) = a_k)$$

represents the probability associated with the transition from state s_i to state s_j under an action $a_k \in A(s_i)$, $k = 1, \dots, M$;

- $V : S \rightarrow \mathbb{R}$ is a cost function, associating to each state a real value.

The *Markov property* of the decision process in (6.2.2) is said to be fulfilled if

$$\begin{aligned} P(s(t+1)|(s(1), s(2), \dots, s(t-1)), s(t) = s_i, a(t) = a_k) \\ = P(s(t+1) = s_j | s(t) = s_i, a(t) = a_k). \end{aligned}$$

The strategy (policy)

$$d_{k|i}(t) \equiv P(a(t) = a_k | s(t) = s_i)$$

represents the probability measure associated with the occurrence of an action $a(t) = a_k$ from state $s(t) = s_i$. The elements of the transition matrix for the controllable Markov chain can be expressed as

$$\begin{aligned} P \{s(t+1) = s_j | s(t) = s_i\} = \\ \sum_{k=1}^M P \{s(t+1) = s_j | s(t) = s_i, a(t) = a_k\} d_{k|i}(t). \end{aligned} \quad (6.2.3)$$

6.2.2 Game Description

Let $\mathcal{N} = \{1, \dots, n\}$ be a set of player indexed by $l = \overline{1, n}$. We use notations $\Delta = \prod_{l \in \mathcal{N}} \Delta^l$ (the mixed strategies profile), and $\Delta^{-l} = \prod_{j \in \mathcal{N} \setminus \{l\}} \Delta^j$ (the mixed strategies profile of all the players except for player l). Let us denote the collection $\{d_{k|i}^l(t)\}$ by Δ_t as follows

$$\Delta_t = \{d_{k|i}^l(t)\}_{k=\overline{1,M}, i=\overline{1,N}, l=\overline{1,n}}.$$

In this chapter we will deal with the class of the, so-called, local-optimal policies (strategies) defined below.

Definition 6.1 A policy $\{\Delta_t^{l,loc}\}_{t \geq 0}$ is said to be **local-optimal** (or **best reply**) if for each $t \geq 0$ it minimizes the conditional mathematical expectation of the individual cost-function $V^l(s^l(t+1))$, $l = \overline{1, n}$, under the condition that the prehistory of the process

$$\mathcal{H}_t^l := \{\Delta_0^l, P\{s^l(0) = s_j^l\} j = \overline{1, N}; \dots; \Delta_{t-1}^l, P\{s^l(t) = s_j^l\} j = \overline{1, N}\}$$

is fixed and can not be changed hereafter, i.e., it realizes the “one-step ahead”conditional optimization rule

$$\Delta_t^{l,loc} := \arg \min_{D_t^l} \mathbb{E}\{V^l(s^l(t+1)) \mid \mathcal{H}_t^l\}, \quad (6.2.4)$$

where $V^l(s^l(t+1))$ is the cost function of the player l at the state $s^l(t+1)$ and D_t^l is the set of admissible strategies $\{d_{k|i}^l(t)\}$.

Remark 6.1 Locally optimal policy is known as a **myopic policy** in the games literature.

A non-cooperative stochastic game is a tuple

$$G = \langle \mathcal{N}, S, (\Delta^l)_{l \in \mathcal{N}}, (\gamma^l)_{l \in \mathcal{N}}, \Pi, (V^l)_{l \in \mathcal{N}} \rangle.$$

For a strategy $d = (d^1, \dots, d^n) \in \Delta$ we denote the complement strategy $d^{-l} = (d^1, \dots, d^{l-1}, d^{l+1}, \dots, d^n)$ and, with an abuse of notation, $d = (d^l, d^{-l})$. The state $d = (d^1, \dots, d^n)$ represents the distribution vector of strategy frequencies and can only move on Δ .

Let us denote by $\mathbf{V}^d(t+1)$ the **vector average cost function** at the state $s^l(t+1)$ and time $(t+1)$ under the fixed strategy $d^l(t) = d^l$, that is,

$$\mathbf{V}^d(t+1) := \left(\mathbf{V}^{1,d^1}(t+1), \dots, \mathbf{V}^{n,d^n}(t+1) \right),$$

where $\mathbf{V}^{l,d^l}(t+1)$ is the average cost function at the state $s^l(t+1)$ and time $(t+1)$ for the player l , namely,

$$\mathbf{V}^{l,d^l}(t+1) := \mathbb{E}(V^l(s^l(t+1))|d^l(t)),$$

and $\mathbb{E}(\cdot|d^l(t))$ is the operator of the conditional mathematical expectation subject to the constraint that at time t the mixed strategy $d^l(t)$ has been applied.

6.3 Problem Formulation

To tackle this problem we proposed representing the state-value function V using a linear (with respect to the control $d \in \Pi$) model. After that we obtain the policy d that results in the minimum trajectory value. Finally, we present V in a recursive matrix format.

6.3.1 The State-Value Function

The probability of the player $l \in \mathcal{N}$ in the game G to find itself in the next state is as follows:

$$P^l(s^l(t+1) = s_j^l | s^l(t) = s_i^l) =$$

$$\sum_{k=1}^M P^l(s^l(t+1) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l) d_{k|i}^l(t) = \sum_{k=1}^M \pi_{j|ik}^l d_{k|i}^l(t).$$

The cost function V^l of any fixed policy $d^l(t)$ is defined over all possible combinations of states and actions, and indicates the expected value when taking action a^l in state s and following policy $d^l(t)$ thereafter. The V -values for all the states of (6.2.2) in open format can be expressed by

$$\begin{aligned} \mathbf{V}^{l,d^l}(t+1) := \mathbb{E}(V^l(s^l(t+1))|d^l(t)) = \\ \left. \sum_{j=1}^N \sum_{i=1}^N \sum_{k=1}^M V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l \right\} \pi_{j|ik}^l d_{k|i}^l(t) P^l(s^l(t) = s_i^l), \end{aligned} \quad (6.3.1)$$

where $V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l$ is a loss value at state s_i^l when the action a_k^l is applied (without loss of generality it can be assumed to be positive) and $P^l(s^l(t))$ for any given $P^l(s^l(0))$ is defined as follows

$$P^l(s^l(t+1)) = s_j^l = \sum_{i=1}^N P^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l,$$

$$P^l(s^l(t) = s_i^l) = \sum_{i=1}^N \left(\sum_{k=1}^M \pi_{j|ik}^l d_{k|i}^l(t) \right) P^l(s^l(t) = s_i^l),$$

or, in matrix format,

$$\mathbf{p}_{n+1}^l = (\cdot_n^l)^\top \mathbf{p}_n^l,$$

$$(\cdot_n^l)_{ij} := \sum_{k=1}^M \sum_{k=1}^M \pi_{j|ik}^l d_{k|i}^l(t).$$

Remark 6.2 We will assume hereafter that

$$V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l > 0$$

for all l . Indeed, by the identity

$$\mathbb{E}(V^l(s^l(t+1))|d^l(t)) =$$

$$\sum_{j=1}^N \sum_{i=1}^N \sum_{k=1}^M V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l) \pi_{j|ik}^l d_{k|i}^l(t) P^l(s^l(t) = s_i^l) =$$

$$\sum_{j=1}^N \sum_{i=1}^N \sum_{k=1}^M [V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l) + c] \cdot \pi_{j|ik}^l d_{k|i}^l(t) P^l(s^l(t) = s_i^l) - c$$

the minimization of the state-value function $\mathbb{E}(V^l(s^l(t+1))|d^l(t))$ is equivalent to the minimization of the function $\mathbb{E}(\tilde{V}^l(s^l(t+1))|d^l(t))$ where

$$\tilde{V}^l(s^l(t+1))|d^l(t)) = V^l(s^l(t+1))|d^l(t)) + c,$$

which is strictly positive if we take

$$c > \max_{1 \leq i \leq N, 1 \leq k \leq M} V^l(s_n = s(i) | s_n = s(i), a_n^l = a^l(k)). \quad (6.3.2)$$

In a vector format, the formula (6.3.1) can be expressed as

$$\mathbf{v}_{t+1}^{l,d^l} := \mathbb{E}(V^l(s^l(t+1))|d^l(t)) =$$

$$\sum_{i=1}^N \left[\sum_{j=1}^N \sum_{k=1}^M V^l(s^l(t+1)) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l) \pi_{j|ik}^l d_{k|i}^l(t) \right] P^l(s^l(t) = s_i^l) = \langle \mathbf{w}_t^l, \mathbf{p}_t^l \rangle,$$

where

$$(\mathbf{w}_t^l)_i := \sum_{j=1}^N \left[\sum_{k=1}^M V_{ij|k}^l \pi_{j|ik}^l d_{k|i}^l(t) \right],$$

$$V_{ij|k}^l := V^l(s^l(t+1) = s_j^l | s^l(t) = s_i^l, a^l(t) = a_k^l),$$

$$(\mathbf{p}_t^l)_i := P^l(s^l(t) = s_i^l).$$

6.3.2 The Recursive Matrix Form

Let us first introduce the following statement about the unit simplex.

Let Δ be the *unit simplex* in \mathbf{R}^M , that is,

$$\Delta = \{u \in \mathbf{R}^M \mid \sum_{k=1}^M u(k) = 1, u(k) \geq 0\}.$$

Then,

$$\min_{u \in \Delta} \sum_{k=1}^M v(k)u(k) = \min_{k=1, \dots, M} v(k) = v(\alpha),$$

and the minimum is achieved at least for $u = \left(\underbrace{0, 0, \dots, 0}_{\alpha}, 1, 0, \dots, 0 \right)$.

Indeed, it is evident that

$$\sum_{k=1}^M v(k)u(k) \geq \sum_{k=1}^M (\min v(k))u(k) = \min v(k) \sum_{k=1}^M u(k) = \min v(k) = v(\alpha),$$

and the equality is achieved at least for $u = \left(\underbrace{0, 0, \dots, 0}_{\alpha}, 1, 0, \dots, 0 \right)$.

As a result we have that

$$\begin{aligned} \mathbf{V}_{t+1}^l &= \langle \mathbf{w}_t^l, \mathbf{p}_t^l \rangle = \sum_{i=1}^N (\mathbf{w}_t^l)_i (\mathbf{p}_t^l)_i \geq \\ &\geq \sum_{i=1}^N \min_{d_{k|i}(t) \in \Delta} (\mathbf{w}_t^l)_i (\mathbf{p}_t^l)_i = \sum_{i=1}^N (\mathbf{p}_t^l)_i \min_{k=1, \dots, M} \left[\sum_{j=1}^N V_{ij|k}^l \pi_{j|ik}^l \right]. \end{aligned}$$

At this point, let us introduce the following general definition of Lyapunov-like function

Definition 6.2 Let $\mathbb{V} : S \rightarrow \mathbb{R}_+$ be a continuous map. Then, \mathbb{V} is said to be a **Lyapunov-like function**¹ iff it satisfies the following properties :

- (1) $\exists s^*$, called below a **Lyapunov equilibrium point**, such that $\mathbb{V}^l(s^*) = 0$,
- (2) $\mathbb{V}^l(s) > 0$ for all $s \neq s^*$ and all $l \in \mathcal{N}$,
- (3) $\mathbb{V}^l(s^i) \rightarrow \infty$ if there exists a sequence $\{s^i\}_{i=1}^\infty$ with $s^i \rightarrow \infty$ as $i \rightarrow \infty$ for all $l \in \mathcal{N}$,
- (4) $\Delta \mathbb{V}^l(s', s) = \mathbb{V}^l(s') - \mathbb{V}^l(s) < 0$ for all $s \neq s' \neq s^*$ and $l \in \mathcal{N}$.

Given fixed history of the process $(\mathbf{p}_0^l, d^l(0), d^l(1), \dots, d^l(t-1))$

$$\min_{d^l(t)} \mathbf{V}_{t+1}^l = \sum_{i=1}^N (\mathbf{p}_t^l)_i \left[\sum_{j=1}^N V_{ij|k^*}^l \pi_{j|ik^*}^l \right]. \quad (6.3.4)$$

(and considering point (4) of Definition 6.2), the identity in (6.3.4) is achieved for the pure and stationary local-optimal policy.

$$d_{k^*(i)}^{l*}(t) = \delta_{k^*(i), i} \quad n = 0, 1, \dots \quad (6.3.5)$$

where $\delta_{k^*(i), i}$ is the Kronecker symbol and $k^*(i)$ is an index for which

$$\sum_{j=1}^N V_{ij|k^*}^l \pi_{j|ik^*}^l \leq \sum_{j=1}^N V_{ij|k}^l \pi_{j|ik}^l := W_{ik}^l, \forall k = 1, \dots, M. \quad (6.3.6)$$

As a result we can state the following lemma.

Definition 6.3 A *Lyapunov game* is a tuple

$$G = \langle \mathcal{N}, S, (\Delta^l)_{l \in \mathcal{N}}, (\mathbb{V}^l)_{l \in \mathcal{N}}, \Pi, (V^l)_{l \in \mathcal{N}} \rangle,$$

where V^l is a Lyapunov-like function (monotonically decreasing in time) and satisfies Definition 6.2.

¹ By the original definition of A. M. Lyapunov the following conditions must be satisfied for an energy function: locally for any small neighborhood Ω_δ of the origin the following inequalities must be satisfied for any $x \in \Omega_\delta$

$$\alpha \|x\|^2 \leq V(x) \leq \beta \|x\|^2, \alpha > 0, \\ V(x_{n+1}) < V(x_n). \quad (6.3.3)$$

If some additional requirements are necessary, or the above conditions hold globally (Lyapunov-Krasovskii) and the second inequality is fulfilled non-strictly, that is, $V(x_{n+1}) \leq V(x_n)$, then in these cases the considered energy function is commonly referred to as “Lyapunov-like function” (see, for example [1]).

Lemma 6.1 *Let*

$$G = \langle \mathcal{N}, S, (\Delta^l)_{l \in \mathcal{N}}, (\Upsilon^l)_{l \in \mathcal{N}}, \Pi, (V^l)_{l \in \mathcal{N}} \rangle$$

be a non-cooperative stochastic game. Given a fixed-local-optimal policy, the V -values for all state-action pairs from (6.3.1) in the recursive matrix format become

$$\boxed{\mathbf{V}_{t+1}^l = \langle \mathbf{w}^{l*}, \mathbf{p}_t^l \rangle} \quad (6.3.7)$$

where $\mathbf{w}^{l*} := ((\mathbf{w}^{l*})_1, \dots, (\mathbf{w}^{l*})_N)$ and

$$(\mathbf{w}^{l*})_i := \sum_{j=1}^N V_{ij|k}^l \pi_{j|ik^*(i)}^l = \min_{k=1, M} W_{ik}^l,$$

$$W_{ik}^l = \sum_{j=1}^N V_{ij|k}^l \pi_{j|ik}^l.$$

Remark 6.3 Under the local-optimal strategy (6.3.5) the probability state-vector \mathbf{p}_t^l satisfies the following relation

$$\boxed{\mathbf{p}_{t+1}^l = (\cdot^{l*})^\top \mathbf{p}_t^l = \left((\cdot^{l*})^\top \right)^{t+1} \mathbf{p}_0^l,} \quad (6.3.8)$$

where

$$\boxed{(\cdot^{l*})_{ij} := \sum_{k=1}^M \pi_{j|ik}^l \delta_{k^*(i), i} = \pi_{j|ik^*(i)}^l.}$$

6.4 Construction of a Lyapunov-Like Function

The aim of this section is to associate to any cost function \mathbf{V}_t^l , governed by (6.3.7), a Lyapunov-like function which monotonically decreases (non-increases) on the trajectories of the given system.

6.4.1 Recurrent Form for the Cost Function

In view of (6.3.2) let us represent \mathbf{V}_{n+1}^l as

$$\begin{aligned}\mathbf{V}_{t+1}^l &= \langle \mathbf{w}^{l*}, \mathbf{p}_t^l \rangle = \langle \mathbf{w}^{l*}, \mathbf{p}_{t-1}^l \rangle + \langle \mathbf{w}^{l*}, \mathbf{p}_n^l - \mathbf{p}_{t-1}^l \rangle \\ &= \mathbf{V}_t^l + \left\langle \mathbf{w}^{l*}, \left[\left((\cdot^{l*})^\top \right)^t - \left((\cdot^{l*})^\top \right)^{t-1} \right] \mathbf{p}_0^l \right\rangle \\ &= \mathbf{V}_t^l + \left\langle \mathbf{w}^{l*}, \left[(\cdot^{l*})^\top - I \right] \left((\cdot^{l*})^\top \right)^{t-1} \mathbf{p}_0^l \right\rangle \\ &= \left[\left(1 + \frac{\langle [\cdot^{l*} - I] \mathbf{w}^{l*}, \mathbf{p}_{t-1}^l \rangle}{\mathbf{V}_t^l} \right) \mathbf{V}_t^l \right],\end{aligned}$$

and denoting

$$\alpha_t^l = \frac{\langle [\cdot^{l*} - I] \mathbf{w}^{l*}, \mathbf{p}_{t-1}^l \rangle}{\mathbf{V}_t^l} = \frac{\langle [\cdot^{l*} - I] \mathbf{w}^{l*}, \mathbf{p}_{t-1}^l \rangle}{\langle \mathbf{w}^{l*}, \mathbf{p}_{t-1}^l \rangle},$$

we get

$$\mathbf{V}_{t+1}^l = (1 + \alpha_t^l) \mathbf{V}_t^l. \quad (6.4.1)$$

6.4.2 The Lyapunov Function Design

Defining $\tilde{\alpha}_t^l$ as

$$\tilde{\alpha}_t^l = \begin{cases} \alpha_t^l & \text{if } \alpha_t^l \geq 0, \\ 0 & \text{if } \alpha_t^l < 0, \end{cases} \quad (6.4.2)$$

we get

$$\mathbf{V}_{t+1}^l = (1 + \alpha_t^l) \mathbf{V}_t^l \leq (1 + \tilde{\alpha}_t^l) \mathbf{V}_t^l, \quad (6.4.3)$$

which leads to the following statement.

Theorem 6.1 *Let*

$$G = \langle \mathcal{N}, S, (\Delta^l)_{l \in \mathcal{N}}, (\Upsilon^l)_{l \in \mathcal{N}}, \Pi, (V^l)_{l \in \mathcal{N}} \rangle$$

be a non-cooperative stochastic game and let the recursive matrix format be represented by (6.4.1). Then, a possible Lyapunov-like function $\mathbf{V}_n^{l,mon}$ (which is monotonically non-increasing) for G has the form

$$\begin{aligned}\mathbf{V}_t^{l,mon} &= \mathbf{V}_n^l \prod_{\tau=1}^{t-1} (1 + \tilde{\alpha}_\tau^l)^{-1} = \frac{1 + \alpha_{t-1}^l}{1 + \tilde{\alpha}_{t-1}^l} \mathbf{V}_{t-1}^{l,mon}, \\ \mathbf{V}_0^{l,mon} &= \mathbf{V}_0^l.\end{aligned} \quad (6.4.4)$$

Proof Let us consider the recursion

$$x_{t+1} \leq (1 + \gamma_t)x_t + \eta_t$$

with $\gamma_t, x_t, \eta_t \geq 0$. Defining

$$\tilde{x}_t := x_t \prod_{\tau=1}^{t-1} (1 + \gamma_\tau)^{-1}, \quad \tilde{\eta}_t := \eta_t \prod_{\tau=1}^t (1 + \gamma_\tau)^{-1}$$

and

$$y_t = \tilde{x}_t - \sum_{\tau=1}^{t-1} \tilde{\eta}_\tau,$$

we obtain $y_{t+1} \leq y_t$. Indeed,

$$\begin{aligned} \tilde{x}_{t+1} &= x_t \prod_{\tau=1}^t (1 + \gamma_\tau)^{-1} \leq \\ x_t \left[\prod_{\tau=1}^t (1 + \gamma_\tau)^{-1} \right] (1 + \gamma_t) + \eta_t \prod_{\tau=1}^t (1 + \gamma_\tau)^{-1} &= \tilde{x}_t + \tilde{\eta}_t \end{aligned}$$

which implies

$$y_{t+1} = \tilde{x}_{t+1} - \sum_{\tau=1}^t \tilde{\eta}_\tau \leq \tilde{x}_t + \tilde{\eta}_t - \sum_{\tau=1}^t \tilde{\eta}_\tau = \tilde{x}_t + \sum_{\tau=1}^{t-1} \tilde{\eta}_\tau = y_t,$$

and therefore $y_{t+1} \leq y_t$. In view of this we have

$$\begin{aligned} \mathbf{V}_{t+1}^{l,mon} &= \mathbf{V}_{t+1}^l \prod_{\tau=1}^t (1 + \tilde{\alpha}_\tau^l)^{-1} \leq (1 + \tilde{\alpha}_t^l) \mathbf{V}_t^l \prod_{\tau=1}^t (1 + \tilde{\alpha}_\tau^l)^{-1} = \\ \mathbf{V}_t^l \prod_{\tau=1}^{t-1} (1 + \tilde{\alpha}_\tau^l)^{-1} &= \mathbf{V}_t^{l,mon}, \end{aligned}$$

that proves the result. \square

Corollary 6.1 Since the sequence $\{\mathbf{V}_t^{l,mon}\}$ is bounded from below and monotonically non-increasing, then by the Weierstrass theorem it converges, that is, there exists a limit

$$\mathbf{V}_\infty^{l,mon} := \lim_{t \rightarrow \infty} \mathbf{V}_t^{l,mon}.$$

Corollary 6.2 If the series $\sum_{\tau=1}^t \tilde{\alpha}_\tau^l$ converges, i.e.,

$$\sum_{\tau=1}^\infty \tilde{\alpha}_\tau^l < \infty,$$

then the product $\prod_{\tau=1}^t (1 + \tilde{\alpha}_\tau^l)$ also converges (by the inequality $1 + x \leq e^x$ application that is valid for any $x \in \mathbf{R}$), namely,

$$\prod_{\tau=1}^\infty (1 + \tilde{\alpha}_\tau^l) < \infty, \tag{6.4.5}$$

which implies the existence of a limit (a convergence) of the sequence $\{\mathbf{V}_t^l\}$ of the given loss-function too, i.e.,

$$\mathbf{V}_\infty^l := \lim_{n \rightarrow \infty} \mathbf{V}_t^l = \mathbf{V}_\infty^{l,mon} \prod_{\tau=1}^{\infty} (1 + \tilde{\alpha}_\tau^l). \quad (6.4.6)$$

Remark 6.4 Notice that by the ergodicity property (see Chap. 1) the infinit product in (6.4.6) always exists for ergodic Markov chains, that is, $\prod_{\tau=1}^{\infty} (1 + \tilde{\alpha}_\tau^l) < \infty$, since by Corollary above

$$\begin{aligned} \prod_{\tau=1}^{\infty} (1 + \tilde{\alpha}_\tau^l) &\leq \exp \left\{ \sum_{\tau=1}^{\infty} \tilde{\alpha}_\tau^l \right\} \leq \exp \left\{ \sum_{\tau=1}^{\infty} \frac{\langle [J^l - I] \mathbf{w}^{l*}, \mathbf{p}_{\tau-1}^l \rangle}{\mathbf{V}_\tau^l} \right\} \leq \\ &\leq \exp \left\{ \sum_{\tau=1}^{\infty} \frac{\langle \mathbf{p}_\tau^l - \mathbf{p}_{\tau-1}^l, \mathbf{w}^{l*} \rangle}{\mathbf{V}_\tau^l} \right\} \leq \exp \left\{ \sum_{\tau=1}^{\infty} \frac{\langle \mathbf{p}_\tau^l - \mathbf{p}_{\tau-1}^l, \mathbf{w}^{l*} \rangle}{\mathbf{V}_\tau^{l,*}} \right\} \leq \\ &\leq \exp \left\{ \frac{\|\mathbf{w}^{l*}\|}{c} \sum_{\tau=1}^{\infty} \|\mathbf{p}_\tau^l - \mathbf{p}_{\tau-1}^l\| \right\} \leq \exp \left\{ \frac{\|\mathbf{w}^{l*}\|}{c} \sum_{\tau=1}^{\infty} \|(\mathbf{p}_\tau^l - \mathbf{p}^{l*}) - (\mathbf{p}_{\tau-1}^l - \mathbf{p}^{l*})\| \right\} \leq \\ &\leq \exp \left\{ 2 \frac{\|\mathbf{w}^{l*}\|}{c} \sum_{\tau=1}^{\infty} \|(\mathbf{p}_\tau^l - \mathbf{p}^{l*})\| \right\} \leq \exp \left\{ 2 \frac{\|\mathbf{w}^{l*}\|}{c} \sum_{\tau=1}^{\infty} C^l \exp \{-D^l \cdot \tau\} \right\} < \infty. \end{aligned}$$

This means that the behavior of the sequence $\{\mathbf{V}_t^{l,mon}\}$ may serve as an indicator of the convergence of the game: the approach of the vector-cost function $\{\mathbf{V}_t^{l,mon}\}$ to its limit point $\mathbf{V}_t^{l,*}$ means that we are close to one of the equilibrium points of the game. Note that this convergence is exponential.

6.5 Examples

6.5.1 Example 1 (Banks Marketing Planning as Prisoner's Dilemma)

Consider the case of two Banks that are planning the marketing campaigns expenditures for the next years. The repeated “Prisoner’s Dilemma” game can be used to represent the problem. If both Banks have an arrangement to leave marketing budgets unchanged, then their profits stay at high levels. However, if one of the Banks defects and increases its marketing expenditures, it may earn a greater income at the expense of the other Bank. But, if both Banks increase their marketing budgets, the increased marketing efforts may balance each other and prove ineffective, resulting in lower profits. Let $N = 2$ be the number of states and let $M = 2$ be the number of actions. Assigning numerical values to the levels of profits of the marketing cam-

paigns, where 10 means profits stay at high levels and 1 implies lower profits, the payoff matrices for $l = 1, 2$ are as shown below:

$$V_{ij}^1 = \begin{bmatrix} 5 & 1 \\ 10 & 3 \end{bmatrix} \text{ for Player 1}, V_{ij}^2 = \begin{bmatrix} 5 & 10 \\ 1 & 3 \end{bmatrix} \text{ for Player 2},$$

and let the transition matrix for $k = 1, 2$ be defined as follows

$$\pi_{j|i1}^1 = \begin{bmatrix} 0.0247 & 0.9753 \\ 0.9756 & 0.0244 \end{bmatrix}, \pi_{j|i2}^1 = \begin{bmatrix} 0.5668 & 0.4332 \\ 0.9960 & 0.0040 \end{bmatrix} \text{ for Player 1},$$

$$\pi_{j|i1}^2 = \begin{bmatrix} 0.1904 & 0.8096 \\ 0.8612 & 0.1388 \end{bmatrix}, \pi_{j|i2}^2 = \begin{bmatrix} 0.0027 & 0.9973 \\ 0.7693 & 0.2307 \end{bmatrix} \text{ for Player 2}.$$

The beginning profile is supposed to be uniform, that is, $P^l(s_0 = j) = 0.5$ for any player $l = 1, 2$ and its state $j = 1, 2$. But as it follows from the statements above in the ergodic case this profile can be arbitrarily selected without any influence to the final equilibrium point.

For d^{1*} and d^{2*} (6.3.5) the fixed local-optimal strategies, and k^* the best-reply strategy the following results have been obtained:

$$\left. \begin{array}{l} \pi^{1*} = \begin{bmatrix} 0.0247 & 0.9753 \\ 0.9756 & 0.0244 \end{bmatrix} \\ \text{for } n_0 = 1, \chi_{erg}^1 = 0.0247, \\ k^{1*} = [1, 1], \\ w^{1*} = [1.0986, 9.8290] \end{array} \right\} \text{for Player 1},$$

$$\left. \begin{array}{l} \pi^{2*} = \begin{bmatrix} 0.1904 & 0.8096 \\ 0.8612 & 0.1388 \end{bmatrix} \\ \text{for } n_0 = 1, \chi_{erg}^2 = 0.1904, \\ k^{2*} = [1, 1], \\ w^{2*} = [9.0482, 1.2775] \end{array} \right\} \text{for Player 2}.$$

- in Figs. 6.1 and 6.3 the state-value function behavior is shown (where during game repetition the states of the players fluctuate according to the given probabilistic dynamics) showing completely non-monotonic behavior;

Fig. 6.1 Non-monotonic behavior of the cost-function for Player 1

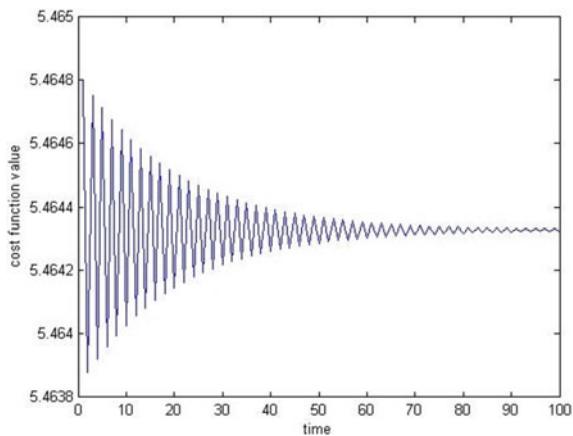
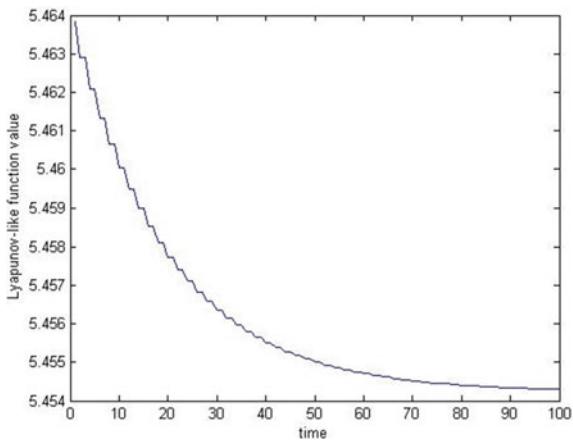


Fig. 6.2 Monotonic behavior of the Lyapunov-like function for Player 1



- in Figs. 6.2 and 6.4 the corresponding Lyapunov-like functions (6.4.4) are plotted definitely demonstrating a monotonic decreasing behavior;
- the results of the two methods clearly show that under the same fixed local-optimal strategy the original cost functions converge non-monotonically to the values 5.4643 (for the first player) and 5.0854 (for the second player) and the corresponding Lyapunov-like functions converge monotonically to the values 5.4543 and 4.9298, respectively which, obviously, are very close.

6.5.2 Example 2 (Duel Game)

In this example we consider the “Duel game” where Player I and Player II each have a gun loaded with exactly one bullet and stand 10 steps apart. The same mathematical interpretation has a military application in other repeated games such as “Fighter-

Fig. 6.3 Non-monotonic behavior of the cost-function for Player 2

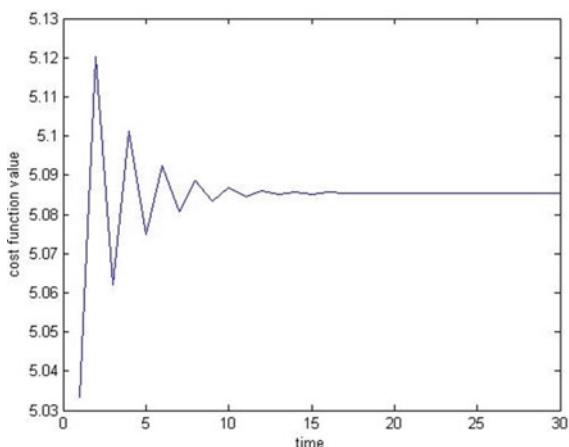
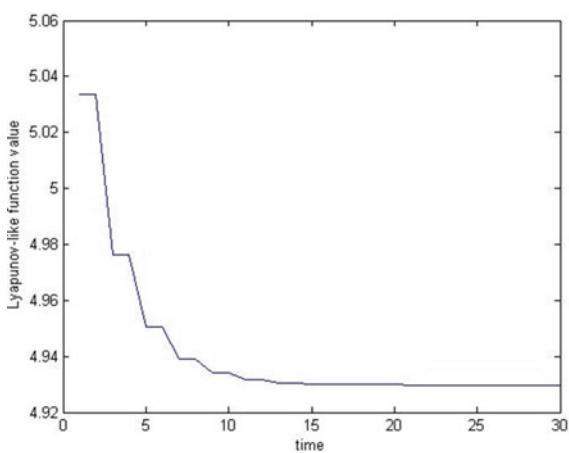


Fig. 6.4 Monotonic behavior of the Lyapunov-like function for Player 2



Bomber Duel Game” and “The Pursuit Game.” Starting with Player I, they take turns deciding whether to fire or not. Each time a player chooses not to fire, the other player takes one step forward before choosing whether to fire in turn. In other words, they start 10 steps apart facing each other and Player I decides whether to take a shot at Player II. If Player I does not, Player II takes a step forward and decides whether to take a shot at Player I. If Player II does not, Player I takes a step forward and decides whether to take a shot at Player II, and so on (players repeat the actions). The situation is grave because if a player fires and misses, the other can then simply not fire until they get next to each other and then shoot the opponent point blank. (Assume that if the players are next to each other, the one whose turn it is to shoot will certainly do so and will certainly hit the opponent.) The probability of hitting the opponent depends on the distance between them and on the skill of the shooter. Let $\sigma = 0, 1, 2, \dots, 10$ be the total number of steps taken by the players. Let ϕ_σ denote

the decision distance, i.e., the distance between the players, after σ steps. Thus, the initial decision distance for Player I is ϕ_0 , the next one is ϕ_1 for Player II, and so on. If players use pure strategies (ϕ_I, ϕ_{II}) where ϕ_I is the decision distance at which Player I opens fire, and ϕ_{II} is the decision distance at which Player II opens fire, then the outcome of the game depends on who fires first and successfully shoots the other one. If $\phi_I > \phi_{II}$ then Player I fires first with probability $p_I(\phi_I)$. If $\phi_I < \phi_{II}$ then Player II fires with probability $1 - p_{II}(\phi_{II})$. Then their payoff functions are given by:

$$V^I(\phi_I, \phi_{II}) = \begin{cases} p_I(\phi_I) & \text{if } \phi_I > \phi_{II}, \\ 1 - p_{II}(\phi_{II}) & \text{if } \phi_I < \phi_{II}, \end{cases} \quad (6.5.1)$$

and

$$V^{II}(\phi_I, \phi_{II}) = 1 - V^I(\phi_I, \phi_{II}). \quad (6.5.2)$$

To obtain a payoff matrix we assign values to the parameters of the game. Let the total distance $\Sigma = 1$ and let $\phi_\sigma = (0.1\sigma)$. Such that $p_I(\phi_\sigma) = 1 - \phi_\sigma$ is the probability of Player I hitting Player II firing at distance ϕ_σ , and $p_{II}(\phi_\sigma) = 1 - (\phi_\sigma)^2$ is the probability of Player II hitting Player I firing at distance ϕ_σ . Let $N = 5$ for the number of states and let $M = 2$ be the number of actions (for simplicity). The payoff functions reflect the players' desire to maximize the probability of survival. Then the payoff matrices for $l = 1, 2$ are as follows:

For Player1

$$V_{ij}^1 = \begin{bmatrix} 0.19 & 0.80 & 0.80 & 0.80 & 0.80 \\ 0.19 & 0.51 & 0.60 & 0.60 & 0.60 \\ 0.19 & 0.51 & 0.75 & 0.40 & 0.40 \\ 0.19 & 0.51 & 0.75 & 0.91 & 0.20 \\ 0.19 & 0.51 & 0.75 & 0.91 & 0.99 \end{bmatrix}.$$

For Player2

$$V_{ij}^2 = \begin{bmatrix} 0.81 & 0.20 & 0.20 & 0.20 & 0.20 \\ 0.81 & 0.49 & 0.40 & 0.40 & 0.40 \\ 0.81 & 0.49 & 0.25 & 0.60 & 0.60 \\ 0.81 & 0.49 & 0.25 & 0.09 & 0.80 \\ 0.81 & 0.49 & 0.25 & 0.09 & 0.01 \end{bmatrix},$$

and let the transition matrices for $k = 1, 2$ be defined as follows

For Player 1

$$\pi_{j|i1}^1 = \begin{bmatrix} 0.0107 & 0.0011 & 0.7950 & 0.0795 & 0.1136 \\ 0.0129 & 0.0012 & 0.9858 & 0.0000 & 0.0001 \\ 0.9869 & 0.0123 & 0.0000 & 0.0006 & 0.0001 \\ 0.6171 & 0.0818 & 0.0755 & 0.0887 & 0.1369 \\ 0.4784 & 0.0066 & 0.0368 & 0.1104 & 0.3679 \end{bmatrix}, \quad \pi_{j|i2}^1 = \begin{bmatrix} 0.0112 & 0.0012 & 0.8514 & 0.0178 & 0.1185 \\ 0.0149 & 0.0000 & 0.9850 & 0.0001 & 0.0000 \\ 0.8117 & 0.0901 & 0.0089 & 0.0893 & 0.0000 \\ 0.3706 & 0.0740 & 0.4810 & 0.0078 & 0.0666 \\ 0.7246 & 0.0000 & 0.2415 & 0.0098 & 0.0241 \end{bmatrix}.$$

For Player 2

$$\pi_{j|i_1}^2 = \begin{bmatrix} 0.0836 & 0.0009 & 0.7397 & 0.0832 & 0.0925 \\ 0.1029 & 0.0013 & 0.8957 & 0.0000 & 0.0001 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3733 & 0.0121 & 0.6137 & 0.0001 & 0.0008 \\ 0.9984 & 0.0000 & 0.0001 & 0.0000 & 0.0015 \end{bmatrix}, \pi_{j|i_2}^2 = \begin{bmatrix} 0.0637 & 0.0233 & 0.8187 & 0.0001 & 0.0943 \\ 0.1418 & 0.0000 & 0.8430 & 0.0001 & 0.0152 \\ 0.8346 & 0.0000 & 0.0002 & 0.1653 & 0.0000 \\ 0.7924 & 0.0900 & 0.0361 & 0.0005 & 0.0810 \\ 0.9988 & 0.0000 & 0.0001 & 0.0000 & 0.0011 \end{bmatrix}.$$

The beginning profile is supposed to be uniform, that is, $P^l(s_0 = j) = 0.5$ for any player $l = 1, 2$ and its states $j = 1, \dots, 5$. But as it follows from the statements above in the ergodic case this profile can be arbitrarily selected without any influence to the final equilibrium point.

For d^{1*} and d^{2*} (6.3.5) the fixed local-optimal strategies, and k^* the best-reply strategy the following results have been obtained:

For Player 1

$$\pi^{1*} = \begin{bmatrix} 0.0112 & 0.0012 & 0.8514 & 0.0178 & 0.1185 \\ 0.0149 & 0.0000 & 0.9850 & 0.0001 & 0.0000 \\ 0.9869 & 0.0123 & 0.0000 & 0.0006 & 0.0001 \\ 0.6171 & 0.0818 & 0.0755 & 0.0887 & 0.1369 \\ 0.7246 & 0.0000 & 0.2415 & 0.0098 & 0.0241 \end{bmatrix}$$

$$\text{for } n_0 = 1, \chi_{erg}^1 = 0.0112,$$

$$k^{1*} = [2, 2, 1, 1, 2],$$

$$w^{1*} = [0.7932, 0.5939, 0.1941, 0.3237, 0.3516].$$

For Player 2

$$\pi^{2*} = \begin{bmatrix} 0.0637 & 0.0233 & 0.8187 & 0.0001 & 0.0943 \\ 0.1029 & 0.0013 & 0.8957 & 0.0000 & 0.0001 \\ 0.8346 & 0.0000 & 0.0002 & 0.1653 & 0.0000 \\ 0.3733 & 0.0121 & 0.6137 & 0.0001 & 0.0008 \\ 0.9984 & 0.0000 & 0.0001 & 0.0000 & 0.0015 \end{bmatrix}$$

$$\text{for } n_0 = 1, \chi_{erg}^2 = 0.0637,$$

$$k^{2*} = [2, 1, 2, 1, 1],$$

$$w^{2*} = [0.2389, 0.4423, 0.7752, 0.4624, 0.8087].$$

- in Figs. 6.5 and 6.7 the state-value function behavior is shown (where during game repetition the states of the players fluctuate according to the given probabilistic dynamics) showing completely non-monotonic behavior;
- in Figs. 6.6 and 6.8 the corresponding Lyapunov-like functions (6.4.4) are plotted definitely demonstrating a monotonic decreasing behavior;

Fig. 6.5 Non-monotonic behavior of the cost-function for Player 1

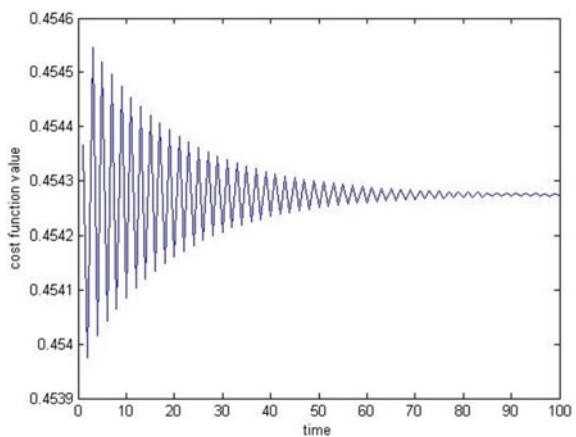


Fig. 6.6 Monotonic behavior of the Lyapunov-like function for Player 1

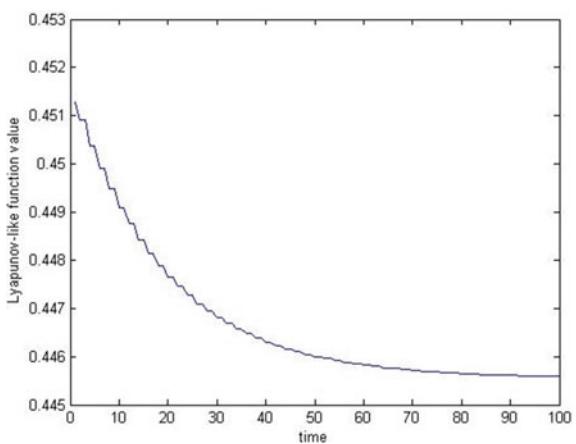


Fig. 6.7 Non-monotonic behavior of the cost-function for Player 2

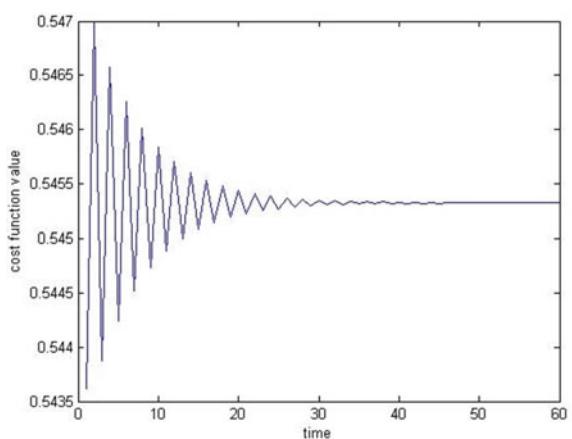
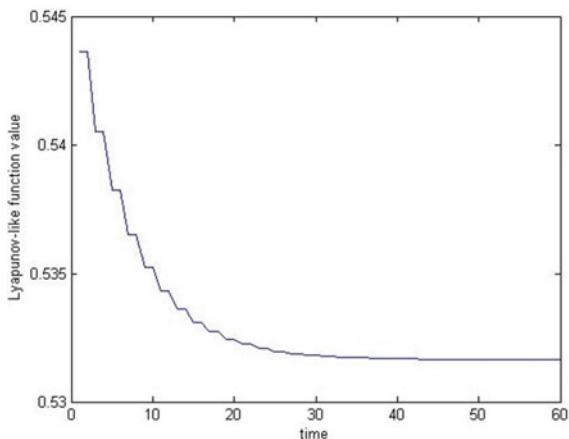


Fig. 6.8 Monotonic behavior of the Lyapunov-like function for Player 2



- the results of the two methods clearly show that under the same fixed local-optimal strategy the original cost functions converge non-monotonically to the values 0.4543 (for the first player) and 0.5453 (for the second player) and the corresponding Lyapunov-like functions converge monotonically to the values 0.4456 and 0.5316, respectively which, obviously, are very close.

As it follows from both examples, the existence of a monotonic decreasing behavior in the constructed Lyapunov-like functions for all players allows to conclude that the considered game has a tendency to evaluate (converge) to an equilibrium point. Conversely, the non-monotonic cost-functions do not permit to get this conclusion: one can not say during the repeating game whether the selected best-reply strategy leads to an equilibrium or not. There also is no guarantee that the non-monotonicity of the cost-functions will converge on an equilibrium point.

From these examples, we also conclude that the proposed method solves a game via the elimination of sequentially unreasonable strategies. A strategy for player l is eventually optimal if, l 's strategy is the optimal strategy among all strategies l could play using a Lyapunov-like function. A strategy for player l is eventually dominant if it is eventually the best-reply against every strategy of the other player. We conclude that the Lyapunov-like function process eliminates strategies that are not the best reply to some strategy profile. Then, there is a best-reply strategy that follows the monotonic decreasing behavior of the Lyapunov-like function which is eventually dominant for the class of rational strategies against regular strategies.

References

1. Bellman, R.: Vector Lyapunov functions. SIAM J. Control Optim. **1**, 32–34 (1962)
2. Bernheim, B.D.: Rationalizable strategic behavior. Econometrica **52**, 1007–1028 (1984)
3. Börgers, T.: Pure strategy dominance. Econometrica **61**, 423–430 (1993)

4. Chen, X., Deng, X.: Setting the complexity of 2-player Nash equilibrium. In: Proceedings of IEEE FOCS (2006)
5. Clempner, J.B.: On Lyapunov game theory equilibrium: static and dynamic approaches. *Int. Game Theory Rev.* **20**(2), 1750033 (2018)
6. Clempner, J.B.: A Lyapunov approach for stable reinforcement learning. *Comput. Appl. Math.* **41**, 279 (2022)
7. Clempner, J.B., Poznyak, A.S.: Convergence method, properties and computational complexity for Lyapunov games. *Int. J. Appl. Math. Comput. Sci.* **21**(2), 349–361 (2011)
8. Clempner, J.B., Poznyak, A.S.: Analysis of best-reply strategies in repeated finite Markov chains games. In: IEEE Conference on Decision and Control (2013)
9. Clempner, J.B., Poznyak, A.S.: Analyzing an optimistic attitude for the leader firm in duopoly models: a strong Stackelberg equilibrium based on a Lyapunov game theory approach. *Econ. Comput. Econ. Cybern. Stud. Res.* **50**(4), 41–60 (2016)
10. Clempner, J.B., Poznyak, A.S.: Convergence analysis for pure stationary strategies in repeated potential games: Nash, Lyapunov and correlated equilibria. *Expert Syst. Appl.* **46**, 474–484 (2016)
11. Clempner, J.B., Poznyak, A.S.: Using the extraproximal method for computing the shortest-path mixed Lyapunov equilibrium in Stackelberg security games. *Math. Comput. Simul.* **138**, 14–30 (2017)
12. Daskalakis, C., Goldberg, P., Papadimitriou, C.: The complexity of computing a Nash equilibrium. In: Proceedings of ACM STOC, pp. 71–78 (2006)
13. Goemans, M., Mirrokni, V., Vetta, A.: Sink equilibria and convergence. In: Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (2005)
14. Guesnerie, R.: Anchoring economic predictions in common knowledge. *Econometrica* **70**, 439–480 (1996)
15. Hernández-Lerma, O., Lasserre, J.B.: Discrete-Time Markov Control Process: Basic Optimality Criteria. Springer, Berlin (1996)
16. Hilas, J., Jansen, M., Potters, J., Vermeulen, D.: Independence of inadmissible strategies and best reply stability: a direct proof. *Int. J. Game Theory* **32**, 371–377 (2003)
17. Hofbauer, J., Sandholm, W.: Stable games and their dynamics. *J. Econ. Theory* **144**(4), 1665–1693 (2009)
18. Moulin, H.: Dominance solvability and Cournot stability. *Math. Social Sci.* **7**, 83–102 (1984)
19. Osborne, M., Rubinstein, A.: A Course in Game Theory. M.I.T. Press, Cambridge, MA (1994)
20. Pearce, D.G.: Raionalizable strategic behavior and the problem of perfection. *Econometrica* **52**, 1029–1050 (1984)
21. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker, New York (2000)
22. Tan, T., Costa Da Werlang, S.R.: On layman's notion of common knowledge, an alternative approach. *J. Econ. Theory* **45**, 370–391 (1988)

Chapter 7

Mechanism Design



Abstract This chapter presents an analytical method for computing Bayesian incentive-compatible mechanisms where the private information is revealed following a class of controllable Markov games. We take into account a dynamic setting where decisions are made after a number of limited time periods. Our approach includes a new variable that denotes the outcome of the distribution vector, the strategies, and the mechanism design. We develop the relationships needed to calculate the relevant variables analytically. The issue becomes computationally tractable with the addition of this variable. The technique uses a Reinforcement Learning (RL) methodology to calculate a mechanism that is nearly optimum and in equilibrium with the game's winning strategy. We employ the Bayesian-Nash equilibrium concept as the default equilibrium idea in the game. There are several equilibria, which presents an intriguing problem because there isn't a single mechanism that is ideal for the goal of profit maximization. To address this issue, we apply Tikhonov's approach and offer a regularization parameter. We show the game's equilibrium and convergence to a single mechanism that is compatible with incentives. This results in numerous game theory issue areas having unique and significantly improved results, as well as incentive-compatible processes that are consistent with the equilibrium of the game. To illustrate the recommended method, we provide a numerical example in the context of a dynamic public finance model with incomplete knowledge.

7.1 Introduction

7.1.1 Brief Review

Informally studied notions for producing specific results in a class of self-interested private information games are known as *mechanism design*. It is an engineering-based approach to resolving problems in game theory. The mechanisms aim to reach a Bayesian-Nash equilibrium and make the assumptions that incentives are provided by monetary transfers and that valuations are private. The players attempt to maximize their own self-interest since they are logical beings. Players who are operating in their own self-interest are not driven to provide accurate information. The presence

of an incentive-compatible mechanism was established by the mechanism design. If a player consistently optimizes his rewards by disclosing his real type, regardless of what the other players do (declare), the mechanism is considered to be honest (subject to an incentive-compatibility limitations). It can be difficult to calculate an efficient process that maximizes rewards. The basic objective is to provide a system for selfish actors that maximizes rewards at equilibrium [5].

There has been a lot of interest in learning how to construct mechanisms and RL approaches during the past ten years. A worldwide difficulty is broken down into smaller, more manageable problems that are handled each time the agents exchange data, according to Goldman and Zilberstein [16]. Jain and Walrandb suggested a method for auctioning collections of various divisible products on a network. Bergemann and Välimäk [2, 6] described a dynamic Vickrey-Clarke-Groves strategy that takes into account quasilinear payoffs in which agents view private information. Pavan et. al. [23] considers dynamical mechanism design in dynamic quasilinear environments where private information arrives over time. Sinha and Anatasopoulos [25] proposed a mechanism design for a network in which fixed groups are formed from strategic agents that are competing for resource allocation. Using extra rewards and punishments based on the actions of the learners, Baumann et al. [4] investigated how an external agent may motivate artificial learners to collaborate. Mguni [20] suggested a method that incorporates stochastic optimization and RL into mechanism design in order to efficiently compute the best incentive compatible mechanisms. Clempner and Poznyak [11] suggested a Bayesian technique for games that expanded the design theory to incorporate mechanism design and joint observer design while accounting for the imperfect information in the Bayesian model and the incomplete knowledge about the states of the Markov system. A Bayes-adaptive RL method based on model-based online planning was proposed by Grover et al. [17]. A BRL model for robots based on an approximative parametric technique, incorporating live Bayesian estimation and planning for an estimated model, was suggested by Senda et al. [24]. Kassab and Simeone [19] suggested a gradient descent method for federated learning in a Bayesian setting. In order to efficiently perform Bayesian estimate, Nolan et al. [22] presented a parameter estimation approach based on a classification task. van Geen and Gerraty [27] devised a technique for obtaining empirical priors using hierarchical Bayesian modeling. A multiscale strategy for model order reduction and machine learning technology were the two separate preconditioning strategies that Vasilyeva et al. [26] provided. Clempner and Poznyak [12] suggested an analytical approach for calculating the mechanism design in the context of a paradigm in which participants in a non-cooperative Markov game with imperfect state knowledge pursue an average utility. For a class of controllable homogeneous Markov games, Clempner [7] developed a dynamic Bayesian-Stackelberg incentive-compatible mechanism in which many agents see private information and learn their behavior through a series of encounters in a repeating game, where it is presumed that leaders may commit to their disclosure approach and method in advance and influence followers' behavior. Clempner [8] presented the Price of Anarchy and the Price of Stability for incomplete information in Bayesian-Markov games. Clempner [13] suggested an analytical method for computing Bayesian incentive-compatible mech-

anisms where the private information is revealed following a class of controllable Markov games. Clempner and Poznyak [15] developed the Price of Anarchy in mechanism design for Pareto-Bayesian-Markov games in which the players privately know their information. The same authors [14] presented a dynamic Bayesian-Stackelberg incentive-compatible mechanism, in which multiple agents observe private information and learn their behavior through a sequence of interactions in a repeated game for a class of controllable homogeneous Markov games.

The main results of this chapter are as follows:

- Suggests the strategy for building an incentive-compatible mechanism in a dynamic environment where participants learn about their preferences via repeated interactions.
- Exposes the private information after a class of controlled Markov games using the analytical approach we suggest for constructing Bayesian incentive-compatible mechanisms .
- Presents a new variable that denotes the outcome of the distribution vector, as well as the strategies, and the mechanism design. We develop the relationships needed to calculate the relevant variables analytically. The problem becomes computationally tractable with the addition of this variable.
- Uses an iterative method divided into two halves: the gradient method and the proximal approach, in order to calculate the equilibrium in Markov games. For our game, the idea of Bayesian-Nash equilibrium serves as the equilibrium concept. With the help of a nonlinear programming solver, this method transforms the game theory issue into a system of equations, which is an independent optimization problem.
- Demonstrates how the strategy will eventually reach the equilibrium point. We also take into account a dynamic setting where participants learn about their preferences via repeated encounters and choose their course of action after a number of limited intervals. The technique uses an RL methodology to compute the nearly optimum mechanism in equilibrium with the ensuing high profit maximization game strategy. In order to compute near-optimal policies, we design a controller exploitation-exploration architecture.
- Employs the controller Kullback-Leibler divergence among the distribution of the policies to trade-off between the exploration and the exploitation processes [1]. Using our method, players adjust their behavior in response to a mechanism that is computed in equilibrium by choosing the best-reply strategies. We demonstrate convergence of the RL technique. There are several equilibria, which presents an intriguing problem because there isn't a single mechanism that is ideal for the goal of profit maximization.
- Applies the regularization Tikhonov's strategy to tackle this issue, which is a well-known method for resolving ill-posed presented minimization problems [9, 10].
- We demonstrate the game's equilibrium as well as the convergence to a singular incentive-compatible mechanism. This generates novel and considerably better answers for many game theory problem areas, as well as incentive-compatible mechanisms that correspond to the game's equilibrium.

- The approach seeks approximate to y by substituting the ill-posed problem given by $\min_{y \in Y_{adm}} \zeta(y)$ introducing a penalized problem given by $\zeta_\delta(y) = \zeta(y) + \frac{\delta}{2} \|y\|^2$ where $\|\cdot\|$ denotes the Euclidean vector norm and the scalar $\delta > 0$ is known as the regularization parameter. The expression $\frac{\delta}{2} \|y\|^2$ penalizes the large values of y . In game theory, regularization plays a fundamental role in order to ensure the convergence to one of the Nash equilibria.

7.2 Description of the Model

We consider a discrete-time repeated game played by a set $\mathcal{N} = \{1, \dots, n\}$ players indexed by $l \in \mathcal{N}$. At each time $t \geq 1$, the player l privately informed about his type (state) $\theta_t^l \in \Theta_t^l$. Players simultaneously take an action (make decisions) $a_t^l \in A_t^l$. Let $A_t = \bigotimes_{l=1}^n A_t^l$. We write Θ_t for $\bigotimes_{l \in \mathcal{N}} \Theta_t^l$, Θ_t^{-l} for $\bigotimes_{h \in \mathcal{N}, h \neq l} \Theta_t^h$ and $\Delta(A)$ for the set of all probability distributions over A . We assume that A_t^l and Θ_t^l are finite sets for all $l \in \mathcal{N}$.

A *social choice function* is a mapping from profile of types to lotteries over alternatives if all players report their type $\theta_t^l \in \Theta_t$, i.e., $f : \Theta_t \rightarrow \Delta(A_t)$. A social choice function represents the goals of the game, e.g., to maximize revenues, etc. Finally, each player has a known *valuation function* $v^l(a_t^l, \theta_t^l) \geq 0$, which determines actual value for the player l . Each $v^l : A \times \Theta \rightarrow \mathbb{R}^+$ is a concave mapping from the alternatives and the set of type profiles to a set of non-negative real numbers. We assume that the type θ_t^l of player l follows a controllable Markov process on the state space Θ_t^l as follows. A *controllable Markov chain* is a sequence of θ -valued random variables θ_t , $t \in T$, satisfying the *Markov condition*:

$$p^l(\theta_{t+1}^l | \theta_t^l, a_t^l, \theta_{t-1}^l, a_{t-1}^l, \dots, \theta_0^l, a_0^l) = p^l(\theta_{t+1}^l | \theta_t^l, a_t^l), \quad (7.2.1)$$

which represents the probability associated with the transition function (or stochastic kernel) from state θ_t^l to state θ_{t+1}^l , under an alternative a_t^l . The common prior (initial distribution) of the state-alternative process for each player l denoted by $\{(\Theta_t, A_t) | t \in T\}$. In this case, the process described by a Markov chain is completely described by the transition function $p^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ and the initial distribution vector $P_0^l(\Delta \Theta_0^l)$ such that $P_0^l(\theta_t^l) \in \Delta \Theta_t^l$, where $\Delta \Theta_t^l$ denotes the set of all probability distributions over Θ_t^l . The Markov chains are mutually independent. We assume that each chain $(P^l, p(\theta_{t+1}^l | \theta_t^l, a_t^l))$ is irreducible and aperiodic, and that P^l is its unique invariant distribution.

Definition 7.1 A *mechanism* μ in the above environment assigns a set of possible messages M_t^l to the player l . At each time t , the player l sends a message m_t^l from this set and the mechanism μ responds with a (possibly randomized) decision that may depend on the entire history of messages sent up to time t , and on past decisions.

Hence, the mechanism $\mu(a_t|m_t)$ has inputs a_t that is the current allocation for the players and $m_t = (m_t^1, \dots, m_t^n)$, which is the joint set of messages made by the player l . Let us consider an *allocation rule* g , which represents the probability measure associated with the occurrence of an alternative a_t from the profile of messages m_t at time $t \in T$. It is a mapping from the message m_t to lotteries over alternatives $\Delta(A_t)$, i.e., $g : M_t \rightarrow \Delta(A_t)$. Formally, a mechanism μ is a pair (M_t, g) where M_t is the set of messages for player l and $g : M_t \rightarrow \Delta(A_t)$ is the allocation rule. Let \mathcal{U} be a set of *admissible mechanisms*. We have that

$$\mathcal{U}_{adm} = \left\{ \mu(a_t|m_t) \geq 0 \mid \sum_{a_t \in A_t} \mu(a_t|m_t) = 1, m_t \in M_t \right\}. \quad (7.2.2)$$

The mechanism μ is interpreted as the probability that a_t will be the outcome if the profile of types θ_t and messages m_t are the players' types. We write

$$\mu(g, M_t^1 \times \dots \times M_t^n) \in \mathcal{U}_{adm}, g(m_t) = \Delta(A_t), \quad (7.2.3)$$

where $m_t \in M_t$.

A dynamic mechanism induces a dynamic game with incomplete information. The following sequence of events takes place in each period t :

- At the beginning of each period t , the designer announces publicly the mechanism to the players and each agent privately learns his current type $\theta_t^l \in \Theta_t^l$ drawn from $p^l(\theta_{t+1}^l|\theta_t^l, a_t^l)$.
- Next, players sent messages m_t^l simultaneously, and the profile of messages is publicly observed m_t .
- The mechanism selects a decision $a_t \in A_t$ according to $\mu(a_t|m_t)$ and the implemented alternative a_t is publicly announced.

With this assumption, the public history in period t is a sequence of messages m_t^l and alternatives $a_t(\theta_t)$ until period $t - 1$. The state of an observer is the vector

$$h_t^l = (m_0^l, a_0^l, \dots, m_{t-1}^l, a_{t-1}^l, m_t^l), \quad (7.2.4)$$

which is a trajectory of length t called the *history* (public history). The public history h_t^l stands for a generic element of \mathbb{H}_t , which is the set of possible public histories in period t :

- (i) \mathbb{H}_t is the set of possible states in period t , captures all information relevant to the decision by the observer in that period, and
- (ii) each $m_t = (m_0^l, \dots, m_t^l)$ is a report profile of the players l .

The sequence of reports by the players is part of the public history and we assume that the past reports of each player are observable to all the players. The *private history* of player l in period t consists of the sequence of private observations and the public history until period t ,

$$\tilde{h}_t^l = (\theta_0^l, m_0^l, a_0^l, \dots \theta_{t-1}^l, m_{t-1}^l, a_{t-1}^l, \theta_t^l). \quad (7.2.5)$$

The set of possible private histories for each player l in period t is denoted by $\mathcal{H}_t^l = \Theta_t \times \mathbb{H}_t$.

Definition 7.2 A (behavioral) *strategy* $\sigma^l(m_t^l | \theta_t^l)$ for player l is a mapping $\sigma^l : \mathcal{H}^l \times \Theta^l \rightarrow \Delta(M^l)$. The set of all admissible policies is denoted by \mathcal{S}_{adm}^l :

$$\mathcal{S}_{adm}^l = \left\{ \sigma^l(m_t^l | \theta_t^l) \geq 0 \mid \sum_{m_t^l \in M_t^l} \sigma^l(m_t^l | \theta_t^l) = 1, \theta_t^l \in \Theta_t^l \right\}. \quad (7.2.6)$$

The *valuation functions* $v^l(a_t^l, \theta_t^l)$ and the *transition functions* $p^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ are all common knowledge at time t . The common prior initial distribution vector $P_0^l(\Delta \Theta_0^l)$ and the transition function $p^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ are assumed to be independent across players. The interaction between players induces a Markov game given by the 5-tuple

$$\Gamma = (\mathcal{N}, \Theta^l, A^l, P^l, U^l)_{l \in \mathcal{N}},$$

where

$$\begin{aligned} U_T^l(\mu, \sigma) &= \sum_{t \in T} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} v^l(a_t^l(\theta_t), \theta_t^l) p^l(\theta_{t+1}^l | \theta_t^l, a_t^l) \prod_{\nu \in \mathcal{N}} \mu(a_\nu | m_\nu) \sigma^l(m_\nu^l | \theta_\nu^l) P^\nu(\theta_\nu^l) = \\ &\sum_{t \in T} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} W^l(a_t^l, \theta_t^l) \prod_{\nu \in \mathcal{N}} \mu(a_\nu | m_\nu) \sigma^l(m_\nu^l | \theta_\nu^l) P^\nu(\theta_\nu^l). \end{aligned}$$

Here

$$W^l(\theta_t^l, a_t^l) = \sum_{\theta_{t+1}^l \in \Theta_t^l} v^l(a_t^l(m_t^l), \theta_t^l) p^l(\theta_{t+1}^l | \theta_t^l, a_t^l).$$

7.3 Mechanism and Equilibrium

We assume that players know their payoffs. A mechanism $\mu(a_t | m_t)$ and the strategy $\sigma^l(m_t^l | \theta_t^l)$ maximize the payoff function $U^l(\mu, \sigma)$ realizing the rule given by

$$(\mu^*, \sigma^*) := \arg \max_{\mu \in \mathcal{U}_{adm}} \max_{\sigma \in \mathcal{S}_{adm}} \sum_{l \in \mathcal{N}} U^l(\mu, \sigma). \quad (7.3.1)$$

The mechanism μ^* and the strategy σ^* satisfies the *Bayesian-Nash equilibrium* fulfilling for all σ the condition

$$U^l(\mu^*, \sigma^*) \geq U^l(\mu, \sigma^l, \sigma^{-l}). \quad (7.3.2)$$

where the **mechanism μ is unique for all participants** and the strategy $\sigma^* = (\sigma^{1*}, \dots, \sigma^{n*})$ is referred to as a *Bayesian-Nash equilibrium* where σ is such that $\sigma^{-l*} = (\sigma^{1*}, \dots, \sigma^{l-1*}, \sigma^{l+1,*}, \dots, \sigma^{n*})$.

The dynamic revelation principle in [18, 21] shows that there is no loss of generality in restricting attention to direct mechanisms where the players report their information truthfully on the equilibrium path: “for any equilibrium of any other coordination game, which the individuals might play, there exists an equivalent incentive-compatible mechanism.” This idea, called the *revelation principle* showed that direct mechanisms are the same as indirect mechanisms ($g(a_t|m_t)) = f(a_t|\theta_t)$).

Let us introduce the ξ -variable as follows

$$\xi^l(\theta_t^l m_t^l a_t^l) := \mu(a_t|m_t) \sigma^l(m_t^l|\theta_t^l) P^l(\theta_t^l). \quad (7.3.3)$$

Formulation of the problem. We will try to find an auxiliary variable $\xi^l(\theta_t^l m_t^l a_t^l)$ which solve the following individual nonlinear programming problem

$$\boxed{\begin{aligned} U^l(\mu, \sigma) &= \sum_{l \in \mathcal{N}} \bar{U}^l(\xi) \rightarrow \max_{\xi \in \Xi_{adm}}, \\ \bar{U}^l(\xi) &= \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} W^l(a_t^l, \theta_t^l) \prod_{i \in \mathcal{N}} \xi^i(\theta_i^l m_i^l a_i^l), \end{aligned}} \quad (7.3.4)$$

where $\xi^l(\theta_t^l m_t^l a_t^l)$ is given in Eq. (7.3.3) and $\Xi_{adm} = \bigotimes_l \Xi_{adm}^l$ with

$$\begin{aligned} \Xi_{adm}^l &:= \left\{ \xi^l(\theta_t^l m_t^l a_t^l) \left| \begin{array}{l} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) = 1, \\ \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) = P^l(\theta_t^l) > 0, \\ \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} [\delta_{\theta_t^l \theta_{t+1}^l} - p^l(\theta_{t+1}^l | \theta_t^l a_t^l)] \xi^l(\theta_t^l m_t^l a_t^l) = 0, \theta_{t+1}^l \in \Theta_{t+1}^l \end{array} \right. \right\}. \end{aligned} \quad (7.3.5)$$

Note that the following relations holds.

$$\sum_{a_t^l \in A_t^l} \mu(a_t|m_t) = 1, \quad \sum_{m_t \in M_t} \sigma(m_t|\theta_t) = 1, \quad \sum_{\theta_t^l \in \Theta_t^l} P^l(\theta_t^l) = 1.$$

It is easy to check that $\xi^l \in \Delta^l$, where

$$\Delta^l := \left\{ \xi^l(\theta_t^l m_t^l a_t^l) \left| \begin{array}{l} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) = 1, \\ \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) = P^l(\theta_t^l) > 0, \end{array} \right. \right\}. \quad (7.3.6)$$

Define the solution of the problem (7.3.4) as ξ^l . The next theorem and lemma clarify how we may recover $\mu^*(a_t|m_t)$, $\sigma^{l*}(m_t^l|\theta_t^l)$ and $P^{l*}(\theta_t^l)$.

Theorem 7.1 *The strategy $\sigma^{l*}(m_t^l|\theta_t^l)$ and the mechanism $\mu^*(a_t|m_t)$ are in a Bayesian-Nash equilibrium where every agent maximizes its expected payoff, for every $l \in \mathcal{N}$,*

$$U^l(\mu^*(a_t|m_t), \sigma^*(m_t^l|\theta_t^l)) \geq U^l(\mu(a_t|m_t), \sigma(m_t^l|\theta_t^l)),$$

if the quantities of $\xi^l(\theta_t^l m_t^l a_t^l)$ satisfies

$$\sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \sum_{\theta_t^l \in \Theta_t^l} [\delta_{\theta_t^l \theta_{t+1}^l} - p^l(\theta_{t+1}^l | \theta_t^l a_t^l)] \xi^l(\theta_t^l m_t^l a_t^l) = 0, \quad \theta_{t+1}^l \in \Theta_{t+1}^l. \quad (7.3.7)$$

Proof Let $\xi^l := [\xi^l(\theta_t^l m_t^l a_t^l)]$ as in (7.3.3) and $\tilde{U}^l(\xi)$ as in (7.3.4). Suppose that the set of strategies \mathcal{E}_{adm}^l satisfies that:

- (a) each $\xi^l(\theta_t^l m_t^l a_t^l)$ represents a mixed joint strategy that belongs to the simplex Δ defined by

$$\Delta = \left\{ \xi^l(\theta_t^l m_t^l a_t^l) \left| \begin{array}{l} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) = 1, \quad \xi^l(\theta_t^l m_t^l a_t^l) \geq 0, \\ \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) > 0 \end{array} \right. \right\},$$

- (b) the joint strategy $\xi^l(\theta_t^l m_t^l a_t^l)$ fulfill the ergodicity constraint and, then it belongs to the convex, closed and bounded set given by

$$\mathcal{E}^l = \left\{ \xi^l(\theta_t^l m_t^l a_t^l) \left| \begin{array}{l} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} p^l(\theta_{t+1}^l | \theta_t^l a_t^l) \xi^l(\theta_t^l m_t^l a_t^l) \\ - \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_{t+1}^l m_t^l a_t^l) = 0, \quad \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) > 0 \end{array} \right. \right\}.$$

Then, $\xi^l(\theta_t^l m_t^l a_t^l) \in \mathcal{E}_{adm}^l := \Delta^l \times \mathcal{E}^l$, and we have that the ergodicity constraints defined in (7.3.7) satisfies

$$\begin{aligned} \sum_{m_t^l \in M_t^l} \left(\sum_{a_t^l \in A_t^l} \sum_{\theta_t^l \in \Theta_t^l} [p^l(\theta_{t+1}^l | \theta_t^l a_t^l) \xi^l(\theta_t^l m_t^l a_t^l) - \xi^l(\theta_{t+1}^l m_t^l a_t^l)] \right) = \\ \sum_{m_t^l \in M_t^l} \left(\sum_{a_t^l \in A_t^l} \sum_{\theta_t^l \in \Theta_t^l} [p^l(\theta_{t+1}^l | \theta_t^l a_t^l) - \delta_{\theta_t^l \theta_{t+1}^l}] \right) \xi^l(\theta_t^l m_t^l a_t^l) = 0. \end{aligned}$$

□

Lemma 7.1 *Let us suppose that the problem (7.3.4) is solved, then the variable $\mu^*(a_t|m_t)$ can be recovered from $\xi^l(\theta_t^l m_t^l a_t^l)$ as follows:*

$$\boxed{\mu^*(a_t|m_t) = \frac{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^l(\theta_t^l m_t^l a_t^l)}{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l)}} \quad (7.3.8)$$

Proof The mechanism design $\mu^*(a_t|m_t)$ may be obtained from Eqs. (7.3.5) and (7.3.6) as follows:

$$\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^l(\theta_t^l m_t^l a_t^l) := \mu^*(a_t|m_t) \sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sigma^{l*}(m_t^l | \theta_t^l) P^{l*}(\theta_t^l).$$

Hence,

$$\mu^*(a_t|m_t) = \frac{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)}{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sigma^{l*}(m_t^l | \theta_t^l) P^{l*}(\theta_t^l)} = \frac{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)}{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)}.$$

To verify that the definition of $\mu^*(a_t|m_t)$ is correct we need to check the fulfilling of Eq. (7.2.2), i.e. $\mu^*(a_t|m_t) \in \mathcal{U}_{adm}$. This property holds directly

$$\mu^*(a_t|m_t) = \frac{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)}{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sigma^{l*}(m_t^l | \theta_t^l) P^{l*}(\theta_t^l)} = \frac{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)}{\sum_{l \in \mathcal{N}} \sum_{\theta_t^l \in \Theta_t^l} \sum_{a_t^l \in A_t^l} \xi^{l*}(\theta_t^l m_t^l a_t^l)} \geq 0. \quad (7.3.9)$$

since $\xi^{l*}(\theta_t^l m_t^l a_t^l) \geq 0$. Summing (7.3.9) by a_t^l directly leads to the property $\sum_{a_t \in A_t} \mu^*(a_t|m_t) = 1$. \square

In order to recover $\sigma^{l*}(m_t^l | \theta_t^l)$, we have that for each player $l = \overline{1, n}$ the quantity of interest is given by

$$\boxed{\sigma^{l*}(m_t^l | \theta_t^l) = \frac{\sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l)}{\sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l)}} \quad (7.3.10)$$

As well as, for distribution $P^{l*}(\theta_t^l)$ we have

$$\boxed{P^{l*}(\theta_t^l) = \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} \xi^l(\theta_t^l m_t^l a_t^l) > 0.} \quad (7.3.11)$$

We have derived the formulas, which solve the problem (7.3.4) based on the variables $\xi^l(\theta_t^l m_t^l a_t^l)$ and the formulas to recover the strategy $\sigma^{l*}(m_t^l | \theta_t^l)$, the mechanism $\mu^*(a_t|m_t)$ and the distribution $P^{l*}(\theta_t^l)$. If the players report their type using some reporting strategy $\sigma^l(m_t^l | \theta_t^l)$, they are maximizing the expected payoff given in problem (7.3.4). The resulting strategy profile $\sigma^{l*}(m_t^l | \theta_t^l)$ is a Bayesian-Nash equilibrium.

A mechanism $\mu^*(a_t|m_t)$ is said to be *incentive compatible* if the relation (7.3.1) is satisfied.

Lemma 7.2 *The obtained mechanism $\mu^*(a_t|m_t)$ and the strategies $\sigma^{l*}(m_t^l|\theta_t^l)$ satisfy the Bayesian-Nash equilibrium defines by Eq. (7.3.2).*

Proof It results straightforward from:

$$\begin{aligned} \max_{\xi \in \mathcal{S}_{adm}} \bar{U}(\xi) &= \bar{U}(\xi^*) = \sum_{l \in \mathcal{N}} \bar{U}^l(\xi^*) = \sum_{l \in \mathcal{N}} U^l(\mu^*, \sigma^*) = \\ &\sum_{l \in \mathcal{N}} \left(\sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} W^l(a_t^l, \theta_t^l, m_t^l) (\mu^*(a_t|m_t))^n \sigma^{l*}(m_t^l|\theta_t^l) P^{l*}(\theta_t^l) \cdot \right. \\ &\left. \prod_{i \neq l \in \mathcal{N}} \sigma^{i*}(m_t^{i*}|\theta_t^{i*}) P^{i*}(\theta_t^{i*}) \right) = \sum_{l \in \mathcal{N}} \left(\max_{\sigma^l \in \mathcal{S}'_{adm}} \sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} (\mu^*(a_t|m_t))^n \cdot \right. \\ &\left. \sum_{\theta_t^l \in \Theta_t^l} W^l(a_t^l, \theta_t^l, m_t^l) \sigma^l(m_t^l|\theta_t^l) P^l(\theta_t^l) \prod_{i \neq l \in \mathcal{N}} \sigma^{i*}(m_t^{i*}|\theta_t^{i*}) P^{i*}(\theta_t^{i*}) \right) \geq \\ &\sum_{l \in \mathcal{N}} \left(\sum_{m_t^l \in M_t^l} \sum_{a_t^l \in A_t^l} (\mu^*(a_t|m_t))^n \sum_{\theta_t^l \in \Theta_t^l} W^l(a_t^l, \theta_t^l, m_t^l) \sigma^l(m_t^l|\theta_t^l) P^l(\theta_t^l) \cdot \right. \\ &\left. \prod_{i \neq l \in \mathcal{N}} \sigma^i(m_t^i|\theta_t^i) P^i(\theta_t^i) \right) = \sum_{l \in \mathcal{N}} U^l(\mu, \sigma^l, \sigma^{-l}) \end{aligned} \quad (7.3.12)$$

From this inequality it follows that

$$\sum_{l \in \mathcal{N}} (U^l(\mu^*, \sigma^*) - U^l(\mu^*, \sigma^l, \sigma^{-l})) \geq 0. \quad (7.3.13)$$

Given that the inequality in Eq. (7.3.13) is valid for all admissible strategies σ , it is valid when $\sigma^j = \sigma^{j*}$ for $j \neq l$, having

$$U^l(\mu^*, \sigma^*) - U^l(\mu^*, \sigma^l, \sigma^{-l}) \geq 0, \quad (7.3.14)$$

which coincides with Eq. (7.3.2) when $\mu = \mu^*$. Lemma is proven. \square

7.4 Reinforcement Learning Approach

In this section, we are interested in the problem of finding a strategy and mechanism that maximizes the expected reward of the players that chooses actions sequentially. We suggest an asymptotic algorithm of the controlled Markov game based on a stochastic approximation method [1]. This method consists of constructing a recurrent procedure which produces the randomized Markov laws such that the expected reward takes its extreme value in the limit ($t \rightarrow \infty$).

Formally, we consider a discrete-time $t \in T$ repeated game played by a set $\mathcal{N} = \{1, \dots, n\}$ players indexed by $l \in \mathcal{N}$. Let us denote \mathcal{H}^∞ the set of complete infinite histories, i.e., the set of sequences of realized types, mechanisms, messages and alternatives. For an history $h \in \mathcal{H}^\infty$ and $(\theta_t^l, a_t^l) \in \Theta_t^l \times A_t^l$ its corresponding finite sequence of realized types and actions and let us denote for each $(\theta_{t+1}^l, \theta_t^l, a_t^l) \in \Theta_t^l \times \Theta_t^l \times A_t^l$

$$\eta^l(\hat{\theta}_{t+1}^l \hat{\theta}_t^l a_t^l) = \sum_{t \in T} \chi(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l)$$

the *average discounted number of times along the history h* where the realized type profile $\hat{\theta}$ implement the action a , as well as, let us denote

$$\eta^l(\hat{\theta}_t^l a_t^l) = \sum_{t \in T} \chi(\hat{\theta}_t^l = \hat{\theta}^l, a_t^l = a^l)$$

the average discounted number of times along the history h where the realized type profile is $\hat{\theta}_t^l$ implement the action a_t such that the indicator function $\chi(\cdot)$ is defined as follows:

- (i) $\chi(e_t) = 1$ if the event e occurs at period t , and
- (ii) $\chi(\cdot) = 0$ otherwise.

We have that

$$\tilde{p}^l(\hat{\theta}_{t+1}^l | \hat{\theta}_t^l, a_t^l) = \frac{\eta^l(\hat{\theta}_{t+1}^l \hat{\theta}_t^l a_t^l)}{\eta^l(\hat{\theta}_t^l a_t^l)}. \quad (7.4.1)$$

The Eq. (7.4.1) denotes the number of periods t where x is implemented considering the periods where the type profile is θ . We can write the discounted payoff of player l as a function of the discounted measure $\eta^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l)$ of states and outcomes,

$$\tilde{v}^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l) = \frac{\sum_{t=0}^T \zeta^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l) \chi(\hat{\theta}_{t+1}^l \hat{\theta}_t^l a_t^l)}{\eta^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l)}, \quad (7.4.2)$$

where $\zeta^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l) = v^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l) + \zeta^l r$ such that $\zeta^l \leq U^l$ and r takes randomly the values -1 or 1 . We estimate $\tilde{v}^l(\hat{\theta}_{t+1}^l, \hat{\theta}_t^l, a_t^l)$ in a finite number of steps.

Let us consider a game whose strategies are denoted by $\psi^l \in \Psi^l$ where Ψ is a convex and compact set where

$$\psi^l := \text{col } (\xi^l(\theta_t^l m_t^l a_t^l)), \Psi^l := \Xi_{adm}^l, \Psi := \bigotimes_{l=1}^n \Psi^l. \quad (7.4.3)$$

Denote by $\psi = (\psi^1, \dots, \psi^n)^\top \in \Psi$ the joint strategy and ψ^l is a strategy of the rest of the players adjoint to ψ^l ,

$$\psi^{\hat{l}} := (\psi^1, \dots, \psi^{l-1}, \psi^{l+1}, \dots, \psi^n)^{\top} \in \Psi^{\hat{l}} := \bigotimes_{m=1, m \neq l}^n \Psi^m$$

such that $\psi = (\psi^l, \psi^{\hat{l}})$ ($l = \overline{1, n}$).

The method of Lagrange multipliers is an approach for finding the local maximum of a function subject to equality constraints given in Eq. (7.4.3). For regularization see [9, 10]. Let us consider the Lagrange function given by

$$\mathbb{L}(\psi, \hat{\psi}(\psi), \lambda) := \alpha G(\psi, \hat{\psi}(\psi)) - \lambda^T \Phi_{eq} \psi - \frac{\delta}{2} \left(\|\psi\|^2 + \|\hat{\psi}\|^2 - \|\lambda\|^2 \right), \quad (7.4.4)$$

where

$$G(\psi, \hat{\psi}(\psi)) := \sum_{l=1}^n \left[u_l \left(\psi^l, \psi^{\hat{l}} - u_l \left(\hat{\psi}^l, \psi^{\hat{l}} \right) \right) \right] \quad (7.4.5)$$

and

$$\hat{\psi}^l := \arg \min_{\psi^l \in \Psi^l} u_l \left(\psi^l, \psi^{\hat{l}} \right), \quad (7.4.6)$$

Φ_{eq} is the restriction matrix of the Markov game, and the parameters α and δ are positive and the Lagrange vector-multipliers $\lambda \in \Lambda$ may have any sign. Here $u_l(\psi^l, \psi^{\hat{l}})$ is the payoff-function of the player l which plays the strategy $\psi^l \in \Psi^l$ and the rest of the players the strategy $\psi^{\hat{l}} \in \Psi^{\hat{l}}$. The optimization problem

$$\mathbb{L}_{\alpha, \delta}(\psi, \hat{\psi}(\psi), \lambda) \rightarrow \max_{\psi \in \Psi_{adm}, \hat{\psi}(\psi) \in \hat{\Psi}_{adm}} \min_{\lambda \in \Lambda} \quad (7.4.7)$$

has a unique saddle-point on ψ since the optimized function (7.4.4) is *strongly concave on ψ* if

$$\frac{\partial^2}{\partial \psi \partial \psi^T} \mathbb{L}_{\alpha, \delta}(\psi, \hat{\psi}(\psi), \lambda) < 0, \forall \psi \in \Psi_{adm}, \hat{\psi}(\psi) \in \hat{\Psi}_{adm} \subset \mathbb{R}^n, \quad (7.4.8)$$

and is *strongly convex* on the Lagrange multipliers λ for any $\delta > 0$.

We shall refer to $(\psi^*(\delta), \hat{\psi}^*(\delta), \lambda^*(\alpha, \delta))$ of this set as the saddle point of the convex-concave Lagrangian function $\mathbb{L}_{\alpha, \delta}(\psi(\delta), \hat{\psi}(\delta), \lambda(\alpha, \delta))$ on the set $\Psi^* \times \hat{\Psi}^* \times \Lambda^*$, which for all $(\psi(\delta), \hat{\psi}(\delta), \lambda(\alpha, \delta)) \in \Psi^* \times \hat{\Psi}^* \times \Lambda^*$ satisfies the system of inequalities

$$\mathbb{L}_{\alpha, \delta}(\psi(\delta), \hat{\psi}(\delta), \lambda^*(\alpha, \delta)) \leq \mathbb{L}_{\alpha, \delta}(\psi^*(\delta), \hat{\psi}^*(\delta), \lambda^*(\alpha, \delta)) \leq \mathbb{L}_{\alpha, \delta}(\psi^*(\delta), \hat{\psi}^*(\delta), \lambda)$$

for any $\lambda \in \Lambda$ with non-negative components and, any $\psi \in \Psi_{adm}$ and $\hat{\psi} \in \hat{\Psi}_{adm}$. Using the estimated values, we propose to update the strategy and the mechanism iteratively following the extragradient method given by:

1. Proximal prediction step:

$$\left. \begin{aligned} \bar{\lambda}_n &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathbb{L}_{\alpha, \delta}(\psi_n, \hat{\psi}_n, \lambda) \right\}, \\ \bar{\psi}_n &= \arg \max_{\psi \in \Psi} \left\{ \frac{1}{2} \|\psi - \psi_n\|^2 + \gamma \mathbb{L}_{\alpha, \delta}(\psi, \hat{\psi}_n, \bar{\lambda}_n) \right\}, \\ \hat{\bar{\psi}}_n &= \arg \max_{\hat{\psi} \in \hat{\Psi}} \left\{ \frac{1}{2} \|\hat{\psi} - \hat{\psi}_n\|^2 + \gamma \mathbb{L}_{\alpha, \delta}(\psi_n, \hat{\psi}, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (7.4.9)$$

2. Gradient approximation step:

$$\left. \begin{aligned} \lambda_{n+1} &= \lambda_n - \gamma \nabla_{\lambda} \mathbb{L}_{\alpha, \delta}(\bar{\psi}_n, \hat{\bar{\psi}}_n, \bar{\mu}_n, \lambda), \\ \psi_{n+1} &= \text{Pr}_{\psi \in \Psi} \left\{ \psi_n + \gamma \nabla_{\psi} \mathbb{L}_{\alpha, \delta}(\psi, \bar{\psi}_n, \bar{\lambda}_n) \right\}, \\ \hat{\psi}_{n+1}(\psi) &= \text{Pr}_{\hat{\psi} \in \hat{\Psi}} \left\{ \hat{\psi}_n + \gamma \nabla_{\hat{\psi}} \mathbb{L}_{\alpha, \delta}(\bar{\psi}_n, \hat{\psi}, \bar{\lambda}_n) \right\}, \end{aligned} \right\} \quad (7.4.10)$$

where Pr is the projection operator. Under proper conditions the extragradient method can be shown to converge to the local maximum (minimum).

Algorithm 1: Learning Processes

Let $\mathcal{N} = \{1, \dots, n\}$ be the set of players ($l \in \mathcal{N}$).

Let $\theta_0^l = \theta^l$ be the initial state for player l .

Let $\hat{p}^l = p^l$ be the initial transition matrix for player l .

Let $\varepsilon > 0$ be the error of the estimated parameters.

do

Compute the strategy $\sigma^{l*}(m_t^l | \theta_t^l)$ and the mechanism $\mu^*(a_t | m_t)$ by applying Eqs. (7.4.9) and (7.4.10).

Select randomly a message m_t^l from the $\sigma^{l*}(m_t^l | \theta_t^l)$.

Select randomly an action a_t from the $\mu^*(a_t | m_t)$.

From the transition matrix $\hat{p}^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ get next state θ_{t+1}^l for each player l .

Increase the values of $\eta^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ and $\eta^l(\theta_t^l | a_t^l)$.

Compute the error mean square error e_t^l .

Estimate $\hat{p}^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ given in Eq. (7.4.1).

Estimate $\hat{\psi}^l(\theta_{t+1}^l | \theta_t^l, a_t^l)$ given in Eq. (7.4.2).

Update $\theta_t^l = \theta_{t+1}^l$ and increase t by 1 ($t = t + 1$).

Until for each player $\varepsilon > e_t^l$ (by Kullback-Leibler)

In the Algorithm 1, the initial time is $t = 0$ and for each player l we let $\theta_0^l = \theta^l$ be the initial state and $\hat{p}^l = p^l$ the initial transition matrix. We fix the error of the estimated parameters by $\varepsilon > 0$. For each iteration, we compute the strategy $\sigma^{l*}(m_t^l | \theta_t^l)$ and the mechanism $\mu^*(a_t^l | m_t^l)$ using Eqs. (7.4.9) and (7.4.10). Next, we select randomly a message m_t^l from the strategy $\sigma^{l*}(m_t^l | \theta_t^l)$ and action a_t^l from the mecha-

nism $\mu^*(a_t^l|m_t^l)$. Then, the system advance by obtaining the state θ_{t+1}^l randomly from $\hat{p}^l(\theta_{t+1}^l|\theta_t^l a_t^l)$ for fixed state θ_t^l and action a_t^l . Next, we update the values of $\eta^l(\theta_{t+1}^l|\theta_t^l a_t^l)$ and $\eta^l(\theta_t^l a_t^l)$. So, we get the estimate of $\hat{p}^l(\theta_{t+1}^l|\theta_t^l a_t^l)$ and $\hat{v}^l(\theta_{t+1}^l|\theta_t^l a_t^l)$ using the estimation rules in Eqs. (7.4.1) and (7.4.2), respectively. Finally, we compute the mean square error $e(t)$ by using

$$e_t^l = \sum_{a_t^l \in A_t^l} ((\hat{p}_{t-1}^l - \hat{p}_t^l)^\top (\hat{p}_{t-1}^l - \hat{p}_t^l)). \quad (7.4.11)$$

The process continue until $\varepsilon > e_t^l$. At the end, we obtain the estimated transition matrices $\hat{p}^l(\theta_{t+1}^l|\theta_t^l a_t^l)$, the estimated values of $\hat{v}^l(\theta_{t+1}^l|\theta_t^l a_t^l)$. The resulting strategy profile $\sigma^*(m_t^l|\theta_t^l)$ is a Bayesian-Nash equilibrium. The resulting mechanism $\mu^*(a_t^l|m_t^l)$ is incentive compatible because it is in equilibrium and satisfies Eq. (7.3.1).

7.5 Risk-Averse Agents Strategies in Contracting Problem

We now show how our method can be used to tackle the contracting problem [3]. In dynamics mechanism design, the main problem for the designer is how to design the most advantageous way of providing the agent with a fixed utility u . With risk-averse agents and a risk-neutral principal, optimal contracts provide some amount of insurance, but incentive compatibility does not provide full insurance. Particularly, we focus on the case of incomplete information where an expected-truthful mechanism is considered as a mechanism that is dominant-strategy incentive compatible for risk-averse agents. In this case, at each time t a risk-averse agent facing a payoff u can contract with a risk-neutral principal. The way of providing dynamic incentives to the agent takes the form $(\mathbb{E}\{u\} - u)$. This form preserves the player's expected payoff and eliminates all associated risk.

The setting consists of the set $\mathcal{N} = \{1, 2, 3\}$ of players ($l \in \mathcal{N}$), and the expected payoff takes the form

$$U^l(\mu, \sigma) = \sum_{t \in T} \sum_{\theta_t^l \in \Theta_t^l} \sum_{m_t^l \in M_t} \sum_{a_t^l \in A_t^l} W^l(a_t^l, \theta_t^l) \prod_{\iota \in \mathcal{N}} \mu(a_t^l | \rho_\iota) \sigma^\iota(m_t^\iota | \theta_t^\iota) P^\iota(\theta_t^\iota).$$

We fix the values of interest $\gamma_0 = 0.1025$, $\delta_0 = 5.0 \times 10^{-4}$ $|\Theta^l| = 4$ and $|A^l| = 2$. The convergence of the strategies is shown in Figs. 7.1, 7.2 and 7.3. Applying our method, we obtain the resulting values of interest:

Fig. 7.1 Convergence of strategies $\xi^1(\theta^1 m^1 a^1)$ for Player 1

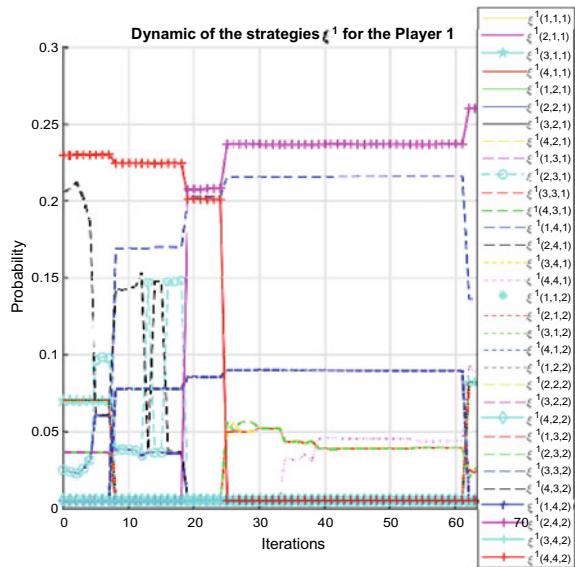


Fig. 7.2 Convergence of strategies $\xi^2(\theta^2 m^2 a^2)$ for Player 2

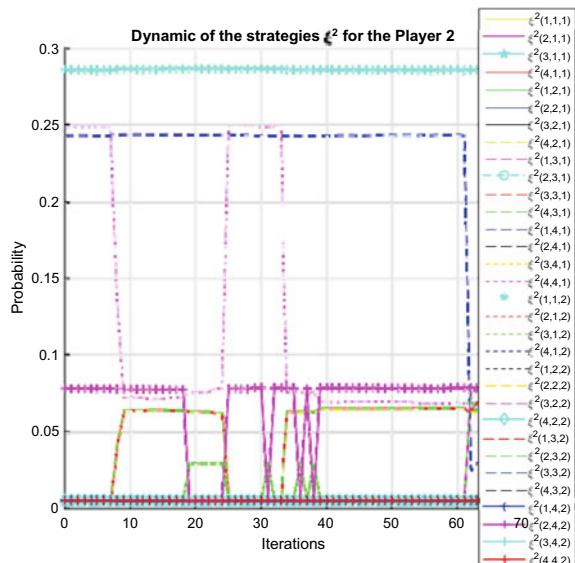
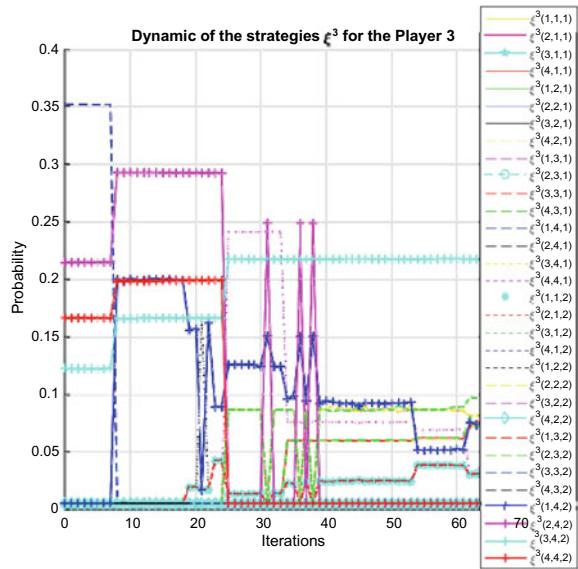


Fig. 7.3 Convergence of strategies $\xi^3(\theta^3 m^3 a^3)$ for Player 3



$$\mu^*(a|m) = \begin{bmatrix} 0.6843 & 0.3157 \\ 0.6709 & 0.3291 \\ 0.6838 & 0.3162 \\ 0.3313 & 0.6687 \end{bmatrix},$$

$$\sigma^{1*}(m|\theta) = \begin{bmatrix} 0.0584 & 0.0584 & 0.0584 & 0.8247 \\ 0.0340 & 0.0340 & 0.0340 & 0.8981 \\ 0.2504 & 0.2504 & 0.2504 & 0.2487 \\ 0.1671 & 0.1538 & 0.1764 & 0.5027 \end{bmatrix}, \quad P^{1*}(\theta) = \begin{bmatrix} 0.1712 \\ 0.2955 \\ 0.3490 \\ 0.1843 \end{bmatrix},$$

$$\sigma^{2*}(m|\theta) = \begin{bmatrix} 0.2921 & 0.2921 & 0.2921 & 0.1237 \\ 0.0886 & 0.0886 & 0.0886 & 0.7342 \\ 0.0321 & 0.0321 & 0.0321 & 0.9037 \\ 0.2416 & 0.2640 & 0.2394 & 0.2549 \end{bmatrix}, \quad P^{2*}(\theta) = \begin{bmatrix} 0.2790 \\ 0.1138 \\ 0.3222 \\ 0.2850 \end{bmatrix},$$

$$\sigma^{3*}(m|\theta) = \begin{bmatrix} 0.1928 & 0.1928 & 0.1928 & 0.4216 \\ 0.3034 & 0.3566 & 0.3048 & 0.0352 \\ 0.0399 & 0.0399 & 0.0399 & 0.8802 \\ 0.2757 & 0.2966 & 0.2714 & 0.1563 \end{bmatrix}, \quad P^{3*}(\theta) = \begin{bmatrix} 0.1859 \\ 0.2851 \\ 0.2526 \\ 0.2763 \end{bmatrix}.$$

This approach presented a dynamical incentive-compatible mechanism for players with risk-averse reward, which results in no loss in the risk-neutral principal's utility, and only increases agents' utilities.

References

1. Asiain, E., Clempner, J.B., Poznyak, A.S.: Controller exploitation-exploration: a reinforcement learning architecture. *Soft Comput.* **23**(11), 3591–3604 (2019)
2. Athey, S., Segal, I.: An efficient dynamic mechanism. *Econometrica* **81**(6), 2463–2485 (2013)
3. Battaglini, M.: Long-term contracting with Markovian consumers. *Am. Econ. Rev.* **95**(3), 637–658 (2005)
4. Baumann, T., Graepel, T., Shawe-Taylor, J.: Adaptive mechanism design: learning to promote cooperation (2019). [ArXiv:1806.04067](https://arxiv.org/abs/1806.04067), v2
5. Bergemann, D., Said, M.: Wiley encyclopedia of operations research and management science, chap. Dynamic Auctions, pp. 1511–1522. Wiley, Hoboken, NJ (2011)
6. Bergemann, D., Välimäki, J.: The dynamic pivot mechanism. *Econometrica* **78**(2), 771–789 (2010)
7. Clempner, J.B.: A Markovian Stackelberg game approach for computing an optimal dynamic mechanism. *Comput. Appl. Math.* **40**(186), 1–25 (2021)
8. Clempner, J.B.: Algorithmic-gradient approach for the price of anarchy and stability for incomplete information. *J. Comput. Sci.* **60**, 101589 (2022)
9. Clempner, J.B., Poznyak, A.S.: A Tikhonov regularization parameter approach for solving Lagrange constrained optimization problems. *Eng. Optim.* **50**(11), 1996–2012 (2018)
10. Clempner, J.B., Poznyak, A.S.: A Tikhonov regularized penalty function approach for solving polylinear programming problems. *J. Comput. Appl. Math.* **328**, 267–286 (2018)
11. Clempner, J.B., Poznyak, A.S.: A nucleus for Bayesian partially observable Markov games: joint observer and mechanism design. *Eng. Appl. Artif. Intell.* **95**, 103876 (2020)
12. Clempner, J.B., Poznyak, A.S.: Analytical method for mechanism design in partially observable Markov games. *Mathematics* **9**(4), 1–15 (2021)
13. Clempner, J.B., Poznyak, A.S.: A dynamic mechanism design for controllable and ergodic Markov games. *Comput. Econ.* **61**, 1151–1171 (2023)
14. Clempner, J.B., Poznyak, A.S.: Mechanism design in Bayesian partially observable Markov games. *Int. J. Appl. Math. Comput. Sci.* **33**(3), 463–478 (2023)
15. Clempner, J.B., Poznyak, A.S.: The price of anarchy as a classifier for mechanism design in a pareto-Bayesian-Nash context. *J. Ind. Manag. Optim.* **19**(9), 6736–6749 (2023)
16. Goldman, C., Zilberstein, S.: Mechanism design for communication in cooperative systems. In: Game Theoretic and Decision Theoretic Agents Workshop at AAMAS'03, Melbourne, Australia, pp. 1–9 (2003)
17. Grover, D., Basu, D., Dimitrakakis, C.: Bayesian reinforcement learning via deep, sparse sampling. In: Chiappa, S., Calandra, R. (eds.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, vol. 108, pp. 3036–3045. PMLR (2020)
18. Groves, T.: Incentives in teams. *Econometrica* **41**, 617–631 (1973)
19. Kassab, R., Simeone, O.: Federated generalized Bayesian learning via distributed stein variational gradient descent (2020). [ArXiv:2009.06419](https://arxiv.org/abs/2009.06419)
20. Mgundi, D.: Efficient reinforcement dynamic mechanism design. In: GAIW: Games, Agents and Incentives Workshops, at AAMAS. Montréal, Canada (2019)
21. Myerson, R.B.: Allocation, Information and Markets, Chap. Mechanism Design, pp. 191–206. The New Palgrave. Palgrave Macmillan, London (1989)
22. Nolan, S., Smerzi, A., Pezzè, L.: A machine learning approach to Bayesian parameter estimation (2020). [arXiv:2006.02369v2](https://arxiv.org/abs/2006.02369v2)
23. Pavan, A., Segal, I., Toikka, J.: Dynamic mechanism design: a Myersonian approach. *Econometrica* **82**(2), 601–653 (2014)
24. Senda, K., Hishinuma, T., Tani, Y.: Approximate Bayesian reinforcement learning based on estimation of plant. *Auton. Rob.* **44**, 845–857 (2020)
25. Sinha, A., Anastopoulos, A.: Mechanism design for resource allocation in networks with intergroup competition and intragroup sharing. *IEEE Trans. Control Netw. Syst.* **5**(3), 1098–1109 (2017)

26. Vasilyeva, M., Tyrylgin, A., Brown, D., Mondal, A.: Preconditioning Markov chain monte Carlo method for geomechanical subsidence using multiscale method and machine learning technique. *J. Comput. Appl. Math.* **392**, 113420 (2021)
27. van Geen, C., Gerraty, R.T.: Hierarchical Bayesian models of reinforcement learning: introduction and comparison to alternative methods (2020). BioRxiv 2020.10.19.345512, <https://doi.org/10.1101/2020.10.19.345512>

Chapter 8

Joint Observer and Mechanism Design



Abstract An intelligent agent suggests an autonomous entity, which manages and learns actions to be taken towards achieving goals. The issue is that it is difficult to build a method that can calculate effective judgments that maximize the overall reward of interacting agents upon an environment with unknown, partial, and uncertain information, according to reports in the literature on Artificial Intelligence (AI). This chapter offers a solution to these problems: a foundation for Bayesian Partially Observable Markov Games (BPOMGs) supported by an AI strategy. The nucleus's structure is governed by three essential concepts: game theory, learning, and inference. The first thing we do is provide a brand-new general Bayesian strategy that is designed for games that take into account both the partial information provided by the Bayesian model and the incomplete information on the states of the Markov system. This approach uses a Partly Observable Markov Game (POMG) to deal with execution uncertainty. Second, we expand design theory to include joint observer design and mechanism design (both unknown). Because agents behave in their own self-interest, the mechanism is created to persuade them not to divulge their personal information and to get a certain result. The purpose of the joint observer design is to depict the possibility that agents may not be motivated to deliver accurate information about their states. The transition matrices, which are also unknown, are estimated by the agents using a model that uses a Reinforcement Learning (RL) technique at each time step. The result is an expanded POMG model that adds a new variable and suggests an analytical method for computing both the observer design and the mechanism design (both unknown). The suggested expansion makes the computationally challenging game theory issue tractable. The variables of relevance for each agent, such as the observation kernels, joint observers, mechanisms, strategies, and distribution vectors, are derived relations in order to retrieve them analytically. By simulating a game-theoretic examination of the patrolling issue that involves developing the mechanism, calculating the observers, and using an RL technique, the utility and efficacy of the suggested nucleus are confirmed.

8.1 Introduction

8.1.1 Brief Problem Analysis

The agent's understanding of its surroundings in real-world settings is unknown, imperfect, or ambiguous. The AI field searches for analytical techniques capable of resolving issues of this nature. In our scenario, a Partly Observable Markov Games (POMGs) [6], which is a controlled stochastic process where it is assumed that the system dynamics are dictated by a Markov process, is used to manage the execution of the uncertainty since the agent cannot directly view the underlying set of states. An agent must revise its belief in the state that the environment may (or may not be in) after acting and perceiving a condition. Agents use a variety of tools to function in these gaming contexts.

- *First*, in order to maximize a reward that depends on the order of system states and the agent's activities, agents should execute learning actions using an RL algorithm in an effort to learn more about their surroundings. The RL method does not (necessarily) presuppose knowledge of a precise mathematical model; see [7] for further information.
- *Second*, the environment is in a certain state at every point in time. Each agent decides on a course of action that probabilistically moves the environment to the next state. The agent also gets an observation at this moment that is based on the altered condition of the environment. He creates a belief about the state the system is in based on these observations. Agents calculate the environment's observation kernel in our unknown information game model.
- *Third*, the mechanism design is a game theory framework for comprehending how an agent might get the “best” results when the primary objective of the game is the agent's self-interest.

This approach arises from the limitations of individual's self-interest incentives and motives, and how these might work in the agent's favor. Mechanism design considers incentives and private information to increase agent rewards and demonstrates how the correct incentives may persuade agents to reveal their private information and produce a certain result. Because agents operate in their own self-interest and are not incentivized to deliver accurate information, POMG is a suitable model for simulating this environment. As a result, the observation kernel may also be created and adopts an engineering methodology to create incentives that lead to desired results (outcomes). The mechanism and joint observer designs both conform to the design theory.

8.1.2 Related Work

Many facets of mechanism design have been covered in earlier research. Mechanism design, which Hurwicz [44] initially described in 1960, has emerged as a technique to design (model, analyze, and solve) game theory issues using an engineering approach, incorporating numerous players who interact rationally [41, 48]. This theory is based on games with incomplete information that attempt to represent how private information might be reported and address the issue of how the information problem limits behavior by allowing social decisions to take the player's preferences into account. Designing a computational mechanism that is both computationally tractable and yet maintains the essential properties of game theory is a difficult task. Based on the idea of game theory, the mechanism design [49, 60] simulates how people interact with one another. The primary focus is on creating games that take into account independent private values and quasilinear payoffs [16, 41]. In these games, participants get messages that contain information about the payoffs [42]. Players in the game report a type that can be a strategic falsehood. Players are paid in accordance with the payment structure following the report. Players commit to a system in the game's dynamics that offers an outcome based on the sort of information that may have been falsely provided. It is essential to remember that the mechanism remains unknown. The social choice function is mapped directly from the (true) type profile to the alternatives of commodities obtained by the mechanism designer. The mechanism implements a social choice function that maximizes some value criterion (profit). In this sense, a mechanism is considered to implement a social-choice function if it leads to a Nash equilibrium for every possible combination of individual preferences.

The 1980s, 1990s, and recent years have seen considerable advancements in mechanism design. See [11, 50] for a survey. Models of negotiations, auctions, public utility regulation, the supply of public goods, nonlinear pricing, and labor market contracts are just a few examples of the many applications for which the mechanism design theory is used. This idea enables the management of player information control and limitation [15, 39, 47]. In this regard, Arrow [5] proposed a method of demand disclosure that maximizes efficiency and prevents resource wastage in incentive payments. In order to construct a mechanism with partial knowledge, d'Aspremont and Gerard-Varet [33] presented two distinct methods. The first method does not incorporate individual beliefs, whereas the second method does. Rogerson [55] developed a generic version of the hold-up issue where numerous participants make relation-specific investments and subsequently agree on some collective action, demonstrating that first-best solutions exist under a range of different assumptions about the nature of information asymmetries. Mailath and Postlewaite [46] proposes a solution for the bargaining problems with asymmetric knowledge involving numerous agents. The model of price discrimination proposed by Courty and Li [31] assumes that customers are aware of the distribution of their values at the time of contracting, and the best mechanisms to use rely on the informational quality of the consumers' initial valuation knowledge. Battaglini [10] provided an ideal infinite-horizon mech-

anism design that takes into account a Markov process in which an agent reacts to a linear model with a continuum of states. Using a two-period state representation to orthogonalize an agent's future information, Eső and Szentes [35] examined a monopolist model in which it sells an indivisible item to numerous agents for optimum information disclosure in auctions. Board [13] used a multi-agent framework with an infinite-time horizon to augment the model described in [35]. Using bandit actions and demonstrating that the ideal mechanism is a variation of the dynamic pivot mechanism, Kakade et al. [45] improved the work presented in [13]. Gershkov and Moldovanu [38] developed a mechanism method that takes into account a dynamic revenue maximization issue in which agents arrive stochastically over time. Wang et al. [61] proposed an ideal double auction method, utilizing a multi-objective model to optimize the predicted total revenue of sellers and buyers. Athey and Segal [9] developed a Bayesian incentive-compatible method for a dynamic framework with quasilinear payments where agents view private information and choices are made across countably many periods. In dynamic quasilinear frameworks, where private knowledge is acquired over time and judgments are made across a number of periods, Pavan et al. [51] constructed a Myersonian process. The work given in [38] was improved by Board and Skrzypacz [14], who proposed a method for creating a mechanism in which agents consider a single piece of private knowledge but arrive stochastically over time. A method for transforming a possibly suboptimal algorithm approach for optimization into a Bayesian incentive-compatible mechanism that improves societal welfare and income was presented by Hartline and Lucier in their paper [43].

In repeated games that take communication into account, useful results on computing the equilibria have been attained. In repeated games that take communication into account, useful results on computing the equilibria have been attained according to Rotemberg and Dutta [32, 56], the play pathways are comparable to those in games with comprehensive knowledge. Citations for non-cooperative repeated models in the literature include [1, 8, 36, 40, 62]. Rahman, Bernheim and Escobar (2018) [12, 34, 54] all provide cooperative models. Clempner and Poznyak presented non cooperative models [17, 27–30].

In this chapter we consider recurrent games involving communication and observable behavior in which the agents' known payoffs change over time in accordance with an irreducible partly observed Markov chain whose transitions are independent of one another. Below we follow [26].

Our chapter's findings are connected to research on the design of mechanisms and observers for partially observable Markov repeating games [19]. We add to this field by proposing a technique for constructing a mechanism with partial state knowledge, the observers design, and describing an essentially equilibrium behavior in recurrent game models that take unfriendly actions on the course of play. Agents are not driven to give truthful information since they behave in their own self-interest. The *mechanism design* considers a specific result based on an agent's self-interest as well as the actions required to attain it. The *observer design*, on the other hand, assumes that agents are not motivated to expose their true states. Last but not least, our game theory simulates how actors could possibly impact different goals.

The main results of the chapter are summarized as follows:

- Provides a game-theory model that uses both the incomplete information of the Bayesian-Markov model and the incomplete information over the states.
- Proposes the joint observer design, which derives a particular (unknown) observation kernel for each agent.
- Extends the design theory: joint observer design and mechanism design.
- Derives the relations to recover the variables of interest for each agent, i.e. observation kernels, observers design, mechanism, (behavior) strategies and distribution vectors.
- Follows a model that employs an RL approach that responds to the game theory model proposed.
- Studies the problem of computing the equilibrium for the game.
- Suggests a new random walk approach.

8.2 Markov Games

Let $MC = (S, A, \{A(s)\}_{s \in S}, \mathbb{K}, P)$ be a Markov chain [20, 53], where S is a finite set of *states*, $S \subset \mathbb{N}$ and A is a finite *set of actions*. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \cup_{s \in S} A(s)$. Whereas, $\mathbb{K} = \{(s, a) | s \in S, a \in A(s)\}$ is the set of *admissible state-action pairs*. The variable $p_{j|ik}$ is a stationary controlled *transition matrix*, where $\pi_{j|ik} := P(X_{t+1} = s_j | X_t = s_i, A_t = a_k) \forall t \in \mathbb{N}$ represents the probability associated with the transition from state s_i to state s_j , $i = \overline{1, N}$ ($i = 1, \dots, N$) and $j = \overline{1, N}$ ($j = 1, \dots, N$), under an action $a_k \in A(s_i)$, $k = \overline{1, K}$ ($k = 1, \dots, K$). The distribution vector is given by $P(X_t = s_i) = P_i$ such that $P_i \in \mathcal{S}^N$, where $\mathcal{S}^N = \{s \in \mathbb{R}^N : \sum_{i=1}^N P_i = 1, P_i \geq 0\}$.

We consider the case where the process is not directly observable [25]. Let us associate with S the observation set Y which takes values in a finite space $\{1, \dots, M\}$, $M \in \mathbb{N}$. The stochastic process $\{Y_t, t \in \mathbb{N}\}$ is called the *observation process*. By observing Y_t at time t information regarding the true value of X_t is obtained. If $X_t = s_i$ and $A_t = a_k$ an observation $Y_t = y_m$ will have a probability $q_{m|ik} := P(Y_t = y_m | X_t = s_i, A_t = a_k)$, that denotes the relationship between the state and the observation when an action $a_k \in A(s_i)$ is chosen at time t . The observation kernel is a stochastic kernel on Y given by $Q = [q_{m|ik}]$. We restrict ourselves to consider $Q = [q_{m|i}]$.

A controllable *partially observable Markov decision process* (**POMDP**) is a tuple

$$POMDP = \{MC, Y, Q, Q_0, P, V\},$$

where:

- MC is a Markov chain;
- Y is the observation set, which takes values in a finite space $\{1, \dots, M\}$, $M \in \mathbb{N}$;
- $Q = [q_{m|i}]_{m=\overline{1,M}, i=\overline{1,N}}$ denotes the observation kernel is a stochastic kernel on Y ; such that

$$\sum_m q_{m|i} = 1;$$

- $Q_0 = [q_{m|i}]_{m=\overline{1,M}, i=\overline{1,N}}$ denotes the initial observation kernel;
- P is the (a priori) initial distribution;
- u_{ijmk} , is the reward function at time t given the state s_i , the observable state y_m , when the action $a_k \in A(s_i, y_m)$ is taken.

A realization of the Bayesian partially observable system at time t is given by

$$(s(0), x(0), y(0), a(0), s(1), x(1), y(1), a(1), \dots) \in \Omega := (SSYA)^\infty,$$

where s_0 has a given by the distribution $P(X_0 = s_0)$ and $\{A_t\}$ is a control sequence in A determined by a control policy. To define a policy we cannot use the (unobservable) states $s(0), s(1), \dots$. Then, we introduce the observable histories $h_0 := (p, Y_0) \in H_0$ and $H_t := H_{t-1}(AY)$, if $t \geq 1$. To define a (behavioral) strategy we cannot also use the (unobservable) states s_0, s_1, \dots . Then, we introduce the observable histories $h_0 \in \mathbb{H}_0$ and

$$h_t := (s(0), x(0), y(0), a(0), \dots, s(t-1), x(t-1), y(t-1), a(t-1), y(t)) \in H_t \text{ for all } t \geq 1,$$

as well as $\mathcal{H}_t = \mathcal{E}_t \times \mathbb{H}_{t-1}$, if $t \geq 1$. Now, a (behavioral) *strategy* $\sigma_{r|i}$ is a mapping $\sigma : S \rightarrow \Delta(S)$. The set of all feasible policies is denoted by \mathcal{S}_{adm} as follows

$$\mathcal{S}_{adm} = \left\{ \sigma_{r|i} \geq 0 \mid \sum_{r=1}^N \sigma_{r|i} = 1, i = \overline{1, N} \right\}. \quad (8.2.1)$$

In the Bayesian form, if $X_t = x_i$ and $A_t = a_k$ an observation $Y_t = y_m$ will have a probability $q_{m|rk} := P(Y_t = y_m | X_t = x_r, A_t = a_k)$, that denotes the relationship between the observed state and the observation when an action $a_k \in A(x_i)$ is chosen at time t . The observation kernel is a stochastic kernel on Y given by $Q = [q_{m|rk}]$. We restrict ourselves to the case where $[q_{m|rk}] = [q_{m|r}]$.

A game consists of a set $\mathcal{N} = \{1, \dots, n\}$ of players (indexed by $l = \overline{1, n}$). We employ l to emphasize the l th player's variables and $-l$ subsumes all the other players' variables. The dynamics is described as follows. At time $t = 0$, the initial (unobservable) state s_0 has a given a priori distribution P_i^l , and the initial observation y_0 is generated according to the initial observation kernel $Q_0^l(y_0|x_0)$. If at time t the state of the system is X_t and the control $A_t^l \in A^l$ is applied, then each strategy is allowed to randomize, with distribution $\sigma_{r|i}^l(t)$, over the state choices X_t^l based on the mechanism $\mu_{k|m}$ over the action choices $A_t^l \in A^l$.

Formally, a **mechanism is any function** $\mu_{k|m}$ such that

$$\mathcal{M}_{adm} = \left\{ \mu_{k|m} \geq 0 \mid \sum_{k=1}^K \mu_{k|m} = 1, m = \overline{1, M} \right\}. \quad (8.2.2)$$

These choices induce immediate utilities $U_T^l(\mu, \sigma)$ where the system tries to maximize the corresponding one-step valuation functions V_{ijmk}^l . The system tries to maximize the corresponding one-step utility. Next, the system moves to the new state $X_{t+1} = s_j$ according to the transition probabilities $P^l(X_{t+1} = s_j | X_t = s_i, A_t = a_k)$. Then, the observation Y_t is generated by the observation kernel $Q^l(Y_t | X_t)$. Based on the obtained utility, the systems adapt its behavior strategy computing $\sigma_{r|i}^l(t+1)$ based on the mechanism $\mu_{k|m}$ for the next selection of the control actions. The one-step valuation functions V_{ijmk}^l and the transition functions $P^l(X_{t+1} = s_j | X_t = s_i, A_t = a_k)$ are all common knowledge at $t = 0$. The common prior initial distribution vector $P^l(X_t = s_i)$ and the transition function $P^l(X_{t+1} = s_j | X_t = s_i, A_t = a_k)$ are assumed to be independent across players. The interaction between players induces a Bayes partially observable Markov game where the average payoff of player l is the expected value of the summed payoff $U_T^l(\mu, o)$, obtained under the mechanism μ and the observer o , is defined as

$$U_T^l(\mu, o) := \sum_{t=1}^T \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^N \sum_{k=1}^K \sum_{j=1}^N u_{ijmk}^l(t) \pi_{j|ik}^l(t) \prod_{\ell=1}^n \mu_{k|m}(t) \sigma_{r|i}^l(t) q_{m|r}^l(t) P_i^l(t) = \\ \sum_{t=1}^T \sum_{m=1}^M \sum_{k=1}^K (\mu_{k|m}(t))^n \sum_{i=1}^N \sum_{r=1}^N W_{imk}^l(t) o_{irm}^l(t) \prod_{\ell \neq l} o_{irm}^{\ell}(t)$$

(8.2.3)

where $W_{imk}^l(t) = \sum_{j=1}^N u_{ijmk}^l(t) \pi_{j|ik}^l(t)$ and the *joint observer* is defined as $o_{irm}^l(t) = \sigma_{r|i}^l(t) q_{m|r}^l(t) P_i^l(t)$ such that $\sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^N o_{irm}^l(t) = 1$. We have that

$$o_{im}^l(t) = \sum_{r=1}^N o_{irm}^l(t) = q_{m|i}^l(t) P_i^l(t)$$

(8.2.4)

as defined in [25]. Let us denote by \mathcal{O}_{adm} the set of “*feasible joint observers*”:

$$\mathcal{O}_{adm} = \left\{ o_{irm}^l(t) \geq 0 \mid \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^N o_{irm}^l(t) = 1, l = \overline{1, n} \right\}. \quad (8.2.5)$$

8.3 Problem Formulation

In this section, we provide an analytical method for computing a mechanism and the joint observer.

8.3.1 Initial Problem

We assume that players know their individual payoffs U^l as are defined in Chaps. 5 and 7.

Problem 8.1 We will find a mechanism $\mu_{k|m}$ and the joint observers o_{irm} , which solve the following nonlinear programming problem

$$(\mu^*, o^*(\mu^*)) = \arg \max_{\mu \in \mathcal{U}_{adm}} \sum_{l=1}^n U^l(\mu, o^*(\mu)), \quad (8.3.1)$$

where for a given mechanism $\mu_{k|m}$ and observer $o^*(\mu)$ the *Bayesian-Nash equilibrium* condition fulfills:

$$U^l(\mu, o^*(\mu)) \geq U^l(\mu, o^l, o^{-l*}). \quad (8.3.2)$$

Remark 8.1 The mechanism μ is **unique** for all participants.

8.3.2 Auxiliary Problem

Now, introduce the z -variable:

$$z_{irmk}^l = \mu_{k|m} \sigma_{r|i}^l q_{m|r}^l P_i^l = \mu_{k|m} o_{irm}^l, \quad (8.3.3)$$

and consider the following auxiliary problem:

Problem 8.2 We will try to find an auxiliary variable z , which solves the following nonlinear programming problem

$$\left. \begin{aligned} \tilde{U}(z) &= \sum_{l=1}^n \tilde{U}^l(z) \rightarrow \max_{z \in \mathcal{Z}_{adm}}, \\ \tilde{U}^l(z) &:= \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K W_{imk}^l \prod_{\ell=1}^n z_{irmk}^\ell, \end{aligned} \right\} \quad (8.3.4)$$

where z_{irmk}^l is given by (8.3.3) and $\mathcal{Z}_{adm} = \otimes_l Z_{adm}^l$ with

$$Z_{adm}^l := \left\{ z_{irmk}^l \geq 0 \mid \begin{array}{l} \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K z_{irmk}^l = 1; \sum_{m=1}^M \sum_{r=1}^N \sum_{k=1}^K z_{irmk}^l = P_i^l > 0; \\ \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K [\delta_{ij} - \pi_{j|ik}^l] z_{irmk}^l, j = \overline{1, N}; \\ \sum_{\rho=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K [\delta_{\rho m} - q_{m|r}^l] z_{ir\rho k}^l, m = \overline{1, M}; \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K q_{\rho|r}^l z_{irmk}^l \geq 0, \rho = \overline{1, M} \end{array} \right\}. \quad (8.3.5)$$

Notice that the following relations hold:

$$\sum_{k=1}^K \mu_{k|m} = 1, \sum_{r=1}^N \sigma_{r|i}^l = 1, \sum_{m=1}^M q_{m|r}^l = 1, \sum_{i=1}^N P_i^l = 1.$$

It is easy to check that Z_{adm}^l includes the simplex Δ^l , namely, $z^l \in \Delta^l \subset Z_{adm}^l$:

$$\Delta^l := \left\{ z_{irmk}^l \geq 0 \mid \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K z_{irmk}^l = 1; \sum_{m=1}^M \sum_{r=1}^N \sum_{k=1}^K z_{irmk}^l = P_i^l > 0 \right\} \quad (8.3.6)$$

Define the solution of the problem (8.3.4) as $z^* = (z^{1*}, \dots, z^{n*})$.

8.4 Relation of Solutions for Initial and Auxiliary Problems

The Lemma below clarifies how we may recover mechanism $\mu_{a|m}^*$.

Theorem 8.1 Suppose that the problem (8.3.4) is solved. Then the mechanism $\mu_{a|m}^*$ can be recovered from z_{irmk}^{l*} as follows:

$$\boxed{\mu_{k|m}^* = \frac{\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K z_{irmk}^{l*}}{\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K z_{irmk}^{l*}}.} \quad (8.4.1)$$

Proof The mechanism $\mu_{k|m}^*$ can be recovered from Eqs. (8.3.5) and (8.3.6) as follows:

$$\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K z_{irmk}^{l*} := \mu_{k|m}^*.$$

Now, we have that

$$\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K \mu_{k|m}^* \sigma_{r|i}^l q_{m|r}^l P_i^{l*} = \mu_{k|m} \sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K \sigma_{r|i}^l q_{m|r}^l P_i^{l*} = \mu_{k|m}^* \sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K o_{irm}^{l*} = \mu_{k|m} \sum_{l=1}^n \sum_{i=1}^N o_{im}^{l*}.$$

Then, one has that

$$\mu_{k|m}^* = \frac{\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K z_{irmk}^{l*}}{\mu_{k|m} \sum_{l=1}^n \sum_{i=1}^N o_{im}^{l*}} = \frac{\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K z_{irmk}^{l*}}{\sum_{l=1}^n \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K z_{irmk}^{l*}}. \quad (8.4.2)$$

Corollary 8.1 *The relation to recover the observer $q_{m|r}^*$ is given by*

$$q_{m|r}^{l*} = \frac{\sum_{i=1}^N \sum_{k=1}^K z_{irmk}^{l*}}{\sum_{i=1}^N \sum_{\rho=1}^M \sum_{k=1}^K z_{ir\rho k}^{l*}}. \quad (8.4.3)$$

So, finally to recover the optimal joint observer we need to use the formula

$$o_{irm}^{l*} = \sigma_{r|i}^l q_{m|r}^* P_i^{l*}. \quad (8.4.4)$$

We have that if $\sigma_{r|i} = \delta_{ir}$ (see [25])

$$o_{im}^l = \sum_{r=1}^N o_{irm}^l = q_{m|i}^l P_i^l. \quad (8.4.5)$$

Variables $\sigma_{r|i}^l$ and P_m^{l*} may be recovered as it is presented below.

Corollary 8.2 *The equilibrium (behavior) strategies $\sigma_{r|i}^{l*}$ are given by*

$$\sigma_{r|i}^{l*} = \frac{\sum_{m=1}^M \sum_{k=1}^K z_{irmk}^{l*}}{\sum_{\rho=1}^N \sum_{m=1}^M \sum_{k=1}^K z_{ir\rho m k}^{l*}}, \quad (8.4.6)$$

$l \in \mathcal{N}$ and the corresponding stationary distributions \bar{P}_m^{l*} are as follows:

$$\bar{P}_m^{l*} = \sum_{i=1}^N \sum_{r=1}^K z_{irmk}^{l*}, l = \overline{1, n}. \quad (8.4.7)$$

Lemma 8.1 *The obtained mechanism $\mu_{k|m}^*$, the observers o_{irm}^{l*} and the strategies $\sigma_{r|i}^{l*}$ satisfy the Bayesian-Nash equilibrium given in Eq. (8.3.2).*

Proof It results directly from the following consideration:

$$\begin{aligned} \max_{z \in \mathcal{Z}_{adm}} \tilde{U}(z) = \tilde{U}(z^*) &= \sum_{l=1}^n \tilde{U}^l(z^*) = \sum_{l=1}^n U^l(\mu^*, o^*(\mu^*)) = \\ &\sum_{l=1}^n \left(\sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K W_{imk}^l (\mu_{k|m}^*)^n o_{irm}^{l*} \prod_{\iota \neq l} o_{ir\iota}^{l*} \right) = \\ &\sum_{l=1}^n \max_{o^l \in \mathcal{O}_{adm}^l} \left(\sum_{m=1}^M \sum_{k=1}^K (\mu_{k|m}^*)^n \sum_{i=1}^N \sum_{r=1}^K W_{imk}^l o_{irm}^l \prod_{\iota \neq l} o_{ir\iota}^{l*} \right) \geq \\ &\sum_{l=1}^n \left(\sum_{m=1}^M \sum_{k=1}^K (\mu_{k|m}^*)^n \sum_{i=1}^N \sum_{r=1}^K W_{imk}^l o_{irm}^l \prod_{\iota \neq l} o_{ir\iota}^{l*} \right) = \sum_{l=1}^n U^l(\mu^*, o^l, o^{-l*}). \end{aligned} \quad (8.4.8)$$

From this inequality it follows that

$$\sum_{l=1}^n (U^l(\mu^*, o^*(\mu^*)) - U^l(\mu^*, o^l, o^{-l*})) \geq 0. \quad (8.4.9)$$

Since the above inequality is valid for all admissible observer o , it is valid when $\sigma^j = \sigma^{j*}$ and $o^j = o^{j*}$ for $j \neq l$, implying

$$U^l(\mu^*, o^*(\mu^*)) - U^l(\mu^*, o^l, o^{-l*}) \geq 0, \quad (8.4.10)$$

which coincides with Eq. (8.3.2) when $\mu = \mu^*$. The lemma is proven. \square

8.4.1 Ergodicity Condition

The next lemma presents the necessary ergodicity conditions that the solutions of the problem (8.3.4) must satisfy. Let us define $Q^l = [q_{m|r}^l]^{-1}$ is the inverse matrix of $[q_{m|r}^l]$ where $M = N$.¹

Lemma 8.2 *If the mechanism $\mu_{k|m}^*$, the observers o_{irm}^{l*} and the strategies $\sigma_{r|i}^{l*}$ are solutions of the problem (8.3.4), which correspond to a Bayesian-Nash equilibrium in Eq. (8.3.2), then variables z_{irmk}^{l*} for all $l \in \mathcal{N}$ satisfy the following ergodicity constraints:*

$$i) \sum_{m=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K [\delta_{ij} - \pi_{j|ik}^l] z_{irmk}^l, j = \overline{1, N}, M = N, \quad (8.4.11)$$

$$ii) \sum_{\rho=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K [\delta_{\rho m} - q_{m|r}^l] z_{ir\rho k}^l, m = \overline{1, M}, M = N, \quad (8.4.12)$$

¹ Notice that $\sum_{m=1}^M Q_{m|h}^l q_{m|r}^l = \delta_{hr}$ - delta Kronecker symbol.

$$iii) \sum_{\rho=1}^M \sum_{i=1}^N \sum_{r=1}^N \sum_{k=1}^K Q_{\rho|h}^l z_{ir\rho k}^l > 0, h = \overline{1, N}. \quad (8.4.13)$$

Proof (i) To show the relation given in Eq. (8.4.11) one has that:

$$P_j^{l*} = \sum_{\alpha=1}^N \left\{ \sum_{\kappa=1}^N \sum_{\beta=1}^M \sum_{\gamma=1}^K \pi_{j|\alpha\gamma}^l z_{\alpha\kappa\beta\gamma}^{l*} \right\}.$$

On the other hand,

$$\sum_{j=1}^N \left\{ \sum_{\kappa=1}^N \sum_{\beta=1}^M \sum_{\gamma=1}^K \pi_{j|\alpha\gamma}^l z_{j\kappa\beta\gamma}^{l*} \right\} = \sum_{\alpha=1}^N \left\{ \sum_{\kappa=1}^N \sum_{\beta=1}^M \sum_{\gamma=1}^K \pi_{j|\alpha\gamma}^l z_{\alpha\kappa\beta\gamma}^{l*} \right\},$$

developing, one obtains that

$$\sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\beta=1}^M \sum_{\gamma=1}^K [\delta_{\alpha'j} - \pi_{j|\alpha\gamma}^l] z_{\alpha\kappa\beta\gamma}^{l*} = 0, , j = \overline{1, N},$$

satisfying

$$z^{l*} \in \mathcal{E} := \left\{ z_{\alpha\kappa\beta\gamma}^{l*} \left| \sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\beta=1}^M \sum_{\gamma=1}^K [\delta_{\alpha j} - \pi_{j|\alpha\gamma}^l] z_{\alpha\kappa\beta\gamma}^{l*}, j = \overline{1, N}, \right. \right\}. \quad (8.4.14)$$

(ii) To prove the relation in Eq. (8.4.12) note that:

$$\sum_{\varrho=1}^M \sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\gamma=1}^K [\delta_{\varrho\beta} - q_{\beta|\kappa}^l] z_{\alpha\kappa\varrho\gamma}^l = \sum_{\varrho=1}^M \sum_{\kappa=1}^N [\delta_{\varrho\beta} - q_{\beta|\kappa}^l] \sum_{\alpha=1}^N \sum_{\gamma=1}^K z_{\alpha\kappa\varrho\gamma}^{l*}.$$

In addition,

$$\sum_{\alpha=1}^N \sum_{\gamma=1}^K z_{\alpha\kappa\varrho\gamma}^{l*} = \sum_{\alpha=1}^N \sum_{\gamma_t=1}^K \mu_{\gamma|\varrho} o_{\alpha\kappa\varrho}^l = \sum_{\alpha=1}^N \sum_{\gamma_t=1}^K \mu_{\gamma|\varrho} \sigma_{\kappa|\alpha}^l q_{\varrho|\kappa}^l P_{\alpha}^l = q_{\varrho|\kappa}^l \sum_{\alpha=1}^N \sum_{\gamma_t=1}^K \mu_{\gamma|\varrho} \sigma_{\kappa|\alpha}^l P_{\alpha}^l,$$

and

$$\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\gamma=1}^K z_{\alpha\kappa\varrho\gamma}^{l*} = \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma|\varrho} \left(\sum_{\kappa=1}^N \sigma_{\kappa|\alpha}^l \right) q_{\varrho|\kappa}^l P_{\alpha}^l = q_{\varrho|\kappa}^l \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma|\varrho} P_{\alpha}^l.$$

Hence,

$$\begin{aligned}
& \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^K \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*} \frac{\sum_{\alpha=1}^N \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*}}{\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^K \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*}} = \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*} \frac{\sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l q_{\varrho | \kappa}^l P_{\alpha}^l}{\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \left(\sum_{\kappa=1}^N \sigma_{\kappa | \alpha}^l \right) q_{\varrho | \kappa}^l P_{\alpha}^l} = \\
& \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^K \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*} \frac{q_{\varrho | \kappa}^l \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l P_{\alpha}^l}{q_{\varrho | \kappa}^l \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \left(\sum_{\kappa=1}^N \sigma_{\kappa | \alpha}^l \right) P_{\alpha}^l} = \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\kappa=1}^N \sum_{\gamma=1}^K z_{\alpha \kappa \varrho \gamma}^{l*} \frac{\sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l P_{\alpha}^l}{\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} P_{\alpha}^l} = \\
& \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \left(\sum_{\kappa=1}^N \sigma_{\kappa | \alpha}^l \right) q_{\varrho | \kappa}^l P_{\alpha}^l \frac{\sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l P_{\alpha}^l}{\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} P_{\alpha}^l} = q_{\varrho | \kappa}^l \sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} P_{\alpha}^l \frac{\sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l P_{\alpha}^l}{\sum_{l=1}^n \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} P_{\alpha}^l} = \\
& q_{\varrho | \kappa}^l \sum_{\alpha=1}^N \sum_{\gamma=1}^K \mu_{\gamma | \varrho} \sigma_{\kappa | \alpha}^l P_{\alpha}^l = q_{\varrho | \kappa}^l \sigma_{\kappa | \alpha}^l.
\end{aligned}$$

Adding, one has

$$\begin{aligned}
& \sum_{\varrho=1}^M \sum_{\kappa=1}^N [\delta_{\varrho \beta} - q_{\beta | \kappa}^l] q_{\varrho | \kappa}^l \sigma_{\kappa | \alpha}^l = \sum_{\varrho=1}^M \sum_{\kappa=1}^N [\delta_{\varrho \beta} q_{\varrho | \kappa}^l \sigma_{\kappa | \alpha}^l] - \sum_{\varrho=1}^M \sum_{\kappa=1}^N [q_{\beta | \kappa}^l q_{\varrho | \kappa}^l \sigma_{\kappa | \alpha}^l] = \\
& \sum_{\kappa=1}^N [q_{\beta | \kappa}^l \sigma_{\kappa | \alpha}^l] - \sum_{\kappa=1}^N [q_{\beta | \kappa}^l \sigma_{\kappa | \alpha}^l] = 0.
\end{aligned}$$

(iii) For showing the relation in Eq. (8.4.13) one has that for any $h = \overline{1, N}$

$$\begin{aligned}
& \sum_{\rho=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K Q_{\rho | h}^l z_{ir \rho k}^l = \sum_{i=1}^N \sum_{r=1}^K \sum_{\rho=1}^M \sum_{k=1}^K Q_{\rho | h}^l \mu_{k | \rho} o_{ir \rho}^l = \\
& \sum_{\rho=1}^M \sum_{i=1}^N \sum_{r=1}^K \sum_{k=1}^K Q_{\rho | h}^l \mu_{k | \rho} \sigma_{r | i}^l q_{\rho | r}^l P_i^l = \sum_{i=1}^N \sum_{r=1}^K \sum_{\rho=1}^M Q_{\rho | h}^l \sigma_{r | i}^l q_{\rho | r}^l P_i^l = \\
& \sum_{i=1}^N \sum_{r=1}^K \sigma_{r | i}^l P_i^l \sum_{\rho=1}^M Q_{\rho | h}^l q_{\rho | r}^l = \sum_{i=1}^N \sum_{r=1}^K \sigma_{r | i}^l P_i^l \delta_{hr} = \sum_{i=1}^N \sigma_{h | i}^l P_i^l > 0.
\end{aligned}$$

8.5 Reinforcement Learning Approach

8.5.1 Iterative Procedure

We propose a model from experiences [7, 18, 52] that is computed by counting the number η_T of private experiences defining the variables recursively as follows: let $h_t \in \mathbb{H}_t$ ($t \in T$) be finite history and $(s^l(t), a^l(t)) \in S_t^l \times A_t^l$ is its corresponding finite sequence of states and actions, and let us denote for each $(s^l(t+1), s^l(t), a^l(t)) \in S_t^l \times S_t^l \times A_t^l$

$$\eta_{j,\hat{i},a(t)|t \in T}^l(T) = \sum_{t \in T} \mathbb{1}(\hat{s}^l(t+1), \hat{s}^l(t), a^l(t))$$

the discounted number of times along the history h_t ($t \in T$) where the state is $\hat{s}(t)$ implement the action $a(t)$, as well as,

$$\eta_{i_t,a(t)|t \in T}^l(T) = \sum_{t \in T} \mathbb{1}(\hat{s}^l(t) = \hat{s}^l, a^l(t) = a^l)$$

is the discounted number of times along the history h_t ($t \in T$), where the profile is \hat{s}_t^l implement the action $a(t)$ where the indicator function $\mathbb{1}(\cdot)$ is defined as:

$$\mathbb{1}(e_t) = \begin{cases} 1 & \text{if the event } e_t \text{ occurs during period } t \\ 0 & \text{otherwise.} \end{cases}$$

The estimated conditional probability $\hat{p}_{j,\hat{i},k}^l$, when all states are observable, is given by

$$\hat{\pi}_{\hat{j},\hat{i},k}^l = \frac{\eta_{\hat{j},\hat{i},k|t \in T}^l}{\eta_{\hat{i},k|t \in T}^l}.$$

We can write the discounted payoff of player l as a function of the discounted measure $\eta_{\hat{j},\hat{i},\hat{m},k}^l$ of states and outcomes,

$$\hat{u}_{\hat{j},\hat{i},\hat{m},k}^l = \frac{\sum_{t \in T} \varsigma_{\hat{j},\hat{i},\hat{m},k}^l \mathbb{1}(\hat{s}^l(t+1), \hat{s}^l(t), \hat{y}^l(t), a^l(t))}{\eta_{\hat{j},\hat{i},\hat{m},k}^l}, \quad (8.5.1)$$

where

$$\varsigma_{\hat{j},\hat{i},\hat{m},k}^l = u_{\hat{j},\hat{i},\hat{m},k}^l \frac{1+r_t}{2}.$$

such that $\varsigma^l \leq u^l$ and r_t takes randomly the values within the interval $[-1, 1]$.

For each player l the mechanism design $\mu_{a|m}^*$, the strategy $\sigma_{r|i}^{l*}$, the distribution P_m^l and the observation kernel $[q_{m|r}^l]$ the construction of the dynamic mechanism design $\mu_{k|m}^*$ is supported by the elements of the estimated transition matrix $\hat{p}_{\hat{j},\hat{i},a}^l$. We suppose that the $\det[q_{m|r}^l] \neq 0$ such that $Q = [q]^{-1}$ exists where $q = [q_{m|r}^l]$.

We propose a model from experiences that is computed by counting the number $\eta(T)$ of unobserved experiences defining the following variables recursively as follows:

$$\eta_{m,k|t \in T}^l(T) = \sum_{t \in T} \mathbb{1}(m, a),$$

$$\eta_{m',m,k|t \in T}^l(T) = \sum_{t \in T} \mathbb{1}(m', m, a),$$

Table 8.1 BPOMG learning process**Algorithm 1:** Learning processes

Let $\mathcal{N} = \{1, \dots, n\}$ be the set of players ($l \in \mathcal{N}$)

Let $s_0^l = s^l$ be the initial state for player l

Let $\hat{\Pi}^l = \Pi^l$ be the initial transition matrix for player l

Let $\varepsilon > 0$ be the error of the estimated parameters

do

Compute the strategy $\sigma_{r|i}^{l*}$, the observers o_{irm}^l and the mechanism $\mu_{k|m}^*$ by applying the game theory solver (see Sect. 8.6)

Select randomly a type r from the $\sigma_{r|i}^{l*}$

Select randomly an m from the $q_{m|r}^l$

Select randomly an action $a(t)$ from the $\mu_{k|m}^*$

From the transition matrix $\hat{\pi}_{j|\hat{i},k}^l$ get next state j for each player l

Increase the values of $\eta_{m',m,k|t \in T}^l(T)$ and $\eta_{m,k|t \in T}^l(T)$

Compute the mean square error e_t^l

Estimate $\hat{\pi}_{j|\hat{i},k}^l$ given in Eq. (8.5.2)

Estimate $\hat{u}_{\hat{j},\hat{i},\hat{m},k}^l$ given in Eq. (8.5.1)

Update $i = j$ and increase t by 1 ($t = t + 1$)

Until for each player $\varepsilon > e_t^l$

where $\eta_{m,k|t \in T}^l(T)$ is the number of visits in the reported state m and, $\eta_{m',m,k|t \in T}^l(T)$ denotes the total number of times that the process evolves from the reported state m to m' applying action a . We have that $\eta_{m,k|t \in T}^l(T) = \sum_{m'=1}^M \eta_{m',m,k|t \in T}^l(T)$. The frequency is defined by $\phi_{m',m,k|t \in T} = t^{-1} \eta_{m',m,k|t \in T}^l(T)$ (Table 8.1).

Remark 8.2 For each player l , the estimated transition matrix $\hat{\pi}_{j|\hat{i},k}^l(T)$ is represented by

$$\hat{\pi}_{j|\hat{i},k}^l(T) = \frac{\sum_{m=1}^M \sum_{m'=1}^M Q_{m'|\hat{r}}^l Q_{m|\hat{r}_{t+1}}^l \eta_{m',m,k|t \in T}^l(T)}{\sum_{\hat{r}_{t+1}=1}^N \left[\sum_{m=1}^M \sum_{m'=1}^M Q_{m'|\hat{r}_{t+1}}^l Q_{m|\hat{r}_{t+1}}^l \eta_{m',m,k|t \in T}^l(T) \right]}. \quad (8.5.2)$$

8.5.2 Learning Algorithm

In the Algorithm 1, the initial time is $t = 0$ and for each player l we let $s_0^l = s^l$ be the initial state and $\hat{\Pi}^l = \Pi^l$ the initial transition matrix. We fix the error of the estimated parameters by $\varepsilon > 0$. For each iteration, we compute the strategy $\sigma_{r|i}^{l*}$, the

observers o_{irm}^l and the mechanism $\mu_{k|m}^*$. Next, we select randomly r from the strategy $\sigma_{r|i}^{l*}$, randomly an m from the $q_{m|r}^l$, and action k from the mechanism $\mu_{k|m}^*$. Then, the system advances by obtaining the state j randomly from $\hat{\pi}_{j|\hat{i},k}^l$ for fixed state i and action k . Next, we update the values of $\eta_{m',m,k|t \in T}^l$ and $\eta_{m,k|t \in T}^l$. So, we get the estimate of $\hat{\pi}_{j|\hat{i},k}^l$ and $\hat{u}_{\hat{j},\hat{i},\hat{m},k}^l$ using the estimation rules in Eqs. (8.5.2) and (8.5.1), respectively. Finally, we compute the mean square error e_t by using

$$e_t^l = \sum_{k=1}^K \text{tr} \left((\hat{\Pi}^l(t-1) - \hat{\Pi}^l(t))^\top (\hat{\Pi}^l(t-1) - \hat{\Pi}^l(t)) \right). \quad (8.5.3)$$

The process continues until $\varepsilon > e_t^l$ for each l . At the end, we obtain the estimated transition matrices $\hat{\pi}_{j|\hat{i},k}^l$, the estimated values of $\hat{u}_{\hat{j},\hat{i},\hat{m},k}^l$. The resulting strategy profile $\sigma_{r|i}^{l*}$ is a Bayesian-Nash equilibrium. The resulting mechanism $\mu_{k|m}^*$ is incentive-compatible because it is in equilibrium and satisfies Eq. (8.3.1).

8.6 Nash Equilibrium as a Solution of a Max-Min Problem

We consider the Nash equilibrium problem [21, 22]. To this end, we first recall the definition of the (standard) Nash equilibrium problem.

We consider a game played by a set $\mathcal{N} = \{1, \dots, n\}$ players indexed by $l \in \mathcal{N}$. Each player $l \in \mathcal{N}$ control the variable $z_{irmk}^l = \mu_{k|m} \sigma_{r|i}^l o_{irm}^l$. Let us consider a game whose strategies are denoted by $v^l \in V^l$ where V^l is a convex and compact set, and $v^l := \text{col}(z_{irmk}^l)$. Let $v = (v^1, \dots, v^n)^\top \in V$, the vector formed by all these decision variables, be the joint strategy of the players and $v^{-l} := (v^1, \dots, v^{l-1}, v^{l+1}, \dots, v^n)^\top \in V^{-l}$ be a strategy of the rest of the players adjoint to $v^l \in V^l$. To emphasize the l -th player's variables within the vector v , we sometimes write $v = (v^l, v^{-l})^\top \in \mathbb{R}^n$ where where v^{-l} subsumes all the other players' variables. We consider a Nash equilibrium problem with n players and denote by $v = (v^l, v^{-l}) \in \mathbb{R}^n$ the vector representing the v -th player's strategy where $v \in V_{adm}$ and $V_{adm} = V_{adm}^l \times V_{adm}^{-l}$, such that $V_{adm} := \{v | v \geq 0, A_{eq}v = b_{eq} \in \mathbb{R}^{r_0}, A_{ineq}v \leq b_{ineq} \in \mathbb{R}^{r_1}\}$.

Let us consider $v^l : \mathbb{R}^n \rightarrow \mathbb{R}$ be the l -th player's reward function (cost function). We assume that these reward functions are at least continuous, and we further assume that the functions $\phi^l(v^l, v^{-l})$ are concave in the variable v^l . The players try to reach one of the non-cooperative equilibrium, that is, they try to find a strategy $v^* = (v^{1*}, \dots, v^{n*})$ satisfying for any v^l and any $l \in \mathcal{N}$ that

$$\bar{U}(v) := \sum_{l \in \mathcal{N}} \left[\left(\max_{v^l \in V^l} \phi^l(v^l, v^{-l}) \right) - \phi^l(v^l, v^{-l}) \right]. \quad (8.6.1)$$

Let us consider that

$$\bar{U}(v) := \sum_{l \in \mathcal{N}} [\phi^l(\bar{v}^l, v^{-l}) - \phi^l(v^l, v^{-l})], \quad (8.6.2)$$

where

$$\bar{v}^l := \arg \max_{v^l \in V^l} \phi^l(v^l, v^{-l}), \quad (8.6.3)$$

and the function ϕ^l satisfies that

$$\phi^l(\bar{v}^l, v^{-l}) - \phi^l(v^l, v^{-l}) \geq 0 \quad (8.6.4)$$

for any $v^l \in V^l$ and $l \in \mathcal{N}$. Then a vector $v^* \in V$ is a *non-cooperative equilibrium*, or a solution of the Nash equilibrium problem if

$$v^* \in \operatorname{Arg} \max_{v \in V} \{\bar{U}(v)\}. \quad (8.6.5)$$

In the case where $\bar{U}(v)$ is a strictly concave function we have that

$$v^* = \arg \max_{v \in V} \{\bar{U}(v)\}.$$

Lagrange's method is employed for finding the local maximum (minimum) of a function, subject to equality constraints [37, 63] such that,

$$\begin{aligned} \mathbb{L}(v, \lambda_{eq}, \lambda_{ineq}) := \\ \theta U(v) - \lambda_{eq}^\top (A_{eq}v - b_{eq}) - \lambda_{ineq}^\top (A_{ineq}v - b_{ineq}) - \frac{\delta}{2} \left(\|v\|^2 - \|\lambda_{eq}\|^2 - \|\lambda_{ineq}\|^2 \right), \end{aligned} \quad (8.6.6)$$

where the parameters θ and δ are positive, the Lagrange vector-multipliers $\lambda_{ineq} \in \mathbb{R}^{r_1}$ are non-negative and the components of $\lambda_{eq} \in \mathbb{R}^{r_0}$ may have any sign ($\lambda \in \Lambda$). The problem

$$\mathbb{L}_{\theta, \delta}(v, \lambda_{eq}, \lambda_{ineq}) \rightarrow \max_{x \in X_{adm}} \min_{\lambda_{eq}, \lambda_{ineq} \geq 0} \quad (8.6.7)$$

has a unique saddle-point on x since the Eq. (8.6.6) is *strongly concave* if the parameters θ and $\delta > 0$ provide the condition that

$$\frac{\partial^2}{\partial v \partial v^\top} \mathbb{L}_{\theta, \delta}(v, \lambda_{eq}, \lambda_{ineq}) < 0 \forall v \in V_{adm} \quad (8.6.8)$$

and it is strongly convex on the Lagrange multipliers $\lambda_{eq}, \lambda_{ineq}$ for any $\delta > 0$. As a result, the Eq. (8.6.6) has the unique saddle point $(v^*(\delta), \lambda_{eq}^*(\theta, \delta), \lambda_{ineq}^*(\theta, \delta))$ (see the Kuhn-Tucker Theorem 21.13 in [52]) for which the following inequalities hold:

$$\begin{aligned} \mathbb{L}_{\theta,\delta} \left(v(\delta), \lambda_{eq}^*(\theta, \delta), \lambda_{ineq}^*(\theta, \delta) \right) &\leq \mathbb{L}_{\theta,\delta} \left(v^*(\delta), \lambda_{eq}^*(\theta, \delta), \lambda_{ineq}^*(\theta, \delta) \right) \leq \\ &\mathbb{L}_{\theta,\delta} \left(v^*(\delta), \lambda_{eq}, \lambda_{ineq} \right) \end{aligned} \quad (8.6.9)$$

for any $\lambda_{eq}, \lambda_{ineq}$ with nonnegative components and any $v \in \mathbb{R}^n$.

We suppose that the parameter θ and the regularizing parameter δ tend to zero ($\theta, \delta \downarrow 0$). It is also assumed that $d\{a; Y\}$ is the Hausdorff distance defined as $d\{c; Y\} = \min_{y \in Y} \|c - y\|^2$. Then, the solutions of $v^*(\theta, \delta)$ and $\lambda_{eq}^*(\theta, \delta), \lambda_{ineq}^*(\theta, \delta)$ of the max-min problem given in Eq. (8.6.7) tend to the set $X^* \otimes \Lambda^* \otimes \Lambda^*$ of all the saddle points of the original game theory problem given in Eq. (8.6.1), that is,

$$d \left\{ v^*(\theta, \delta), \lambda_{eq}^*(\theta, \delta), \lambda_{ineq}^*(\theta, \delta); V^* \otimes \Lambda^* \otimes \Lambda^* \right\} \xrightarrow[\theta, \delta \downarrow 0]{} 0. \quad (8.6.10)$$

Then, the following theorem holds.

Theorem 8.2 *The solution x^* of the problem*

$$\mathbb{L}_{\theta,\delta} (v, \lambda_{eq}, \lambda_{ineq}) \rightarrow \max_{v \in V_{adm}} \min_{\lambda_{eq}, \lambda_{ineq} \geq 0}$$

is a (unique) Nash equilibrium point.

Proof Straightforward following [23, 24]. □

8.7 Application: Patrolling

This section present an application of our results and methods.

8.7.1 Description of the Patrolling Problem

Patrolling an environment can be considered as finding efficient ways of performing visits to target locations of a given area [2, 3, 59]. The challenge of protecting different locations employing limited patrolling resources, makes it impossible to protect all targets properly. This problem is usually represented by a game theory approach, which seems to be appropriate to solve many real-world security situations. This task can be considered inherently multiagent-like since in most cases the process will be started in a distributed manner, by a group of agents. The players react to each other's changes until an equilibrium is achieved. A Nash equilibrium occurs when defenders and attackers react to each other's strategic changes until their actions reach an equilibrium. An approach for solving this problem is to design a mechanism where its equilibrium state coincides with the state of the environment. It is possible

to consider different designs of mechanisms as strategic games that analyses the existence of Nash equilibrium in dominating strategies (agents may not be motivated to provide accurate information and they are induced to reveal private information and create a particular outcome). This example presents a game-theoretic analysis of mechanisms and observers using an RL approach.

For representing the patrolling game, we will consider three players, an attacker ($h = 1$) that try to make the most of the expected damage and two defenders ($l = 1, 2$) that try to stop the attacker. In the dynamics of the game the players take alternate turns: defenders commit first to a strategy and then, the attacker strategy is played. Let the number of states of each player be $\theta = 4$, and the number of actions of each player be $a = 3$. We will recover analytically the variables of interest for each agent, i.e. observation kernels (q), joint observers (o), mechanism (μ), behavior strategies (σ), and distribution vectors (P).

8.7.2 Solver

The Lagrange function $\mathbb{L}_{\theta, \delta}(v^l, v^{-l}, \lambda)$ is poly-linear in $v^l \in V^l$, and therefore cannot be solved analytically. So, we need to apply an iterative method to find a minimizing solution which, additionally, may be not unique. The general format iterative version ($n = 0, 1, \dots$) of the iterative step method for computing the Nash equilibrium for Markov chains [4, 57, 58] with some fixed admissible initial values ($v_0 \in V, v_0^{-l}(v) \in \hat{V}$) is as follows

1. Proximal prediction step:

$$\left. \begin{aligned} \bar{\lambda}_n &= \arg \min_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathbb{L}_{\theta, \delta}(x_n^l, x_n^{-l}, \lambda) \right\}, \\ \bar{v}^l_n &= \arg \min_{v^l \in V^l} \left\{ \frac{1}{2} \|v^l - v_n^{-l}\|^2 + \gamma \mathbb{L}_{\theta, \delta}(v^l, v_n^{-l}, \bar{\lambda}_n) \right\}, \\ \bar{v}_n^{-l} &= \arg \min_{v^{-l} \in V^{-l}} \left\{ \frac{1}{2} \|v^{-l} - v_n^{-l}\|^2 + \gamma \mathbb{L}_{\theta, \delta}(v_n, v^{-l}, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (8.7.1)$$

2. Gradient approximation step:

$$\left. \begin{aligned} \lambda_{n+1} &= \arg \min_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathbb{L}_{\theta, \delta}(\bar{v}_n^l, \bar{v}_n^{-l}, \lambda) \right\}, \\ v^l_{n+1} &= \arg \min_{v^l \in V^l} \left\{ \frac{1}{2} \|v^l - v_n^{-l}\|^2 + \gamma \mathbb{L}_{\theta, \delta}(v^l, \bar{v}_n^{-l}, \bar{\lambda}_n) \right\}, \\ \bar{v}_n^{-l} &= \arg \min_{v^{-l} \in V^{-l}} \left\{ \frac{1}{2} \|v^{-l} - v_n^{-l}\|^2 + \gamma \mathbb{L}_{\theta, \delta}(\bar{v}_n, v^{-l}, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (8.7.2)$$

The formulas in Eq. (8.7.1) and Eq. (8.7.2) have an interpretation: they involve two nonlinear systems, corresponding to evaluation of the extraproximal operators. Evaluating the Eq. (8.7.1) and Eq. (8.7.2) of the objective involves solving two related optimization problems, one for each possible sequence of outcomes. The first step computes the direction of the future evolution of the optimization problem at a given

point and, the second step makes the gradient step from the same point along the predicted direction of the optimization problem.

Conditions to the parameters [23, 24] are given by

$$\delta_n = \begin{cases} \delta_0 & \text{if } n \leq n_0 \\ \delta_0 \frac{[1 + \ln(n - n_0)]}{(1 + n - n_0)^{\beta_\delta}} & \text{if } n > n_0 \end{cases}, \theta_n = \begin{cases} \theta_0 & \text{if } n < n_0 \\ \frac{\theta_0}{(1 + n - n_0)^{\beta_\theta}} & \text{if } n \geq n_0, \end{cases}$$

$$\beta_\delta = 1/2, \beta_\theta = 3/4.$$

The optimal orders δ , θ and γ are as follows

$$\gamma = \gamma^* = \frac{1}{2}, \delta = \delta^* = \frac{1}{2}, \theta^* = \frac{3}{4}, \varkappa^* = 1.$$

providing the order of the convergence $O\left(\frac{1}{n^{\varkappa^*}}\right) = O\left(\frac{1}{n}\right)$.

8.7.3 Random Walk Problem Formulation

For the simulation of the security game, we take into account two defenders and one attackers in which, at each time instant n , the defenders and the attacker select randomly a type r from the $\sigma_{r|l}^{l*}$ and the $\sigma_{r|l}^{h*}$. Then, choose an observable state $y^l(n)$ and $y^h(n)$ from $q_{m|l}^l$ and $q_{m|l}^h$. Next, they select randomly stochastic actions k from $\mu_{k|m}^*$. From the transition matrix $\hat{p}_{j|\hat{i},k}^l$ get next state j for each player l , so as to respectively minimize and maximize the probability of damage jumping from the partially observed state $y^l(n)$ and $y^h(n)$ at the next time instant $n + 1$ (given by $y^l(n + 1)$ and $y^h(n + 1)$). The defenders and attacker attempt finishing the game at the next time instant $n + 1$: this until the realization of the Markov game satisfies the game-over or capture condition given by

$$\sum_{l=1}^2 \sum_{h=1}^1 \sum_{m=1}^M \mathbb{1}(y^l(n) = y_m) \mathbb{1}(y^h(n) = y_m). \quad (8.7.3)$$

The algorithm for the partially observable random walk is given by:

Algorithm 8.1: Random walk

-
1. Select randomly an initial state i^l for each player ι (defender l or attacker h) from P_i^{l*} .
 2. Let $\mu_{k|m}^*$ be the mechanism design, strategies $\sigma_{r|i}^{l*}$ and kernel $q_{m|r}^{l*}$ of the security game .
 3. **For** each player ι **Do**
 4. **For** select randomly a type r from the $\sigma_{r|i}^{l*}$
 5. **For** the matrix $q_{m|r}^{l*}$ select randomly an observable state $y(t) = y_m$ with random $m \in \{1, M\}$ distributed according to the vector $q_{m|r}^{l*}$.
 6. **For** the matrix $\mu_{k|m}^*$ choose randomly an action $a(t) = a_k$.
 7. **For** the matrix $p_{j|ik}^l$ find the state s_j with random $j \in \{1, N\}$ distributed according to the stochastic vector $\pi_{j|ik}^l$ for a fixed $i \in \{1, N\}$ and action $a = a_k$ for $k \in \{1, K\}$.
 8. Add s_j to the random walk and update the initial value of i with j .
 9. Repeat steps (3) to (7) until the capture condition $\mathbb{1}(\omega : s^l(n) = s_j) \mathbb{1}(\omega : s^h(n) = s_j)$ is satisfied, ($l = 1, 2$) and ($h = 1$).
-

8.7.4 Resulting Values

Applying the proposed method, the resulting mechanism is given by

$$\mu_{k|m}^* = \begin{bmatrix} 0.2915 & 0.7085 \\ 0.5061 & 0.4939 \\ 0.3985 & 0.6015 \\ 0.4775 & 0.5225 \end{bmatrix}.$$

The resulting (behavior) strategies of the attacker are given by

$$\sigma_{r|i}^{1*} = \begin{bmatrix} 0.2304 & 0.2306 & 0.2304 & 0.3085 \\ 0.2500 & 0.2500 & 0.2500 & 0.2500 \\ 0.1962 & 0.1962 & 0.1962 & 0.4115 \\ 0.5454 & 0.1516 & 0.1515 & 0.1516 \end{bmatrix},$$

and the behavior) strategies of the defenders are given by

$$\sigma_{r|i}^{2*} = \begin{bmatrix} 0.2400 & 0.2400 & 0.2400 & 0.2800 \\ 0.1479 & 0.1479 & 0.1479 & 0.5563 \\ 0.2499 & 0.2499 & 0.2499 & 0.2504 \\ 0.2489 & 0.2504 & 0.2504 & 0.2504 \end{bmatrix}, \quad \sigma_{r|i}^{3*} = \begin{bmatrix} 0.2401 & 0.2401 & 0.2401 & 0.2797 \\ 0.1180 & 0.1180 & 0.1180 & 0.6459 \\ 0.2127 & 0.2127 & 0.2133 & 0.3613 \\ 0.4690 & 0.1770 & 0.1770 & 0.1770 \end{bmatrix}.$$

The resulting observers of the attacker are given by

$$q_{m|r}^{1*} = \begin{bmatrix} 0.1224 & 0.2026 & 0.5090 & 0.1660 \\ 0.1995 & 0.3304 & 0.1995 & 0.2706 \\ 0.1996 & 0.3302 & 0.1996 & 0.2705 \\ 0.3153 & 0.3300 & 0.1506 & 0.2042 \end{bmatrix},$$

and the resulting observers of the defenders are given by

$$q_{m|r}^{2*} = \begin{bmatrix} 0.1862 & 0.1860 & 0.1860 & 0.4425 \\ 0.2982 & 0.2978 & 0.2978 & 0.2285 \\ 0.3125 & 0.3133 & 0.3133 & 0.1997 \\ 0.2032 & 0.2029 & 0.2029 & 0.1293 \end{bmatrix}, \quad q_{m|r}^{3*} = \begin{bmatrix} 0.1648 & 0.2330 & 0.4375 & 0.1647 \\ 0.2265 & 0.3202 & 0.2269 & 0.2264 \\ 0.2263 & 0.3207 & 0.2268 & 0.2262 \\ 0.6540 & 0.1434 & 0.1014 & 0.1012 \end{bmatrix}.$$

The distribution vectors are as follows:

$$P_i^{1*} = \begin{bmatrix} 0.2765 \\ 0.1959 \\ 0.2048 \\ 0.3228 \end{bmatrix}, \quad P_i^{2*} = \begin{bmatrix} 0.3304 \\ 0.2724 \\ 0.2329 \\ 0.1643 \end{bmatrix}, \quad P_i^{3*} = \begin{bmatrix} 0.2253 \\ 0.3437 \\ 0.2026 \\ 0.2284 \end{bmatrix}.$$

The resulting joint observers are given by

$$o_{im}^{1*} = \begin{bmatrix} 0.0601 & 0.0831 & 0.0707 & 0.0625 \\ 0.0410 & 0.0584 & 0.0519 & 0.0446 \\ 0.0475 & 0.0625 & 0.0492 & 0.0456 \\ 0.0565 & 0.0841 & 0.1165 & 0.0657 \end{bmatrix},$$

$$o_{im}^{2*} = \begin{bmatrix} 0.0615 & 0.0985 & 0.1033 & 0.0671 \\ 0.0506 & 0.0811 & 0.0853 & 0.0553 \\ 0.0433 & 0.0694 & 0.0730 & 0.0473 \\ 0.0727 & 0.0376 & 0.0328 & 0.0213 \end{bmatrix}, \quad o_{im}^{3*} = \begin{bmatrix} 0.0371 & 0.0525 & 0.0985 & 0.0371 \\ 0.0778 & 0.1101 & 0.0780 & 0.0778 \\ 0.0459 & 0.0650 & 0.0460 & 0.0458 \\ 0.1494 & 0.0327 & 0.0232 & 0.0231 \end{bmatrix}.$$

The convergence of the strategies are given in Figs. 8.1, 8.2 and 8.3. The convergence of the error of the transition matrices are given in Figs. 8.4, 8.5 and 8.6. The convergence of the error of the utility matrices are given in Figs. 8.7, 8.8 and 8.9.

We consider the random walk process presented in Algorithm 8.1. A simulation of the game is shown in Fig. 8.10 where the attacker (black) is caught in state 2 after 16 steps by the defender 1 (blue), so the game is over. A different realization of the game is shown in Fig. 8.11. Here, we have that the attacker (black) is caught in state 1 after 11 steps by the defender 1 (blue) and the defender 2 (red) at the same time, then the game is over.

Fig. 8.1 Convergence of the Strategies z^1 of the Attacker

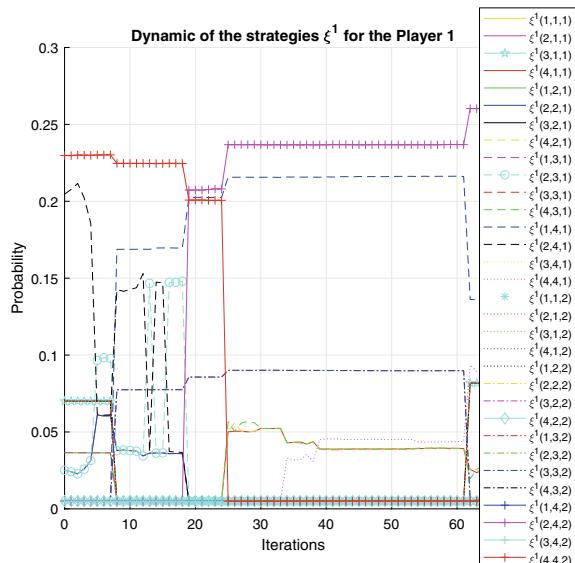


Fig. 8.2 Convergence of the Strategies z^2 of the Defender

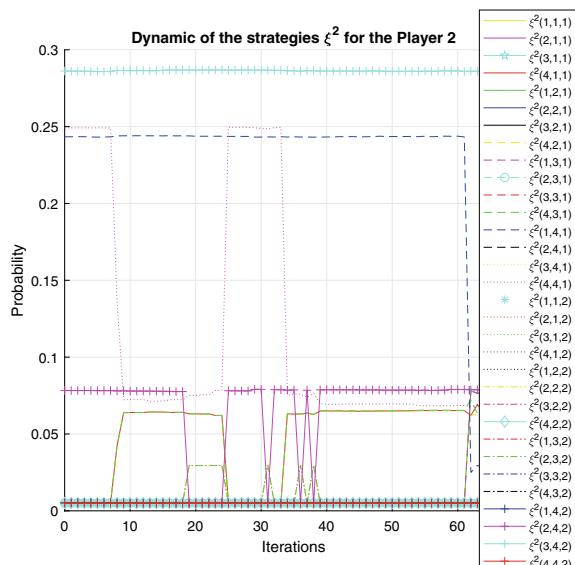


Fig. 8.3 Convergence of the Strategies z^3 of the Defender 2

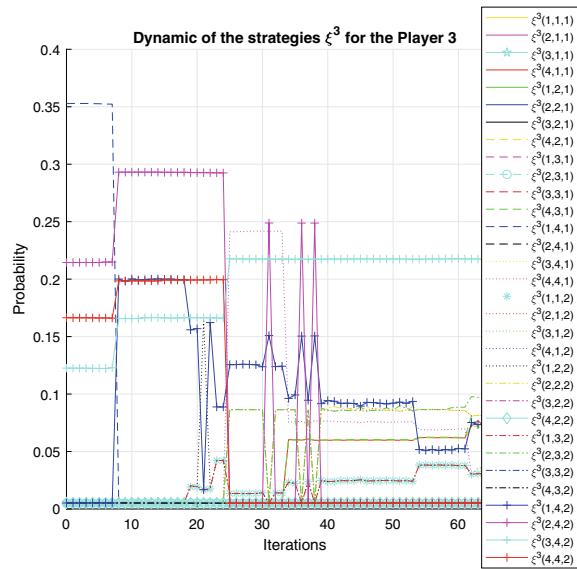


Fig. 8.4 Convergence of the error p^1 of the Attacker

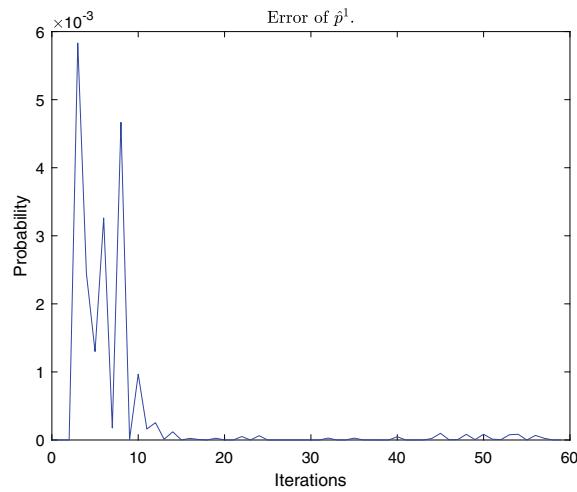


Fig. 8.5 Convergence of the error p^2 of the Defender 1

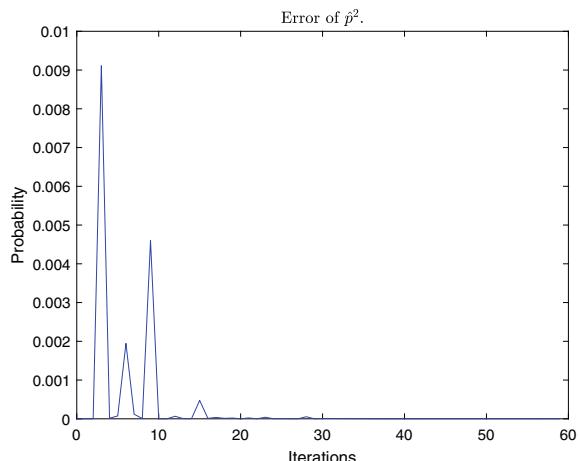


Fig. 8.6 Convergence of the error p^3 of the Defender 2

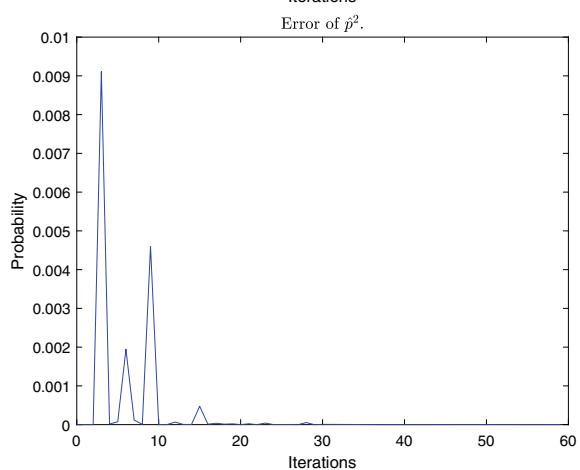


Fig. 8.7 Convergence of the error u^1 of the Attacker

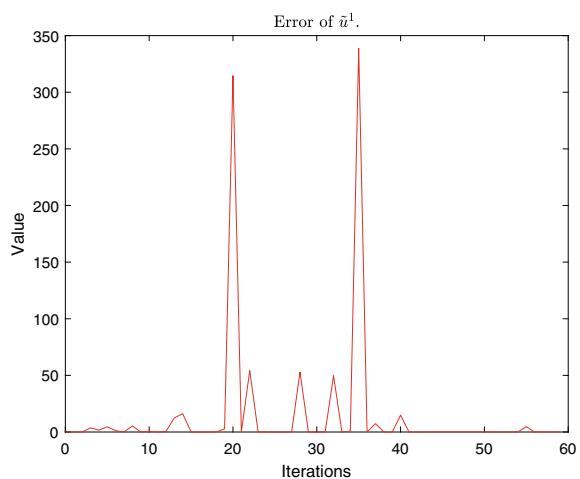


Fig. 8.8 Convergence of the error u^2 of the Defender 1

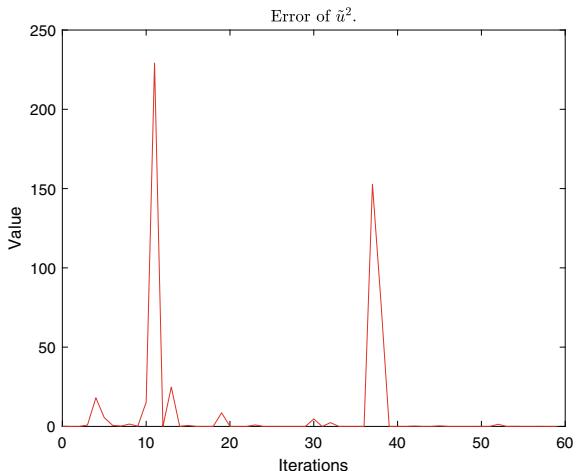


Fig. 8.9 Convergence of the error u^3 of the Defender 2

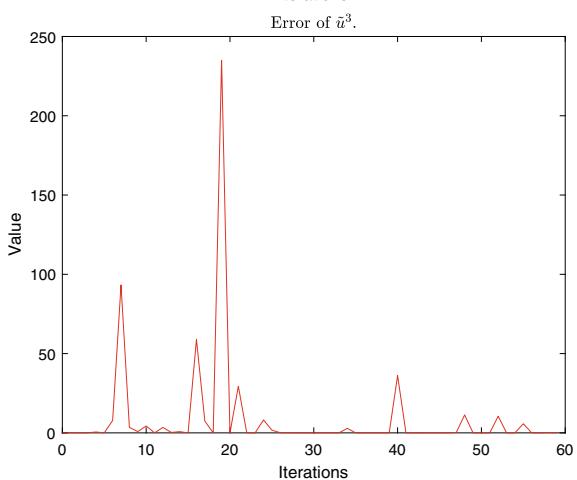


Fig. 8.10 Realization 1 of the game. The attacker (black) is caught by the defender 1 (blue)

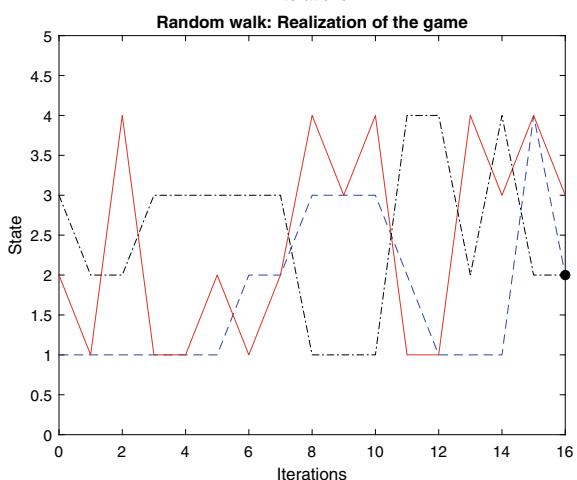
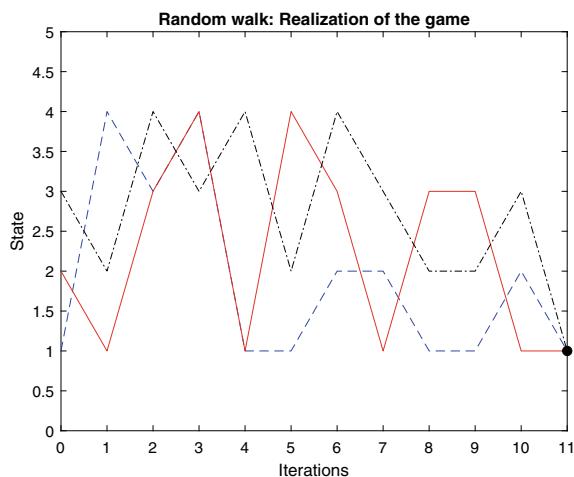


Fig. 8.11 Realization 2 of the game. The attacker (black) is caught by the defender 1 (blue) and the defender 2 (red)



References

1. Abreu, D., Pearce, D., Stacchetti, E.: Optimal cartel equilibria with imperfect monitoring. *J. Econ. Theory* **39**, 251–269 (1986)
2. Albaran, S., Clempner, J.B.: A stackelberg security markov game based on partial information for strategic decision making against unexpected attacks. *Eng. Appl. Artif. Intell.* **81**, 408–419 (2019)
3. Alcantara-Jimenez, G., Clempner, J.B.: Repeated stackelberg security games: Learning with incomplete state information. *Reliab. Eng. Syst. Saf.* **195**, 106695 (2020)
4. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. *Comput. Math. Math. Phys.* **45**(11), 1893–1914 (2005)
5. Arrow, K.: Economics and human welfare, chapter. In: *The property rights doctrine and demand revelation under incomplete information*, pp. 23–39. Academic, New York (1979)
6. Asiain, E., Clempner, J.B., Poznyak, A.S.: A reinforcement learning approach for solving the mean variance customer portfolio for partially observable models. *Int. J. Artif. Intell. Tools* **27**(8), 1850034–1–1850034–30 (2018)
7. Asiain, E., Clempner, J.B., Poznyak, A.S.: Controller exploitation-exploration reinforcement learning architecture for computing near-optimal policies. *Soft. Comput.* **23**(11), 3591–3604 (2019)
8. Athey, S., Bagwell, K.: Optimal collusion with private information. *RAND J. Econ* **32**, 428–465 (2001)
9. Athey, S., Segal, I.: An efficient dynamic mechanism. *Econometrica* **81**(6), 2463–2485 (2013)
10. Battaglini, M.: Long-term contracting with markovian consumers. *Am. Econ. Rev.* **95**(3), 637–658 (2005)
11. Bergemann, D., Välimäki, J.: Dynamic mechanism design: an introduction. *J. Econ. Perspect.* (2018). Forthcoming
12. Bernheim, B., Madsen, E.: Price cutting and business stealing in imperfect cartels. *Am. Econ. Rev.* **107**, 387–424 (2017)
13. Board, S.: Selling options. *J. Econ. Theory* **136**, 324–340 (2007)
14. Board, S., Skrzypacz, A.: Revenue management with forward-looking buyers forward-looking buyers. *J. Polit. Econ.* **124**(4), 1046–1087 (2016)
15. Börgers, T.: *An Introduction to the Theory of Mechanism Design*. Oxford University Press, Oxford (2015)

16. Clarke, E.: Multi-part pricing of public goods. *Pub. Choice* **11**, 17–23 (1971)
17. Clemmpner, J.B.: Algorithmic-gradient approach for the price of anarchy and stability for incomplete information. *J. Comput. Sci.* **60**, 101589 (2022)
18. Clemmpner, J.B.: A bayesian reinforcement learning approach in markov games for computing near-optimal policies. *Ann. Math. Artif. Intell.* **91**, 675–690 (2023)
19. Clemmpner, J.B.: Revealing perceived individuals' self-interest. *J. Oper. Res. Soc.* 1–10 (2023). To be published. <https://doi.org/10.1080/01605682.2023.2195878>
20. Clemmpner, J.B., Poznyak, A.S.: Simple computing of the customer lifetime value: a fixed local-optimal policy approach. *J. Syst. Sci. Syst. Eng.* **23**(4), 439–459 (2014)
21. Clemmpner, J.B., Poznyak, A.S.: Computing the strong nash equilibrium for markov chains games. *Appl. Math. Comput.* **265**, 911–927 (2015)
22. Clemmpner, J.B., Poznyak, A.S.: Convergence analysis for pure and stationary strategies in repeated potential games: nash, lyapunov and correlated equilibria. *Expert Syst. Appl.* **46**, 474–484 (2016)
23. Clemmpner, J.B., Poznyak, A.S.: A tikhonov regularization parameter approach for solving lagrange constrained optimization problems. *Eng. Optim.* **50**(11), 1996–2012 (2018)
24. Clemmpner, J.B., Poznyak, A.S.: A tikhonov regularized penalty function approach for solving polylinear programming problems. *J. Comput. Appl. Math.* **328**, 267–286 (2018)
25. Clemmpner, J.B., Poznyak, A.S.: Observer and control design in partially observable finite markov chains. *Automatica* **110**, 108587 (2019)
26. Clemmpner, J.B., Poznyak, A.S.: A nucleus for bayesian partially observable markov games: joint observer and mechanism design. *Eng. Appl. Artif. Intell.* **95**, 103876 (2020)
27. Clemmpner, J.B., Poznyak, A.S.: Analytical method for mechanism design in partially observable markov games. *Mathematics* **9**(4), 1–15 (2021)
28. Clemmpner, J.B., Poznyak, A.S.: A dynamic mechanism design for controllable and ergodic markov games. *Comput. Econ.* **61**, 1151–1171 (2023)
29. Clemmpner, J.B., Poznyak, A.S.: Mechanism design in bayesian partially observable markov games. *Int. J. Appl. Math. Comput. Sci.* **33**(3), 463–478 (2023)
30. Clemmpner, J.B., Poznyak, A.S.: The price of anarchy as a classifier for mechanism design in a pareto-bayesian-nash context. *J. Ind. Manag. Optim.* **19**(9), 6736–6749 (2023)
31. Courty, P., Li, H.: Sequential screening. *Rev. Econ. Stud.* **67**, 697–717 (2000)
32. Dutta, P.: A folk theorem for stochastic games. *J. Econ. Theory* **66**, 1–32 (1995)
33. d'Aspremont, C., Gerard-Varet, L.: Incentives and incomplete information. *J. Public Econ.* **11**(1), 25–45 (1979)
34. Escobar, J.F., Llanes, G.: Cooperation dynamics in repeated games of adverse selection. *J. Econ. Theory* **176**, 408–443 (2018)
35. Eső, P., Szentes, B.: Optimal information disclosure in auctions and the handicap auction. *Rev. Econ. Stud.* **74**(3), 705–731 (2007)
36. Fudenberg, D., Maskin, E.: The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533–554 (1986)
37. Garcia, C.B., Zangwill, W.I.: Pathways to Solutions. Fixed Points and Equilibria. Prentice-Hall, Englewood Cliffs (1981)
38. Gershkov, A., Moldovanu, B.: Dynamic revenue maximization with heterogenous objects: A mechanism design approach. *Am. Econ. J.: Microecon.* **1**(2), 168–198 (2009)
39. Gershkov, A., Moldovanu, B.: Dynamic Allocation and Pricing: A Mechanism Design Approach. MIT Press (2014)
40. Green, E., Porter, R.: Noncooperative collusion under imperfect price information. *Econometrica* **52**, 87–100 (1984)
41. Groves, T.: Incentives in teams. *Econometrica* **41**, 617–631 (1973)
42. Harsanyi, J.C.: Games with incomplete information played by bayesian players. part i: the basic model. *Manage. Sci.* **14**, 159–182 (1967)
43. Hartline, J.D., Lucier, B.: Non-optimal mechanism design. *Am. Econ. Rev.* **105**(10), 3102–3124 (2015)

44. Hurwicz, L.: Optimality and informational efficiency in resource allocation processes. In: K.J. Arrow, S. Karlin, P. Suppes (eds.), *Mathematical Methods in the Social Sciences: Proceedings of the First Stanford Symposium*, pp. 27–46. Stanford University Press, California (1960)
45. Kakade, S., Lobel, I., Nazerzadeh, H.: Optimal dynamic mechanism design and the virtual-pivot mechanism. *Oper. Res.* **61**(4), 837–854 (2013)
46. Mailath, G., Postlewaite, A.: Asymmetric information bargaining problems with many agents. *Rev. Econ. Stud.* **57**, 351–360 (1990)
47. Menicucci, D.: Efficient mechanisms for a partially public good. *Decis. Econ. Finance* **25**(1), 71–77 (2002)
48. Myerson, R.B.: Allocation, information and markets, chapter. In: *Mechanism Design*, pp. 191–206. The New Palgrave. Palgrave Macmillan, London (1989)
49. Nobel: The sveriges riksbank prize in economic sciences in memory of alfred nobel 2007: Scientific background. Technical report, The Nobel Foundation, Stockholm, Sweden (2007)
50. Pavan, A.: Dynamic mechanism design: Robustness and endogenous types. In: Honore, M.P.B., Pakes, A., Samuelson, L. (eds.), *Advances in Economics and Econometrics: 11th World Congress* (2017)
51. Pavan, A., Segal, I., Toikka, J.: Dynamic mechanism design: a myersonian approach. *Econometrica* **82**(2), 601–653 (2014)
52. Poznyak, A.S.: Advanced mathematical tools for automatic control engineers. In: *Deterministic Technique*, vol. 1. Elsevier, Amsterdam (2008)
53. Poznyak, A.S., Najim, K., Gómez-Ramírez, E.: *Self-learning Control of Finite Markov Chains*. Marcel Dekker, Inc. (2000)
54. Rahman, D.: The power of communication. *Am. Econ. Rev.* (2014)
55. Rogerson, W.: Contractual solutions to the hold-up problem. *Rev. Econ. Stud.* **59**, 777–793 (1992)
56. Rotemberg, J., Saloner, G.: A supergame-theoretic model of price wars during booms. *Am. Econ. Rev.* **76**, 390–407 (1986)
57. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the stackelberg/nash equilibria using the extraproximal method: convergence analysis and implementation details for markov chains games. *Int. J. Appl. Math. Comput. Sci.* **25**(2), 337–351 (2015)
58. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the lp-strong nash equilibrium for markov chains games. *Appl. Math. Modell.* **41**, 399–418 (2017)
59. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Adapting attackers and defenders preferred strategies: a reinforcement learning approach in stackelberg security games. *J. Comput. Syst. Sci.* **95**, 35–54 (2018)
60. Vickrey, W.: Counterspeculation, auctions, and competitive sealed tenders. *J. Finance* **16**(1), 8–37 (1961)
61. Wang, X., Chin, K.S., Yin, H.: Design of optimal double auction mechanism with multi-objectives. *Expert Syst. Appl.* **38**, 13749–13756 (2011)
62. Yang, X., Yin, S.: Variational bayesian inference for fir models with randomly missing measurements. *IEEE Trans. Ind. Electron.* **64**(5), 4217–4225 (2017)
63. Zangwill, W.I.: *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs (1969)

Chapter 9

Bargaining Games or How to Negotiate



Abstract The term “bargaining game” describes a scenario in which participants may reach a win-win agreement. There is a conflict of interest here over whether an agreement should be reached or whether no agreement should be reached without the consent of each player. Amazingly, negotiating protocols, duopoly market games, business transactions, arbitration, and other key scenarios have all used bargaining and its game-theoretic solutions. The underpinning for all of these research applications is equilibrium computation. This chapter explores the theory behind bargaining games and offers a way for solving the game-theoretic models of bargaining put out by Nash and Kalai-Smorodinsky, which suggest a beautiful axiomatic solution to the problem based on several fairness standards. We consider a class of continuous-time, controllable, and ergodic Markov games. The Nash bargaining solution is first introduced and axiomatized. The Kalai-Smorodinsky approach, which adds the monotonicity postulate to the Nash’s model, is then presented. We recommend using a bargaining solver to resolve the issue, which is carried out iteratively using a set of nonlinear equations represented by the Lagrange principle and the Tikhonov regularization approach to guarantee convergence to a particular equilibrium point. Each equation in this solver is an optimization problem for which the necessary condition of a minimum is solved using the projection gradient method. This chapter’s key finding illustrates how equilibrium calculations work in bargaining games. We specifically discuss the convergence analysis and rate of convergence of the suggested technique. A numerical example contrasting the Nash and Kalai-Smorodinsky bargaining solution problem is used to illustrate the value of the considered method.

9.1 Introduction

9.1.1 Brief Review

Numerous significant scenarios, such as arbitration, supply chain agreements, duopoly market games, negotiating protocols, etc., have made use of the *bargaining paradigm*. It presents an idea for a cooperative game’s solution and is related to

negotiation and group decision-making procedures. Collaboration refers to alliances of two or more players acting jointly to achieve a specified goal while keeping in mind the need to maximize each player's personal profit. While discussing the mechanics of a bargaining game, it's important to note that it's possible for players to reach a win-win agreement in certain circumstances. There is a conflict of interest here over whether an agreement should be reached or whether no agreement should be reached without the consent of each player. Nash [27] and Kalai-Smorodinsky [19] are two theoretical viewpoints that offer a solution for cooperative game-theoretic bargaining models that use the axiomatic method to analyze bargaining. It is important to recall that in Nash's model, the disagreement point and the two approaches to a bargaining solution have the same viable payoff set.

Using the game theory framework outlined by von Neumann and Morgenstern in their famous 1944 article [29], John Nash first presented the bargaining model as a game in 1950's important study [27]. According to the von Neumann and Morgenstern hypothesis, when players establish a coalition, they anticipate that a complementary coalition will react by doing the worst possible harm to them. In the literature, there are arguments against this claim. By establishing axioms that define a singular outcome and a solution to the problem known as the Nash bargaining solution, Nash extended the work of von Neumann and Morgenstern in this way. A possible set of utility allocations established through collaboration and the disagreement point that arises when players do not collaborate make up the formal description's two primary parts. A function that chooses a workable utility allocation for each problem is a solution. It's noteworthy to note that in the game theory literature (see [18, 30]), bargaining is one of the first cases of a conflict of interest.

The Nash technique [27], as described in [19], is different from the Kalai-Smorodinsky bargaining strategy. The Kalai-Smorodinsky technique fits monotonicity, but the Nash solution complies with independence of irrelevant alternatives. This is the key distinction between the two approaches. According to Kalai and Smorodinsky, every alternative must have an impact on the decision made.

9.1.2 Related Work

The Nash bargaining issue and the Kalai-Smorodinsky technique have drawn the interest of scholars from several academic fields, and it is still a topic that is interesting to game theory practitioners and academics alike. Merlo and Wilson developed a n -player sequential bargaining model for Markov processes, [23]. They looked at the efficiency and uniqueness of the equilibrium results, the reasons why agreement takes longer, and the benefit of suggesting. The sets of stationary subgame perfect and subgame perfect equilibria that describe the size of the cake and the sequence in which players move define the sequential bargaining model. A model for the negotiation process based on an offer model with an exogenous risk of breakdown for Markov games was proposed by Bolt and Houba, [6]. In a dynamic setting, they examined a modified version of the variable threat game without commitment. An

alternating-offer bargaining game described by Cai [7] shown that the suggested model has a limited number of Markov Perfect Equilibria, some of which display unnecessary delays. The number of delay periods that Markov Perfect Equilibria can sustain rises in the order of the square of the number of participants, and these findings hold true even when the Markov constraints are relaxed and more flexible surplus functions are used. In a market where pairs of agents interested in carrying out a transaction are brought together by a stochastic process and, upon meeting, initiate a bargaining process over the terms of the transaction, Rubinstein and Wolinsky [36] developed a bargaining problem treated with a strategic approach. They calculated the steady-state equilibrium agreements and examined how much they depended on the market. Kalandrakis [20] studied an infinitely repeated divide-the-dollar bargaining game with an endogenous reversion point characterizing a Markov equilibrium which is such that irrespective of the discount factor or the initial division of the dollar, the proposer eventually extracts the whole dollar in all periods. They demonstrated that the status quo's weak continuity of the offered strategies and the failure of the correspondence of the voters' acceptance set to satisfy lower hemicontinuity. A risk-neutral buyer and seller engage in an alternating offer negotiation in which the value of the product being exchanged is determined by a Markov process, according to Cripps [14]. He demonstrated that if the buyer lacks patience compared to the seller, there would be delays in the parties coming to a consensus, the buyer will be compelled to follow a poor consuming strategy, and the equilibrium will be ex-ante inefficient. Moreover, the author demonstrated that there exists a particular and effective equilibrium in which agreement occurs instantly if the buyer is more patient than the seller. Kenan [21] presented repeated contract negotiations involving the same buyer and seller where the size of the surplus being divided, is specified as a two-state Markov chain with transitions that are synchronized with contract negotiation dates. The contracts are linked because the buyer has persistent private information. Since there is persistence in the Markov chain generating the surplus, a successful demand induces the seller to make another aggressive demand in the next negotiation, since the buyer's acceptance reveals that the current surplus is large. An alternating offers-bargaining model was developed by Coles and Muthoo [13], in which the set of potential utility pairings changes over time in a non-stationary fashion. They demonstrated the existence of a unique subgame perfect equilibrium in the limit when the gap between two successive offers grows arbitrarily small and developed a description of the unique subgame perfect equilibrium payoffs. In a class of games that contains contract games and the Nash demand game, Naidu et al. [26] explored purposeful idiosyncratic play in a typical stochastic evolutionary model of equilibrium selection, demonstrating the existence and uniqueness of a stochastically stable bargaining result under such play. Abreu and Manea [1] investigated the Markov perfect equilibria of an infinite horizon game in which players are randomly matched into pairs. They proved that Markov perfect equilibria exist and demonstrated that Markov perfect equilibria payoffs are not always exclusive. A technique for creating pure strategy Markov perfect equilibria with high discount factors was also devised. In a model of social learning, Agastya [2] examined the twin problems of coalition formation and allocation, demonstrating

that all self-perpetuating allocations realized from a straightforward bargaining game must be core allocations, even though players simultaneously demand surplus and only on their own behalf. Additionally, regardless of the society's early history, they offered a suitable circumstance under which it finally learns to allocate the excess according to some fundamental principle.

The Nash equilibrium regularity property was used by Anant et al. [4] to generalize the Kalai-Smorodinsky result and demonstrate that it is true only if the feasibility sets also happen to be Nash equilibrium regular. A one parameter class of asymmetric Kalai-Smorodinsky solutions is defined by restricted independence, scale invariance, Pareto optimality, and the individual monotonicity axiom of Kalai-Smorodinsky. Dubra [16] established a restricted version of Nash's Independence that resolves its major criticisms. For the Kalai-Smorodinsky bargaining solution, Köbberling and Peters [22] showed that utilitarian risk aversion and probabilistic risk aversion can have very different effects on bargaining solutions with a weak monotonicity characteristic. To demonstrate that n -player negotiating problems have a distinct self-supporting result under the Kalai-Smorodinsky solution, Driesen et al. [15] investigated bargaining problems under the premise that players are loss-averse. In addition to proposing a novel axiom termed proportionate concession invariance, they created the bargaining solutions that provide these precise results and described them using the conventional axioms of scale invariance, individual monotonicity, and strong individual rationality. According to Roth [35], no solution, regardless of what it is, can possess the other properties that characterize in the two-person case. This is because the Kalai-Smorodinsky game for two-person bargaining cannot be simply transformed into a general n -person bargaining game, demonstrating that the solution is not Pareto Optimal. In this sense, Peters and Tijs [31] established the existence of a single bargaining solution, defined on the entire class of two-person bargaining games, with the following characteristics: individual rationality, Pareto optimality independence of equivalent utility representation, individual monotonicity and symmetry. The four axioms of the Kalai-Smorodinsky solution are provided, and they demonstrate that precisely one of these solutions is symmetric by introducing a sizable subclass of n -person bargaining games. They demonstrated the risk-sensitivity of each of these strategies as well. Using dictatorial fractions as a threat to accept the winner's proposal, all non-winners of the auction participate in Moulin's [24] negotiating game. A subgame perfect equilibrium was how Moulin defined the equilibrium behavior. The Kalai-Smorodinsky bargaining approach is when businesses and unions just bargain on pay, and businesses decide the level of employment in order to maximize profits given the negotiated wage, according to Alexander [3]. The author examined the scenario in which the union's risk aversion and the pay elasticity of employment are both constant, demonstrating that the Kalai-Smorodinsky and Nash solutions in this situation have a clear link with one another. Some applications are presented in [9, 10].

9.1.3 Nash Versus Kalai-Smorodinsky

In this chapter, following [37, 38, 42], we analyze the bargaining solutions presented by Nash [27, 28] and Kalai-Smorodinsky [19], which depend on different principles of fairness. The bargaining solution allows players to solve the bargaining problem that result in a “fair” improved position.

Following Nash [27], a solution to the bargaining problems \mathcal{B} is a function f that takes as input any bargaining problem and returns a vector of utilities that belongs to the set of possible agreements Φ . Several solutions can be proposed for solving the problem, but some of them can present inconsistencies. For example, one solution can go against symmetry by proposing a total improvement of the position of one player obtaining a point in the Pareto frontier of the utility and the other player receives no improvement [12]. A different solution of the problem could be a disagreement point. The first solution violates symmetry, so the solution is unfair, and the second solution is not Pareto-efficient, and does not take advantage of the cooperation related to an agreement situation. For solving the inconsistencies in the solution of the problem, Nash [27] proposed several axioms:

- (a) *Invariant to affine transformations* (or Invariant to equivalent utility representations): an affine transformation of the utility and disagreement point should not alter the outcome of the bargaining process;
- (b) *Pareto optimality*: the solution selects a point of the Pareto frontier such that the players can be made “better” off without making other players “worse off”;
- (c) *Symmetry*: if the players are indistinguishable, the solution should not discriminate between them; and
- (d) *Independence of irrelevant alternatives*: if the solution is chosen from a feasible set which is an element of a subset of the original set but containing the point selected earlier by the solution, then the solution must still assign the same point chosen from the subset.

As a result, Nash [27] proposed the Nash bargaining solution: we say that there is a unique solution b to the bargaining problem that satisfies the four axioms (a to d) which is given by the point that maximizes the product of utilities of the players.

While three of Nash’s axioms are quite uncontroversial, the fourth one (*Independence of irrelevant alternatives*) raised some criticism, which lead to a different line of research. Kalai and Smorodinsky [19] looked for characterizations of an alternative solution which do not use the controversial axiom. The solution idea can be represented geometrically in the following way. Let $a(\Phi)$ be the utopia point, typically not feasible, which gives the maximum payoff. Now, connect the point of disagreement d and that ideal point $a(\Phi)$ by a line segment. The Kalai-Smorodinsky solution is the maximal point in Φ on that line segment. They replaced Nash arguable fourth axiom by a monotonicity axiom:

- (e) If the set of possible agreements Φ is enlarged such that the maximum utilities of the players remain unchanged, then neither of the players must not suffer from it.

Then, Kalai and Smorodinsky [19] proposed the following solution: we say that there is a unique solution b to the bargaining problem that satisfies the four axioms (a, b, c and e) which is given by the intersection point of the Pareto frontier and the straight line segment connecting d and the utopia point $a(\Phi)$.

Nash [27] showed that there exists a unique standard independent solution for the bargaining model, while Kalai and Smorodinsky [19] showed that a different solution is the unique standard monotonic one.

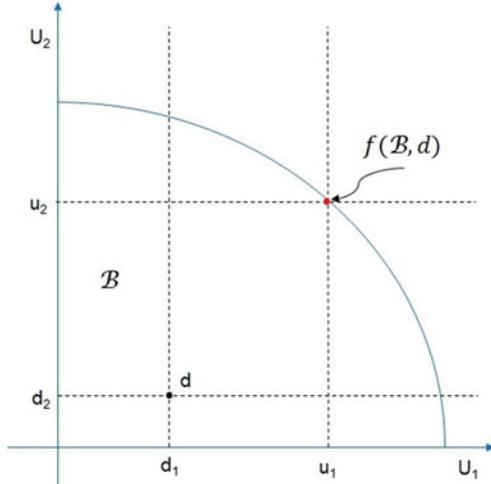
9.2 Motivation

The most basic definition of bargaining refers to a socio-economic class of problems involving several players who cooperate in terms of obtaining a mutually better position of a desirable surplus whose distribution is in conflict. The features of the cooperation of the players in terms of reaching an agreement and the initial situations of the players in the status-quo before an agreement has effect will determine how the surplus will be distributed. Several social, political and economic problems are related to the bargaining problem.

For instance, let us consider the case of selling a used car. When it comes to selling the car, the seller naturally wants to obtain the most money possible. It is practical to trade the car at a dealer or make a quick sale to a used car dealership, but these options usually leave the seller with significantly less than what the car is actually worth. Selling a car by himself allows the seller to get its full value. Then, the seller values his car at 3,000 which is the minimum price at which he would sell it. On the other hand, there exists a buyer that values the car at 5,000 which is the maximum price at which he would buy it. If trade occurs, the price lies between 3,000 and 5,000, then both the seller and the buyer would become better-off and a conflict of interests arises. In any trade the seller and the buyer have the possibility of achieving a mutually beneficial agreement by having conflicting interests over the terms of the trade.

The formal theory of bargaining originated with John Nash's work in the early 1950's [27, 28]. The term bargaining is usually employed to refer to a situations in which players have the possibility of achieving a mutually beneficial agreement, there is a conflict of interests about which an agreement should conclude, and no agreement may be imposed on any individual without their approval.

Let us consider two players $l = 1, 2$. A *bargaining* problem is a pair $\mathcal{B} = (\Phi, d)$ in the utility space were Φ is a set of possible agreements in terms of utilities u that player 1 and player 2 can yield. The player's utility function u^l is strictly increasing and concave. The set of possible agreements is Φ , which is a compact and convex set of \mathbb{R}^2 . An element of Φ is a pair $u = (u^1, u^2) \in \Phi$ and $d = (d^1, d^2)$ is called the

Fig. 9.1 Bargaining model

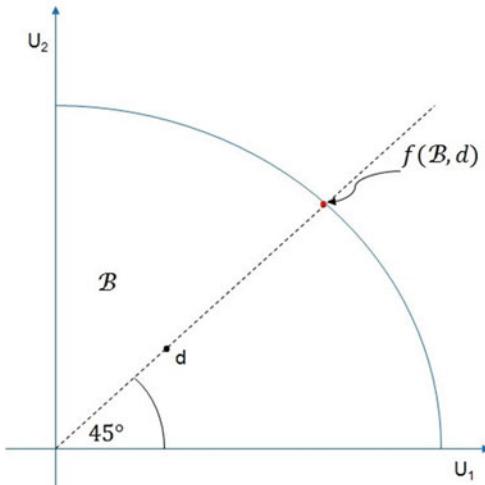
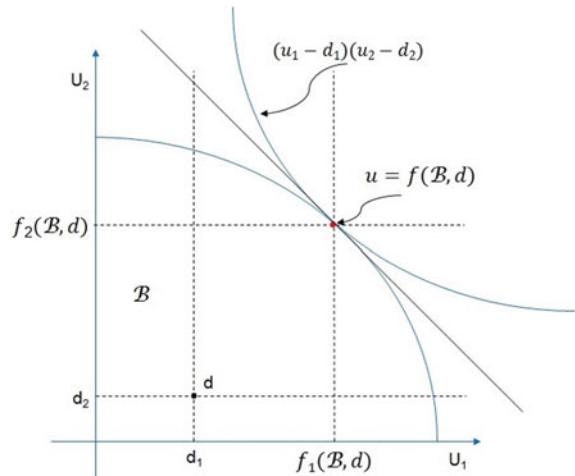
disagreement utility point. Compactness arises from the assumptions related to closed production sets and bounded factor endowments. Convexity is obtained from the fact that expected utility over outcomes. Also, the set Φ involves points that dominate the disagreement point, i.e., there is a positive surplus to be enjoyed if agreement is reached. The function f takes as input any bargaining problem and returns a pair of utilities $u = (u^1, u^2) \in \Phi$. When we need to refer to the components of f , we write $u^1 = f^1(\mathcal{B})$ and $u^2 = f^2(\mathcal{B})$. The interpretation is that given a bargaining problem $\mathcal{B} = (\Phi, d)$ there exists an agreement $u = f(\Phi, d) \in \Phi$ such that $u^1 \geq d^1$ and $u^2 \geq d^2$ which ensures that there exists a mutually beneficial agreement. Figure 9.1 shows the bargaining problem.

Two fundamental axioms impose the most important restrictions over the solution of the bargaining problem (see Fig. 9.2). Pareto optimality: the function $f(\Phi, d)$ has the property that there does not exist a point $u = (u^1, u^2) \in \Phi$ such that $u^1 \geq f^1(\Phi, d)$ and $u^2 \geq f^2(\Phi, d)$ such that $(u^1, u^2) \neq f(\Phi, d)$. Symmetry: suppose that \mathcal{B} is such that U is symmetric around the 45° line and $d^1 = d^2$, then $f^1(\mathcal{B}) = f^2(\mathcal{B})$. The rest of the axioms will be presented in the formalization of the model.

Nash's bargaining solution of the bargaining situation described above is the unique pair of utilities, denoted by $(u^1, u^2) \in \Phi$, that solves the following maximization problem

$$\max_{u^1, u^2 \in \Theta} (u^1 - d^1)(u^2 - d^2)$$

where $\Theta \equiv \{(u^1, u^2) \in \Phi \mid u^1 \geq d^1 \text{ and } u^2 \geq d^2\}$. The maximization problem stated above has a unique solution, because the result of $(u^1 - d^1)(u^2 - d^2)$, which is referred to as the Nash product, is continuous and strictly concave. Figure 9.3 illustrates Nash's bargaining solution.

Fig. 9.2 Bargaining axioms**Fig. 9.3** Nash's bargaining solution

9.3 Preliminaries

In this section we introduce the (continuous-time, time-homogeneous) game model in which we are interested and the formulation of the problem [8]. As usual, \mathbb{R} and \mathbb{N} stand for the sets of real numbers and nonnegative integers, respectively.

Throughout the remainder

$$\mathcal{G} = (\mathcal{N}, S^l, A^l, \{A^l(s)\}_{s \in S}, V^l, Q^l)_{l=\overline{1, \mathcal{N}}} \quad (9.3.1)$$

stands for a *continuous-time Markov game* (CTMG), where $\mathcal{N} = \{1, \dots, n\}$ is the set of players and each player is indexed by $l = \overline{1, n}$, the state space S^l is a *finite* set $\{s_{(1)}^l, \dots, s_{(N)}^l\}$, $N \in \mathbb{N}$, endowed with the discrete topology and the action set A^l is the action (or control) space, a *finite space* endowed with the corresponding Borel σ -algebra $\mathcal{B}(A^l)$.

For each $s^l \in S^l$, $A^l(s^l) \subset A^l$ is the nonempty set of admissible actions at s^l and we shall suppose that it is *compact*. Whereas, the set $\mathbb{K}^l := \{(s^l, a^l) : s^l \in S^l, a^l \in A^l(s^l)\}$ is the class of admissible pairs, which is considered as a topological subspace of $S^l \times A^l$ and, similarly, the set $\mathbb{K} := \{\mathbf{k} : \mathbf{k} \in \times_{l=1}^n \mathbb{K}^l\}$. $V^l \in \mathcal{B}(\times_{l=1}^n S^l \times \mathbb{K}^l)$ is the (measurable) one-stage cost function.

The function Q^l in (9.3.1) is the matrix $[q_{j_l|i_l k_l}^l]$ of the game's transition rates, satisfying $q_{j_l|i_l k_l}^l \geq 0$ for all $(s^l, a^l) \in \mathbb{K}^l$ and $i \neq j$ such that

$$[q_{j_l|i_l k_l}^l] = \begin{cases} -\sum_{i \neq j}^N \lambda_{i,j}^l(a^l), & \text{if } i = j, \\ \lambda_{i,j}^l(a^l), & \text{if } i \neq j, \end{cases}$$

where $\lambda_{i,j}^l$ is a transition rate between state i and j , $\lambda_i^l = \sum_{j \neq i}^N \lambda_{i,j}^l$. This matrix is assumed to be conservative $\sum_{j_l=1}^{N_l} q_{j_l|i_l k_l}^l = 0$ and stable, which means that

$$q_{(i_l)}^{l*} := \sup_{a^l \in A^l} q_{(i_l)}^l(a^l) < \infty \quad \forall i_l \in S^l,$$

where $q_{i_l}^l(a_l) := -q_{i_l, i_l}^l(a_l) \geq 0$ for all $a^l \in A^l$.

Now, we denote the probability *transition matrix* by

$$\Pi^l(t) = [\pi_{s,i_l,\tau,j_l,k_l}^l]_{i_l,j_l,k_l}, \quad \tau \geq s$$

such that,

$$\pi_{s,i_l,\tau,j_l,k_l}^l = \pi_{0,i_l,t,j_l,k_l}^l, \quad t = \tau - s \quad \forall i_l, j_l \in S^l$$

and $\sum_{j_l=1}^{N_l} \pi_{j_l|i_l k_l}^l = 1$.

The Kolmogorov forward equations can be written as the matrix differential equation as follows:

$$\Pi'(t) = \Pi(t)Q; \quad \Pi(0) = I,$$

$\Pi(t) \in \mathbb{R}^{N \times N}$, $I \in \mathbb{R}^{N \times N}$ is the identity matrix. This system can be solved by

$$\Pi(t) = \Pi(0)e^{Qt} = e^{Qt} := \sum_{t=0}^{\infty} \frac{t^n Q^n}{n!}, \quad (9.3.2)$$

and at the stationary state, the probability transition matrix is defined as

$$\Pi^* = \lim_{t \rightarrow \infty} \Pi(t).$$

Definition 9.1 The vector $P^l \in \mathbb{R}^N$ is called *stationary distribution vector* if

$$(\Pi^{l\top})^* P^l = P^l,$$

where $\sum_{i_l=1}^{N_l} P_{(i_l)}^l = 1$.

This vector can be seen as the long-run proportion of time that the process is in state $i_l \in S^l$.

Theorem 9.1 *The following statements are equivalent:*

- $Q^{l\top} P^l = 0$
- $\Pi^{l\top}(t) P^l = P^l; \forall t \geq 0$

The proof of this fact is easy in the case of a finite state space, recalling the Kolmogorov backward equation.

A *strategy* for player l is then defined as a sequence $d^l = \{d^l(t), t \geq 0\}$ of stochastic kernels $d^l(t)$ such that:

- (a) for each time $t \geq 0$, $d_{k_l|i_l}^l(t)$ is a probability measure on A^l such that $d_{A^l(i_l)|i_l}^l(t) = 1$ and,
- (b) for every $E^l \in \mathcal{B}(A^l)$ $d_{E^l|i_l}^l(t)$ is a Borel-measurable function in $t \geq 0$.

We denoted by D^l the family of all strategies for player l . A multistrategy is a vector $\mathbf{d} = (d^1, \dots, d^n) \in D := \bigotimes_{l=1}^n D^l$. From now on, we will consider only stationary strategies $d_{k_l|i_l}^l(t) = d_{k_l|i_l}^l$. For each strategy $d_{k_l|i_l}^l$ the associated transition rate matrix is defined as:

$$Q^l(d^l) := [q_{j_l|i_l}^l(d^l)] = \sum_{k_l=1}^{M_l} q_{j_l|i_l k_l}^l d_{k_l|i_l}^l$$

such that on a stationary state distribution for all $d_{k_l|i_l}^l$ and $t \geq 0$ we have that $\Pi^{l*}(d) = \lim_{t \rightarrow \infty} e^{Q^l(d^l)t}$, where $\Pi^{l*}(d^l)$ is a stationary transition controlled matrix.

The cost function of each player, depending on the states and actions of all the other players, is given by the values $W_{i_1, k_1; \dots; i_n, k_n}^l$, so that the “average cost function” \mathbf{V}^l in the stationary regime can be expressed as

$$\mathbf{V}^l(\mathbf{d}) := \sum_{i_1, k_1}^{N_1, M_1} \dots \sum_{i_n, k_n}^{N_n, M_n} W_{i_1, k_1; \dots; i_n, k_n}^l \prod_{l=1}^n d_{k_l|i_l}^l P^l(s^l = s_{(i_l)}) ,$$

where

$$W_{i_1, k_1, \dots, i_n, k_n}^l = \sum_{j_1}^{N_1} \dots \sum_{j_n}^{N_n} V_{i_1, j_1, k_1, \dots, i_n, j_n, k_N}^l \prod_{l=1}^n \pi_{j_l | i_l k_l}^l.$$

Given that

$$c_{i_l k_l}^l = d_{k_l | i_l}^l P^l(s^l = s_{(i_l)}),$$

we have

$$\mathbf{V}^l(\mathbf{c}) := \sum_{i_1, k_1}^{N_1, M_1} \dots \sum_{i_n, k_n}^{N_n, M_n} W_{i_1, k_1, \dots, i_n, k_n}^l \prod_{l=1}^n c_{i_l k_l}^l, \quad (9.3.3)$$

$$\mathbf{c} = (c^1, \dots, c^n).$$

The variable $c_{i_l k_l}^l$ satisfies the following restrictions [11, 33]:

1. Each vector from the matrix $c^l := [c_{i_l k_l}^l]_{i_l=1, N_l; k_l=1, M_l}$ that represents a stationary mixed-strategy that belongs to the simplex

$$\mathcal{S}^{N_l \times M_l} := \begin{cases} c_{i_l k_l}^l \in \mathbb{R}^{N_l \times M_l} \text{ for } c_{i_l k_l}^l \geq 0, \\ \text{where } \sum_{i_l k_l}^{N_l, M_l} c_{i_l k_l}^l = 1. \end{cases} \quad (9.3.4)$$

2. The variable $c_{i_l k_l}^l$ satisfies the continuous time and the ergodicity constraints, and belongs to the convex, closed and bounded set defined as follows:

$$c^l \in C_{adm}^l = \begin{cases} h_{j_l}^l(c^l) = \sum_{i_l k_l}^{N_l, M_l} \pi_{j_l | i_l k_l}^l c_{i_l k_l}^l - \sum_{k_l}^{M_l} c_{j_l, k_l}^l = 0, \\ \sum_{i_l k_l}^{N_l, M_l} q_{j_l | i_l k_l}^l c_{i_l k_l}^l = 0. \end{cases} \quad (9.3.5)$$

Notice that by (9.3.5) it follows that

$$P^l(s^l = s_{(i_l)}) = \sum_{k_l}^{M_l} c_{i_l k_l}^l, \quad d_{k_l | i_l}^l = \frac{c_{i_l k_l}^l}{\sum_{k_l}^{M_l} c_{i_l k_l}^l}. \quad (9.3.6)$$

In the ergodic case $\sum_{k_l}^{M_l} c_{i_l k_l}^l > 0$ for all $l = \overline{1, n}$.

9.4 The Nash Bargaining Model

The *Nash bargaining* solution is based on a model in which the players are assumed to negotiate on which point of the set of feasible payoffs $\Phi \subset \mathbb{R}^n$ will be agreed upon and realized by concerted actions of the members of the coalition $l = 1, \dots, n$. A pivotal element of the model is a fixed disagreement vector $d \in \mathbb{R}^n$ which plays the role of a deterrent. If negotiations break down and no agreement is reached, then the disagreement point will take effect. The players are committed to the disagreement point in the case of failing to reach a consensus on which feasible payoff to realize. Thus the whole bargaining problem \mathcal{B} will be concisely given by the pair $\mathcal{B} = (\Phi, d)$. We will call this form the condensed form of the bargaining problem (see [17, 27]).

A bargaining problem can be derived from the normal form of an n -person game $\mathcal{G} = C^1, \dots, C^n; u^1, \dots, u^n$ in a natural way. The set of all feasible payoffs (outcomes) is defined as

$$\Theta = u : u = (u^1(c), \dots, u^n(c)), c \in C$$

where $C = C^1 \times \dots \times C^n$.

Given a disagreement vector $d \in \mathbb{R}^n$, $\mathcal{B} = (\Theta, d)$ it is a bargaining problem in condensed form. We can derive another bargaining problem $\mathcal{B} = (\Phi, d)$ from \mathcal{G} by extending the set of feasible outcomes Θ to its convex hull Φ . Notice that any element $\varphi \in \Phi$ can be represented as

$$\varphi = \sum_{l=1}^n \lambda^l u^l.$$

where $u = (u^1(c), \dots, u^n(c))$, $(c \in C)$, $\lambda^l \geq 0$ for all l , and $\sum_{l=1}^n \lambda^l = 1$.

The payoff φ can be realized by playing the strategies c with probability λ , and so φ is the expected payoff of the players. Thus, when the players face the bargaining problem \mathcal{B} the question is, which point of Φ should be selected, taking into account the different position and strength of the players that is reflected in the set Φ of extended payoffs and the disagreement point d .

Nash approached this problem by assigning a one-point solution to \mathcal{B} in an axiomatic manner. Let \mathcal{B} denote the set of all pairs (Φ, d) such that

1. $\Phi \subset \mathbb{R}^n$ is compact, convex;
2. there exists at least one $u \in \Phi$ such that $u > d$.

A Nash solution to the bargaining problem is a function $f : \mathcal{B} \rightarrow \mathbb{R}^n$ such that $f(\Phi, d) \in \Phi$. We shall confine ourselves to functions satisfying the following axioms and we still call them functions solution (see [17, 19, 25, 27]).

1. Feasibility: $f(\Phi, d) \in \Phi$.
2. Rationality: $f(\Phi, d) \geq d$.

3. Pareto Optimality: For every $(\Phi, d) \in \mathcal{B}$ there is $u \in \Phi$ such that $u \geq f(\Phi, d)$ and imply $u = f(\Phi, d)$.
4. Symmetry: If for a bargaining problem $(\Phi, d) \in \mathcal{B}$, there exist indices i, j such that $\varphi = (\varphi^1, \dots, \varphi^n) \in \Phi$ if and only if $\bar{\varphi} = (\bar{\varphi}^1, \dots, \bar{\varphi}^n) \in \Phi$, $(\bar{\varphi}^l = \varphi^l, l \neq i, l \neq j, \bar{\varphi}^i = \varphi^j, \bar{\varphi}^j = \varphi^i)$ and $d^i = d^j$ for $d = (d^1, \dots, d^n)$, then $f^i = f^j$ for the solution vector $f(\Phi, d) = (f^1, \dots, f^n)$.
5. Invariance with respect to affine transformations of utility: Let $\alpha^l > 0, \beta^l, (l = 1, \dots, n)$ be arbitrary constants and let

$$d' = (\alpha^1 d^1 + \beta^1, \dots, \alpha^n d^n + \beta^n) \quad \text{with } d = (d^1, \dots, d^n)$$

and

$$\Phi' = (\alpha^1 \varphi^1 + \beta^1, \dots, \alpha^n \varphi^n + \beta^n) : (\varphi^1, \dots, \varphi^n) \in \Phi.$$

Then $f(\Phi', d') = (\alpha^1 d^1 + \beta^1, \dots, \alpha^n d^n + \beta^n)$, where $f(\Phi, d) = (f^1, \dots, f^n)$.

6. Independence of irrelevant alternatives: If (Φ, d) and (T, d) are bargaining pairs such that $\Phi \subset T$ and $f(T, d) \in \Phi$, then $f(T, d) = f(\Phi, d)$.

Theorem 9.2 *There is a unique function f satisfying axioms 1-6, furthermore for all $(\Phi, d) \in \mathcal{B}$, the vector $f(\Phi, d) = (f^1, \dots, f^n) = (u^1, \dots, u^n)$ is the unique solution of the optimization problem*

$$\begin{aligned} &\text{maximize} && g(u) = \prod_{l=1}^n (u^l - d^l) \\ &\text{subject to} && u \in \Phi, u \geq d. \end{aligned} \tag{9.4.1}$$

The objective function of problem (9.4.1) is usually called the *Nash product*.

Proof See [17]. □

Remark 9.1 There are exactly two solutions satisfying axioms 1, 2, 4, 5, and 6. One is the Nash's solution and the other is the disagreement solution.

For the next conjectures consider a bargaining problem as a pair (Φ, d) where $\Phi \subset \mathbb{R}^2$ and $d \in \mathbb{R}^2$.

Conjecture 9.4.1 The Pareto frontier \mathcal{Q}^e of the set Φ is the graph of a concave function, denoted by h , whose domain is a closed interval $\mathcal{B} \subseteq R$. Furthermore, there exists $f^1 \in \mathcal{B}$ such that $u^1 > d^1$ and $h(u^1) > d^2$ [25].

Conjecture 9.4.2 The set \mathcal{Q}^w of weakly Pareto efficient utility pairs is closed [25].

9.5 The Kalai-Smorodinsky Bargaining Model

With the property of independence of irrelevant alternatives, Nash's solution is not sensitive to the range of outcomes contained in the feasible set, for instance, by the utopia point $a(\Phi) = (a^1(\Phi), \dots, a^n(\Phi))$ defined by

$$a^l(\Phi) = \max\{u^l | u^l \in \Phi, u \geq d\}$$

this point is the highest possible utility payoff player l can attain in the bargaining problem (Φ, d) . Raiffa [34] proposed a solution for two-player games which is sensitive to changes in $a(\Phi)$, he proposed the solution u for two-player games such that $u = f(\Phi, d)$ is the Pareto-optimal point at which $(u^1 - d^1)/(a^1 - d^1) = (u^2 - d^2)/(a^2 - d^2)$. The solution u selects the maximal point on the line joining d to a , yielding each player the largest reward consistent with the constraint that the players' actual gains should be in proportion to their maximum gains, as measured by the ideal point $a(\Phi)$.

The Kalai-Smorodinsky solution of the bargaining problem amounts to normalizing the utility function of each agent in such a way that it is worth zero at the status-quo and one at this agent's best outcome, given that all others get at least their status quo utility level; and to sharing equally the benefit from cooperation. This solution has been axiomatically characterized when society n contains only two agents, i.e., $l = 1, 2$. To every two-person game we associate a pair (Φ, d) , where d is a point in the plane $d = (d^1, d^2)$ is the status quo and Φ is a subset of the plane, every point $u = (u^1, u^2) \in \Phi$ represents levels of utility for players 1 and 2 that can be reached by an outcome of the game which is feasible for the two players when they do cooperate.

Let \mathcal{B} denote the set of all pairs (Φ, d) such that

1. $\Phi \subset \mathbb{R}^2$ is compact, convex;
2. there exists at least one point $u \in \Phi$ such that $u^l > d^l$, for $l = 1, 2$.

A solution to the bargaining problem is a function $f : \mathcal{B} \rightarrow \mathbb{R}^2$ such that $f(\Phi, d) \in \Phi$. We shall confine ourselves to functions satisfying the following axioms and we still call them functions solution (see [19]).

1. Pareto Optimality: For every $(\Phi, d) \in \mathcal{B}$ there is no $u \in \Phi$ such that $u \geq f(\Phi, d)$ and imply $u \neq f(\Phi, d)$.
2. Symmetry: We let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be defined by $T((u^1, u^2)) = (u^2, u^1)$ and we require that for every $(\Phi, d) \in \mathcal{B}$, $f(T(\Phi), T(d)) = T(f(\Phi, d))$
3. Invariance with respect to affine transformations of utility: A is an affine transformation of utility if $A = (A^1, A^2) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, $A((u^1, u^2)) = (A^1(u^1), A^2(u^2))$, and the maps $A^l(u)$ are of the form $c^l u + d^l$ for some positive constant c^l and some constant d^l . We require that for such a transformation A , $f(A(\Phi), A(d)) = A(f(\Phi, d))$.
4. Monotonicity: For a pair $(\Phi, d) \in \mathcal{B}$, let $a(\Phi) = (a^1(\Phi), a^2(\Phi))$ and $g_\Phi(u^1)$ be a function defined for $u^1 \leq a^1(\Phi)$ in the following way

$$\begin{aligned} g_{\Phi}(u^1) &= u^2 \text{ if } (u^1, u^2) \text{ is the Pareto of } (\Phi, d) \\ &= a^2(\Phi) \text{ if there is no such } u^2. \end{aligned}$$

If (Φ^2, d) and (Φ^1, d) are bargaining pairs such that $a^1(\Phi^1) = a^1(\phi^2)$ and $g_{\Phi^1} \leq g_{\Phi^2}$, then $f^2(\Phi^1, d) \leq f^2(\Phi^2, d)$, where $f(\Phi, d) = (f^1(\Phi, d), f^2(\Phi, d))$.

The axiom of monotonicity states that if, for every utility level that player 1 may demand, the maximum feasible utility level that player 2 can simultaneously reach is increased, then the utility level assigned to player 2 according to the solution should also be increased.

Theorem 9.3 *Let f be a bargaining solution. Then f satisfies Pareto optimality, symmetry, invariance with respect to affine transformations of utility and monotonicity if, and only if, f is Kalai-Smorodinsky solution. (See the proof in Roth [34]).*

9.5.1 The n -Person Kalai-Smorodinsky Solution

Kalai and Smorodinsky [19] defined their solution only on two-player bargaining problems. We consider the set of all n -player bargaining problems defined by Peters and Tijs [31], and on this set we define a class of asymmetric n -person Kalai-Smorodinsky solutions. The set of players is denoted by $l = (1, \dots, n)$, with $n \geq 2$. A set $\Phi \subseteq \mathbb{R}^n$ is comprehensive if $x \in \Phi$ and $x \geq y$ imply $y \in \Phi$, for all $x, y \in \mathbb{R}^n$. A bargaining problem for n is a pair (Φ, d) where:

1. $\Phi \subseteq \mathbb{R}^n$ is compact, convex, and comprehensive,
2. there exists a $u \in \Phi$ such that $u > d$ and $d \in \Phi$,

We talk about comprehensiveness in the sense that any player can choose a lower utility payoff without this leading to an infeasible outcome. Players seek agreement on an outcome $u \in \Phi$, yielding utility u^l to player l . In case no agreement is reached the disagreement outcome d results. For all bargaining problem $(\Phi, d) \in \mathcal{B}$ we define the Pareto set of Φ as

$$P(\Phi) = \{u \in \Phi \mid \text{for all } x \in \mathbb{R}^n, \text{ if } x \geq u \text{ and } x \neq u, \text{ then } x \notin \Phi\}.$$

A bargaining solution is a map $f : \mathcal{B} \rightarrow \mathbb{R}^n$ that assigns to each bargaining problem $(\Phi, d) \in \mathcal{B}$ a single point $f(\Phi, d) \in \Phi$.

Raiffa [34] and Kalai and Smorodinsky [19] defined and characterized the Kalai-Smorodinsky solution for two-person bargaining problems. Roth [35] observed that the n -player extension of the solution is not Pareto-optimal on all bargaining problems in \mathcal{B} , i.e., does not assign an element of $P(\Phi)$ to each $(\Phi, d) \in \mathcal{B}$. Therefore, Peters and Tijs [31] introduced a subclass of bargaining problems in \mathcal{B} for which this shortcoming does not occur.

Theorem 9.4 *For bargaining games with three or more players, no solution exists which possesses the properties of Pareto optimality, symmetry, and restricted monotonicity. (See the proof in Roth [35]).*

Condition For all $u \in \Phi, u \geq d, l = (1, \dots, n)$: $u \notin P(\Phi)$ and $u^l < a^l(\Phi) \Rightarrow \exists \varepsilon > 0$ with $u + \varepsilon e^l \in \Phi$, where the vector e^l in \mathbb{R}^n has the l -th coordinate equal to 1 and all other coordinates equal to 0.

If a feasible outcome u is not Pareto optimal, then for any player l who receives less than his utopia payoff it is possible to increase his utility while all other players still receive u . Let $\mathcal{I} \subseteq \mathcal{B}$ consist of all bargaining problems satisfying Condition 9.5.1. The class of bargaining problems $(\Phi, 0) \in \mathcal{I}$ is denoted by \mathcal{I}_0 .

Peters and Tijs [31] defined the n -player extension of the solution by making use of monotonic curves. A monotonic curve for n is a map

$$\psi : [1, n] \rightarrow \left\{ u \in \mathbb{R}_+^n \mid u^l \leq 1 \text{ for all player } l, \text{ and } 1 \leq \sum_{l=1}^n u^l \right\}$$

such that for all $1 \leq s \leq t \leq n$ we have $\psi(s) \leq \psi(t)$ and $\sum_{l=1}^n \psi^l(s) = s$. The set of all monotonic curves for n is denoted by Ψ .

Lemma 9.1 *For each $\psi \in \Psi$ and $(\Phi, 0) \in \mathcal{I}_0$ with $a(\Phi, 0) = e^n$, the set*

$$P(\Phi) \cap \{\psi(t) \mid t \in [1, n]\}$$

contains exactly one point (see [31]).

Let ψ be some monotonic curve in Ψ . Following the Lemma 9.1 we can define $\rho^\psi : \mathcal{I} \rightarrow \mathbb{R}^n$, the solution associated with ψ . Let $(\Phi, 0) \in \mathcal{I}_0$; if $a(\Phi, 0) = e^n$, then

$$\{\rho^\psi(\Phi, 0)\} := P(\Phi) \cap \{\psi(t) \mid t \in [1, n]\}$$

and if $a(\Phi, 0) = a$, then $\rho^\psi(\Phi, 0) := a\rho^\psi(a^{-1}\Phi)$. For $(\Phi, d) \in \mathcal{I}$, we define

$$\rho^\psi(\Phi, d) = d + \rho^\psi(\Phi \hat{a}' - d).$$

The class of all solutions associated with a monotonic curve in Ψ is referred to as the class of individually monotonic bargaining solutions, the Kalai–Smorodinsky solution is an element of this class. Observe that $\hat{\psi}$ defines a straight line in \mathbb{R}^n , which for bargaining games $(\Phi, 0) \in \mathcal{I}_0$ with $a(\Phi, 0) = e^n$, coincides with the line connecting the disagreement point 0 and the utopia point e^n . For general bargaining problems $(\Phi, d) \in \mathcal{I}$, the solution is the intersection of the Pareto set $P(\Phi)$ and the straight line that connects the disagreement point d and the utopia point $a(\Phi, d)$.

9.6 The Bargaining Solver

9.6.1 The Nash Bargaining Solver

Stated in general terms, a n -person bargaining situation is a situation in which n players have a common interest to cooperate, but have conflicting interests over exactly how to cooperate. This process involves the players making offers and counteroffers to each other.

Consider a n -person bargaining problem. Let us denote the disagreement utility that depends on the strategies $c_{(i_l, k_l)}^l$ as $d^l(c^1, \dots, c^n)$ for each player ($l = 1, \dots, n$), and the solution for the Nash bargaining problem as the point (u^1, \dots, u^n) . Following (9.3.3) the utilities u^l , in the same way that the disagreement utilities, are for Markov chains as follows

$$u^l = u^l(c^1, \dots, c^n) := \sum_{i_1, k_1}^{N_1, M_1} \dots \sum_{i_n, k_n}^{N_n, M_n} W_{i_1, k_1, \dots, i_n, k_n}^l \prod_{l=1}^n c_{i_l k_l}^l, \quad (9.6.1)$$

where the matrices W^l represent the behavior of each player. This point is better than the disagreement point, therefore must satisfy that $u^l > d^l$.

The function for finding the solution to the Nash bargaining problem is

$$g(c^1, \dots, c^n) = \prod_{l=1}^n (u^l - d^l)^{\alpha^l} \chi(u^l > d^l), \quad (9.6.2)$$

where $\alpha^l \geq 0$ and $\sum_l \alpha^l = 1$, ($l = 1, \dots, n$), which are weighting parameters for each player. We can rewrite (9.6.2) for purposes of implementation as follows

$$\tilde{g}(c^1, \dots, c^n) = \sum_{l=1}^n \alpha^l \chi(u^l > d^l) \ln(u^l - d^l). \quad (9.6.3)$$

Thus, the strategy x^* , which is the vector $x^* = (c^1, \dots, c^n) \in X_{adm} := \bigotimes_{l=1}^n C_{adm}^l$, is the solution for the Nash bargaining problem

$$x^* \in \operatorname{Arg} \max_{x \in X_{adm}} \{\tilde{g}(c^1, \dots, c^n)\},$$

the strategies c^l satisfy the restrictions (9.3.4) and (9.3.5). Applying the Lagrange principle, (see, for example, [32, 33]) let us introduce the Lagrange function

$$\mathcal{L}(x, \mu, \xi, \eta) = \tilde{g}(c^1, \dots, c^n) - \sum_{l=1}^n \sum_{j_l=1}^{N_l} \mu_{(j_l)}^l h_{(j_l)}^l(c^l) -$$

$$\sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{j_l=1}^{N_l} \sum_{k_l=1}^{M_l} \xi_{(j_l)}^l q_{j_l|i_l k_l}^l c_{i_l k_l}^l - \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \eta^l (c_{i_l k_l}^l - 1).$$

The approximative solution obtained by the Tikhonov's regularization with $\delta > 0$ (see [33]) is given by

$$x^*, \mu^*, \xi^*, \eta^* = \arg \max_{x \in X_{adm}} \min_{\mu, \xi, \eta \geq 0} \mathcal{L}_\delta(x, \mu, \xi, \eta),$$

where

$$\begin{aligned} \mathcal{L}_\delta(x, \mu, \xi, \eta) &= \tilde{g}(c^1, \dots, c^n) - \sum_{l=1}^n \sum_{j_l=1}^{N_l} \mu_{(j_l)}^l h_{(j_l)}^l(c^l) - \\ &\quad \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{j_l=1}^{N_l} \sum_{k_l=1}^{M_l} \xi_{(j_l)}^l q_{j_l|i_l k_l}^l c_{i_l k_l}^l - \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \eta^l (c_{i_l k_l}^l - 1) - \\ &\quad \frac{\delta}{2} (\|x\|^2 - \|\mu\|^2 - \|\xi\|^2 - \|\eta\|^2). \end{aligned} \tag{9.6.4}$$

Notice that the Lagrange function (9.6.4) satisfies the saddle-point [32] condition, namely, for all $x \in X_{adm}$ and $\mu, \xi, \eta \geq 0$ we have

$$\mathcal{L}_\delta(x_\delta, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_\delta(x_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_\delta(x_\delta^*, \mu_\delta, \xi_\delta, \eta_\delta).$$

9.6.2 Kalai-Smorodinsky Solver

Consider a n -person bargaining problem. Let us denote the disagreement utility that depends on the strategies $c_{i_l k_l}^l$ as $d^l(c^1, \dots, c^n)$ for each player ($l = 1, \dots, n$), and the solution for the bargaining problem as the point (u^1, \dots, u^n) . Following (9.3.3) the utilities u^l are for Markov chains as follows

$$u^l = u^l(c^1, \dots, c^n) := \sum_{i_1, k_1}^{N_1, M_1} \dots \sum_{i_n, k_n}^{N_n, M_n} W_{(i_1, k_1, \dots, i_n, k_n)}^l \prod_{l=1}^n c_{i_l k_l}^l, \tag{9.6.5}$$

where the matrices W^l represent the behavior of each player.

The Kalai-Smorodinsky solution chooses the maximum individually rational payoff profile at which each player's payoff has the same proportion from disagreement point to the utopia point. For solving the bargaining problem we consider that there exists an optimal solution that is a strong Pareto optimal point and it is the closest

solution to the utopia point. To find the Pareto optimal solution, we formulate the problem as the L_p -norm that reduces the distance to the utopian point in the Euclidian space. Following [41], the function for finding the solution to the bargaining problem is

$$g(c^1, \dots, c^n) = \left[\sum_{l=1}^n \left| \lambda^l \frac{(u^l - d^l)^{\alpha^l} \chi(u^l > d^l)}{(a^l - d^l)^{\alpha^l} \chi(a^l > d^l)} \right|^p \right]^{1/p}, \quad (9.6.6)$$

where a^l is the utopia point, $\alpha^l \geq 0$ are weighting parameters for each player, and $\lambda \in \Lambda^n$ such that

$$\Lambda^n := \left\{ \lambda \in \mathbb{R}^n : \lambda \in [0, 1], \sum_{l=1}^n \lambda^l = 1 \right\}.$$

We can rewrite (9.6.6) for purposes of implementation as follows

$$\tilde{g}(c^1, \dots, c^n) = \left[\sum_{l=1}^n \left| \lambda^l \left(\alpha^l \chi(u^l > d^l) \ln(u^l - d^l) - \alpha^l \chi(a^l > d^l) \ln(a^l - d^l) \right) \right|^p \right]^{1/p}.$$

Thus, the strategy x^* , which is the vector $x^* = (c^1, \dots, c^n) \in X_{adm} := \bigotimes_{l=1}^n C_{adm}^l$, is the solution for the bargaining problem

$$x^* \in \arg \max_{x \in X_{adm}, \lambda \in \Lambda^n} \{ \tilde{g}(c^1, \dots, c^n) \},$$

the strategies c^l satisfy the restrictions (9.3.4) and (9.3.5). Applying the Lagrange principle let us introduce the Lagrange function

$$\mathcal{L}(x, \lambda, \mu, \xi, \eta) = \tilde{g}(c^1, \dots, c^n) - \sum_{l=1}^n \sum_{j_l=1}^{N_l} \mu_{(j_l)}^l h_{(j_l)}^l(c^l) -$$

$$\sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{j_l=1}^{N_l} \sum_{k_l=1}^{M_l} \xi_{(j_l)}^l q_{j_l|i_l k_l}^l c_{i_l k_l}^l - \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \eta^l (c_{i_l k_l}^l - 1),$$

The approximative solution obtained by the Tikhonov's regularization with $\delta > 0$ is given by

$$x^*, \lambda^*, \mu^*, \xi^*, \eta^* = \arg \max_{x \in X_{adm}, \lambda \in \Lambda^n} \min_{\mu, \xi, \eta \geq 0} \mathcal{L}_\delta(x, \lambda, \mu, \xi, \eta)$$

where

$$\begin{aligned} \mathcal{L}_\delta(x, \lambda, \mu, \xi, \eta) = & \tilde{g}(c^1, \dots, c^n) - \sum_{l=1}^n \sum_{j_l=1}^{N_l} \mu_{(j_l)}^l h_{(j_l)}^l(c^l) - \\ & \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{j_l=1}^{N_l} \sum_{k_l=1}^{M_l} \xi_{(j_l)}^l q_{j_l|i_l k_l}^l c_{i_l k_l}^l - \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \eta^l (c_{i_l k_l}^l - 1) - \\ & \frac{\delta}{2} (\|x\|^2 + \|\lambda\|^2 - \|\mu\|^2 - \|\xi\|^2 - \|\eta\|^2). \end{aligned} \quad (9.6.7)$$

Notice that the Lagrange function (9.6.7) satisfies the saddle-point condition, namely, for all $x \in X_{adm}$, $\lambda \in \Lambda^n$ and $\mu, \xi, \eta \geq 0$ we have

$$\mathcal{L}_\delta(x_\delta, \lambda_\delta, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_\delta(x_\delta^*, \lambda_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_\delta(x_\delta^*, \lambda_\delta^*, \mu_\delta, \xi_\delta, \eta_\delta).$$

9.6.3 The Extraproximal Solver Method

In the proximal format (see, [5]) the relation (9.6.4) can be expressed as

$$\left. \begin{aligned} \mu_\delta^* &= \arg \min_{\mu \geq 0} \left\{ \frac{1}{2} \|\mu - \mu_\delta^*\|^2 + \gamma \mathcal{L}_\delta(x_\delta^*, \mu, \xi_\delta^*, \eta_\delta^*) \right\}, \\ \xi_\delta^* &= \arg \min_{\xi \geq 0} \left\{ \frac{1}{2} \|\xi - \xi_\delta^*\|^2 + \gamma \mathcal{L}_\delta(x_\delta^*, \mu_\delta^*, \xi, \eta_\delta^*) \right\}, \\ \eta_\delta^* &= \arg \min_{\eta \geq 0} \left\{ \frac{1}{2} \|\eta - \eta_\delta^*\|^2 + \gamma \mathcal{L}_\delta(x_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta) \right\}, \\ x_\delta^* &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_\delta^*\|^2 + \gamma \mathcal{L}_\delta(x, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \right\}. \end{aligned} \right\} \quad (9.6.8)$$

For the relation (9.6.7) the proximal format will be extended with $\mathcal{L}_\delta(x, \lambda, \mu, \xi, \eta)$ and the following equation

$$\lambda_\delta^* = \arg \max_{\lambda \in \Lambda^n} \left\{ -\frac{1}{2} \|\lambda - \lambda_\delta^*\|^2 + \gamma \mathcal{L}_\delta(x_\delta^*, \lambda, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \right\},$$

where the solutions $x_\delta^*, \lambda_\delta^*, \mu_\delta^*, \xi_\delta^*$ and η_δ^* depend on the parameters $\delta > 0$ and $\gamma > 0$.

The Extraproximal Method for the conditional optimization problems was suggested in [5, 40]. We design the method for the static bargaining game in a general format with some fixed admissible initial values ($x_0 \in X$, $\lambda_0 \in \Lambda^n$ and $\mu_0, \xi_0, \eta_0 \geq 0$), considering that we want to maximize the function as follows:

1. The *first half-step* (prediction):

$$\left. \begin{aligned} \bar{\mu}_n &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 - \gamma \mathcal{L}_\delta(x_n, \mu, \xi_n, \eta_n) \right\}, \\ \bar{\xi}_n &= \arg \max_{\xi \geq 0} \left\{ -\frac{1}{2} \|\xi - \xi_n\|^2 - \gamma \mathcal{L}_\delta(x_n, \bar{\mu}_n, \xi, \eta_n) \right\}, \\ \bar{\eta}_n &= \arg \max_{\eta \geq 0} \left\{ -\frac{1}{2} \|\eta - \eta_n\|^2 - \gamma \mathcal{L}_\delta(x_n, \bar{\mu}_n, \bar{\xi}_n, \eta) \right\}, \\ \bar{x}_n &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_\delta(x, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}. \end{aligned} \right\} \quad (9.6.9)$$

2. The *second half-step* (basic)

$$\left. \begin{aligned} \mu_{n+1} &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 - \gamma \mathcal{L}_\delta(\bar{x}_n, \mu, \bar{\xi}_n, \bar{\eta}_n) \right\}, \\ \xi_{n+1} &= \arg \max_{\xi \geq 0} \left\{ -\frac{1}{2} \|\xi - \xi_n\|^2 - \gamma \mathcal{L}_\delta(\bar{x}_n, \bar{\mu}_n, \xi, \bar{\eta}_n) \right\}, \\ \eta_{n+1} &= \arg \max_{\eta \geq 0} \left\{ -\frac{1}{2} \|\eta - \eta_n\|^2 - \gamma \mathcal{L}_\delta(\bar{x}_n, \bar{\mu}_n, \bar{\xi}_n, \eta) \right\}, \\ x_{n+1} &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_\delta(x, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}. \end{aligned} \right\} \quad (9.6.10)$$

For the Kalai-Smorodinsky solution the presented extraproximal method will be extended employing the relation (9.6.7) and the following equations:

1. The *first half-step* (prediction):

$$\bar{\lambda}_n = \arg \max_{\lambda \in \Lambda^n} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathcal{L}_\delta(x_n, \lambda, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}.$$

2. The *second half-step* (basic)

$$\lambda_{n+1} = \arg \max_{\lambda \in \Lambda^n} \left\{ -\frac{1}{2} \|\lambda - \lambda_n\|^2 + \gamma \mathcal{L}_\delta(\bar{x}_n, \lambda, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}.$$

The following theorem presents the convergence conditions of (9.6.9) and (9.6.10) and gives the estimate of its rate of convergence for the bargaining equilibrium. As well, we prove that the extraproximal method converges to a unique equilibrium point. Let us define the following extended vectors

$$\tilde{x} = \begin{pmatrix} x \\ \lambda \end{pmatrix} \in \tilde{X} := X \times \mathbb{R}^+, \quad \tilde{\mu} = \begin{pmatrix} \mu \\ \xi \\ \eta \end{pmatrix} \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+.$$

Then, the regularized Lagrange function can be expressed as

$$\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{\mu}) := \mathcal{L}_\delta(x, \lambda, \mu, \xi, \eta).$$

The equilibrium point that satisfies (9.6.8) can be expressed as

$$\begin{aligned}\tilde{\mu}_\delta^* &= \arg \min_{\tilde{\mu} \geq 0} \left\{ \frac{1}{2} \|\tilde{\mu} - \tilde{\mu}_\delta^*\|^2 + \gamma \tilde{\mathcal{L}}_\delta(\tilde{x}_\delta^*, \tilde{\mu}) \right\}, \\ \tilde{x}_\delta^* &= \arg \max_{\tilde{x} \in \tilde{X}} \left\{ -\frac{1}{2} \|\tilde{x} - \tilde{x}_\delta^*\|^2 + \gamma \tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{\mu}_\delta^*) \right\}.\end{aligned}$$

Now, let us introduce the following variables

$$\tilde{y} = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} \in \tilde{X} \times \mathbb{R}^+, \quad \tilde{z} = \begin{pmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{pmatrix} \in \tilde{X} \times \mathbb{R}^+.$$

and let define the Lagrange function in term of \tilde{y} and \tilde{z}

$$L_\delta(\tilde{y}, \tilde{z}) := \mathcal{L}_\delta(\tilde{y}_1, \tilde{z}_2) - \mathcal{L}_\delta(\tilde{z}_1, \tilde{y}_2).$$

For $\tilde{y}_1 = \tilde{x}$, $\tilde{y}_2 = \tilde{\mu}$, $\tilde{z}_1 = \tilde{z}_1^* = \tilde{x}_\delta^*$ and $\tilde{z}_2 = \tilde{z}_2^* = \tilde{\mu}_\delta^*$ we have

$$L_\delta(\tilde{y}, \tilde{z}^*) := \tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{\mu}_\delta^*) - \tilde{\mathcal{L}}_\delta(\tilde{x}_\delta^*, \tilde{\mu}).$$

In these variables the relation (9.6.8) can be represented by

$$\tilde{z}^* = \arg \max_{\tilde{y} \in \tilde{X} \times \mathbb{R}^+} \left\{ -\frac{1}{2} \|\tilde{y} - \tilde{z}^*\|^2 + \gamma L_\delta(\tilde{y}, \tilde{z}^*) \right\}. \quad (9.6.11)$$

Finally, we have that the extraproximal method can be expressed by

1. First step

$$\hat{y}_n = \arg \max_{\tilde{z} \in \tilde{X} \times \mathbb{R}^+} \left\{ -\frac{1}{2} \|\tilde{z} - \tilde{y}_n\|^2 + \gamma L_\delta(\tilde{z}, \tilde{y}_n) \right\}. \quad (9.6.12)$$

2. Second step

$$\tilde{y}_{n+1} = \arg \max_{\tilde{z} \in \tilde{X} \times \mathbb{R}^+} \left\{ -\frac{1}{2} \|\tilde{z} - \tilde{y}_n\|^2 + \gamma L_\delta(\tilde{z}, \hat{y}_n) \right\}. \quad (9.6.13)$$

Lemma 9.2 *Let $\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{\mu})$ be differentiable in \tilde{x} and $\tilde{\mu}$, whose partial derivative with respect to μ satisfies the Lipschitz condition with positive constant K_0 . Then,*

$$\|\tilde{z}_{n+1} - \hat{z}_n\| \leq \gamma K_0 \|\tilde{z}_n - \hat{z}_n\|.$$

Proof See [39, 41]. □

Lemma 9.3 *Let us consider the set of regularized solutions of a non-empty game. The behavior of the regularized function is described by the following inequality:*

$$L_\delta(\tilde{y}, \tilde{y}) - L_\delta(\tilde{z}_\delta^*, \tilde{y}) \geq \delta \|\tilde{y} - \tilde{z}_\delta^*\|$$

for all $\tilde{y} \in \{\tilde{y} \mid \tilde{y} \in X \times \mathbb{R}^+\}$ and $\delta > 0$.

Proof See [41]. □

Theorem 9.5 (Convergence and rate of convergence). *Let $\tilde{\mathcal{L}}_\delta(\tilde{x}, \tilde{\mu})$ be differentiable in \tilde{x} and $\tilde{\mu}$, whose partial derivative with respect to $\tilde{\mu}$ satisfies the Lipschitz condition with positive constant K . Then, for any $\delta > 0$ there exists a small-enough*

$$\gamma_0 = \gamma_0(\delta) < K := \min \left\{ \frac{1}{\sqrt{2}K_0}, \frac{1 + \sqrt{1 + 2(K_0)^2}}{2(K_0)^2} \right\}$$

such that, for any $0 < \gamma \leq \gamma_0$, sequence $\{\tilde{z}_n\}$, which generated by the equivalent extraproximal procedure (9.6.9) and (9.6.10), monotonically converges with exponential rate $r \in (0, 1)$ to a unique equilibrium point \tilde{z}^* , i.e.,

$$\|\tilde{z}_n - \tilde{z}^*\|^2 \leq e^{n \ln r} \|\tilde{z}_0 - \tilde{z}^*\|^2,$$

where

$$r = 1 + \frac{4(\delta\gamma)^2}{1+2\delta\gamma-2\gamma^2K^2} - 2\delta\gamma < 1,$$

and r_{\min} is given by

$$r_{\min} = 1 - \frac{2\delta\gamma}{1+2\delta\gamma} = \frac{1}{1+2\delta\gamma}.$$

Proof Following Theorem 18 in [41] we obtain that

$$r = 1 - 2\gamma\delta + \frac{(2\gamma\delta)^2}{1+2\gamma\delta-2\gamma^2K^2} < 1.$$

Iterating over the previous inequality we have

$$\|\tilde{z}_\delta^* - \tilde{z}_{n+1}\|^2 \leq r \|\tilde{z}_\delta^* - \tilde{z}_n\|^2 \leq \dots \leq e^{n+1 \ln r} \|\tilde{z}_\delta^* - \tilde{z}_0\|^2. \quad (9.6.14)$$

That implies that the series converge and also that the trajectories are bounded. Then, by (9.6.14) we have that

$$\|\tilde{z}_\delta^* - \tilde{z}_{n+1}\|^2 \xrightarrow{n \rightarrow \infty} 0.$$

Given that \tilde{z} is a bounded sequence, by the Weierstrass Theorem there exists a point \tilde{z}' such that any subsequence \tilde{z}_{n_i} satisfies that $\tilde{z}_{n_i} \xrightarrow{n_i \rightarrow \infty} \tilde{z}'$. In addition, we have

that $\|\tilde{z}_{n_i} - \tilde{z}_{n_i+1}\|^2 \rightarrow 0$. Fixing, $n = n_i$ in (9.6.11) and computing the limit when $n_i \rightarrow \infty$ we have

$$\tilde{z}' = \arg \min_{\tilde{y} \in \tilde{X} \times \mathbb{R}^+} \left\{ \frac{1}{2} \|\tilde{y} - \tilde{z}'\|^2 + \gamma L_\delta(\tilde{y}, \tilde{z}') \right\}.$$

Then, we have that $\tilde{z}' = \tilde{z}_\delta^*$, i.e., any limit point of the sequence \tilde{z}_n is a solution of the problem. Given that $\|\tilde{z}_n - \tilde{z}_\delta^*\|^2$ is monotonically decreasing then, there exists a unique limit point (equilibrium point). As a consequence, we have that the sequence \tilde{z}_n satisfies that $\tilde{z}_n \xrightarrow{n \rightarrow \infty} \tilde{z}_\delta^*$ with a convergence velocity of $e^{n \ln r}$.

See the complete proof in [41]. \square

Remark 9.2 The exponential rate $r \in (0, 1)$ satisfies

$$r \simeq r_0 \left(1 + \frac{1}{N^2}\right).$$

9.7 The Model for the Disagreement Point

A pivotal element of the model is a fixed disagreement vector (sometimes also called as status quo or threat point). If negotiations break down and no agreement is reached, then inevitably the disagreement point will take effect. The player are committed to the disagreement point in the case of failing to reach a consensus on which feasible payoff to realize.

Let us introduce the variables (see [40])

$$x := \text{col } c^l, \quad \hat{x} := \text{col } \hat{c}^l, \quad (l = 1, \dots, n).$$

The strategies of the players are denoted by the vector x , and \hat{x} is a strategy of the rest of the players adjoint to x . For reaching the goal of the game, players try to find a join strategy $x^* = (c^1, \dots, c^n) \in X_{adm} := \bigotimes_{l=1}^n C_{adm}^l$ satisfying

$$g(x, \hat{x}) := \sum_{l=1}^n \left[d^l(c^l, \hat{c}^l) - \left(\max_{c^l \in C^l} d^l(c^l, \hat{c}^l) \right) \right]. \quad (9.7.1)$$

Here $d^l(c^l, \hat{c}^l)$ (see 9.3.3) is the utility-function of the player l which plays the strategy $c^l \in C^l$ and the other plays the strategy $\hat{c}^l \in \hat{C}^l$. If we consider the utopia point

$$\bar{c}^l := \arg \max_{c^l \in C^l} d^l(c^l, \hat{c}^l),$$

then, we can rewrite (9.7.1) as follows

$$g(x, \hat{x}) := \sum_{l=1}^n \left[d^l \left(c^l, \hat{c}^l \right) - d^l \left(\bar{c}^l, \hat{c}^l \right) \right]. \quad (9.7.2)$$

The functions $d^l \left(c^l, \hat{c}^l \right)$ ($l = 1, \dots, n$) are assumed to be concave in all their arguments.

Condition The function $g(x, \hat{x})$ satisfies the **Nash condition**

$$d^l \left(c^l, \hat{c}^l \right) - d^l \left(\bar{c}^l, \hat{c}^l \right) \leq 0,$$

for any $c^l \in C^l$ and all $l = 1, \dots, n$

Definition 9.2 A strategy x^* is said to be a Nash equilibrium if

$$x^* \in \arg \max_{x \in X_{adm}} \{g(x, \hat{x})\}.$$

Applying the regularized Lagrange principle we have the solution for the Nash equilibrium

$$x^*, \hat{x}^*, \mu^*, \xi^*, \eta^* = \arg \max_{x \in X, \hat{x} \in \hat{X}} \min_{\mu, \xi, \eta \geq 0} \mathcal{L}_{\theta, \delta}(x, \hat{x}, \mu, \xi, \eta),$$

where

$$\begin{aligned} \mathcal{L}_{\theta, \delta}(x, \hat{x}, \mu, \xi, \eta) := & (1 - \theta)g(x, \hat{x}) - \sum_{l=1}^n \sum_{j_l=1}^{N_l} \mu_{(j_l)}^l h_{(j_l)}^l(c^l) - \\ & \left. \left\{ \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{j_l=1}^{N_l} \sum_{k_l=1}^{M_l} \xi_{(j_l)}^l q_{j_l | i_l k_l}^l c_{i_l k_l}^l - \sum_{l=1}^n \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \eta^l (c_{i_l k_l}^l - 1) - \right. \right. \\ & \left. \left. \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 - \|\mu\|^2 - \|\xi\|^2 - \|\eta\|^2) \right. \right. \end{aligned} \quad (9.7.3)$$

Notice also that the Lagrange function (9.7.3) satisfies the saddle-point condition, namely, for all $x \in X, \hat{x} \in \hat{X}$, and $\mu, \xi, \eta \geq 0$ we have

$$\mathcal{L}_{\theta, \delta}(x_\delta, \hat{x}_\delta, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \leq \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}_\delta^*, \mu_\delta, \xi_\delta, \eta_\delta).$$

9.7.1 The Extraproximal Solver Method

In the proximal format the relation (9.7.3) can be expressed as

$$\left. \begin{aligned} \mu_\delta^* &= \arg \min_{\mu \geq 0} \left\{ \frac{1}{2} \|\mu - \mu_\delta^*\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}_\delta^*, \mu, \xi_\delta^*, \eta_\delta^*) \right\}, \\ \xi_\delta^* &= \arg \min_{\xi \geq 0} \left\{ \frac{1}{2} \|\xi - \xi_\delta^*\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}_\delta^*, \mu_\delta^*, \xi, \eta_\delta^*) \right\}, \\ \eta_\delta^* &= \arg \min_{\eta \geq 0} \left\{ \frac{1}{2} \|\eta - \eta_\delta^*\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta) \right\}, \\ x_\delta^* &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_\delta^*\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x, \hat{x}_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \right\}, \\ \hat{x}_\delta^* &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\frac{1}{2} \|\hat{x} - \hat{x}_\delta^*\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x_\delta^*, \hat{x}, \mu_\delta^*, \xi_\delta^*, \eta_\delta^*) \right\}, \end{aligned} \right\}$$

where the solutions x_δ^* , $\hat{x}_\delta^*(u)$, μ_δ^* , ξ_δ^* and η_δ^* depend on the parameters $\delta > 0$ and $\gamma > 0$.

We design the method for the static Nash game in a general format with some fixed admissible initial values ($x_0 \in X$, $\hat{x}_0 \in \hat{X}$, and $\mu_0, \xi_0, \eta_0 \geq 0$), considering that we want to maximize the function, as follows:

1. The *first half-step*:

$$\left. \begin{aligned} \bar{\mu}_n &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(x_n, \hat{x}_n, \mu, \xi_n, \eta_n) \right\}, \\ \bar{\xi}_n &= \arg \max_{\xi \geq 0} \left\{ -\frac{1}{2} \|\xi - \xi_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(x_n, \hat{x}_n, \bar{\mu}_n, \xi, \eta_n) \right\}, \\ \bar{\eta}_n &= \arg \max_{\eta \geq 0} \left\{ -\frac{1}{2} \|\eta - \eta_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(x_n, \hat{x}_n, \bar{\mu}_n, \bar{\xi}_n, \eta) \right\}, \\ \bar{x}_n &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x, \hat{x}_n, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}, \\ \bar{\hat{x}}_n &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\frac{1}{2} \|\hat{x} - \hat{x}_n\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x_n, \hat{x}, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}. \end{aligned} \right\} \quad (9.7.4)$$

2. The *second half-step*

$$\left. \begin{aligned} \mu_{n+1} &= \arg \max_{\mu \geq 0} \left\{ -\frac{1}{2} \|\mu - \mu_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(\bar{x}_n, \bar{\hat{x}}_n, \mu, \bar{\xi}_n, \bar{\eta}_n) \right\}, \\ \xi_{n+1} &= \arg \max_{\xi \geq 0} \left\{ -\frac{1}{2} \|\xi - \xi_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(\bar{x}_n, \bar{\hat{x}}_n, \bar{\mu}_n, \xi, \bar{\eta}_n) \right\}, \\ \eta_{n+1} &= \arg \max_{\eta \geq 0} \left\{ -\frac{1}{2} \|\eta - \eta_n\|^2 - \gamma \mathcal{L}_{\theta, \delta}(\bar{x}_n, \bar{\hat{x}}_n, \bar{\mu}_n, \bar{\xi}_n, \eta) \right\}, \\ x_{n+1} &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_{\theta, \delta}(x, \bar{\hat{x}}_n, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}, \\ \hat{x}_{n+1} &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\frac{1}{2} \|\hat{x} - \hat{x}_n\|^2 + \gamma \mathcal{L}_{\theta, \delta}(\bar{x}_n, \hat{x}, \bar{\mu}_n, \bar{\xi}_n, \bar{\eta}_n) \right\}. \end{aligned} \right\} \quad (9.7.5)$$

9.8 Numerical Illustration

Consider a two-person bargaining problem in a class of continuous-time controllable Markov chains. Let us denote the disagreement cost that depends on the strategies $c_{(i_l, k_l)}^l$ ($l = 1, 2$) for players 1 and 2 as $d^1(c^1, c^2)$ and $d^2(c^1, c^2)$ respectively, and the solution for the bargaining problem as the point (u^1, u^2) .

The process to solve the bargaining problem consists of two main steps, firstly to find the disagreement point we define it as the Nash equilibrium point of the problem [28]; while for the solution of the bargaining process we follow the models presented by Nash and Kalai-Smorodinsky.

Let the states $N_1 = N_2 = 6$, and the number of actions $M_1 = M_2 = 3$. The individual utility for each player are defined by

$$U_{(i,j|1)}^1 = \begin{bmatrix} 34 & 45 & 1 & 28 & 7 & 23 \\ 27 & 43 & 25 & 47 & 26 & 24 \\ 15 & 45 & 14 & 15 & 43 & 48 \\ 36 & 47 & 12 & 17 & 20 & 5 \\ 20 & 41 & 22 & 43 & 35 & 14 \\ 29 & 29 & 18 & 18 & 32 & 23 \end{bmatrix}, U_{(i,j|1)}^2 = \begin{bmatrix} 31 & 1 & 30 & 38 & 2 & 17 \\ 18 & 41 & 10 & 13 & 42 & 11 \\ 5 & 8 & 34 & 33 & 12 & 31 \\ 2 & 44 & 13 & 43 & 3 & 40 \\ 25 & 5 & 22 & 5 & 28 & 10 \\ 13 & 18 & 7 & 29 & 48 & 3 \end{bmatrix},$$

$$U_{(i,j|2)}^1 = \begin{bmatrix} 30 & 44 & 14 & 47 & 25 & 31 \\ 44 & 24 & 45 & 37 & 11 & 30 \\ 24 & 25 & 12 & 20 & 32 & 22 \\ 22 & 25 & 44 & 50 & 12 & 33 \\ 38 & 12 & 36 & 33 & 27 & 22 \\ 24 & 5 & 44 & 45 & 37 & 1 \end{bmatrix}, U_{(i,j|2)}^2 = \begin{bmatrix} 15 & 15 & 43 & 9 & 18 & 14 \\ 13 & 13 & 2 & 36 & 32 & 30 \\ 25 & 25 & 15 & 42 & 18 & 22 \\ 39 & 23 & 45 & 2 & 11 & 5 \\ 18 & 41 & 27 & 38 & 40 & 2 \\ 29 & 5 & 7 & 18 & 17 & 25 \end{bmatrix},$$

$$U_{(i,j|3)}^1 = \begin{bmatrix} 28 & 27 & 48 & 8 & 16 & 27 \\ 43 & 47 & 33 & 24 & 22 & 28 \\ 21 & 37 & 19 & 28 & 15 & 42 \\ 24 & 29 & 24 & 3 & 50 & 42 \\ 42 & 49 & 46 & 33 & 31 & 42 \\ 50 & 42 & 51 & 45 & 13 & 11 \end{bmatrix}, U_{(i,j|3)}^2 = \begin{bmatrix} 14 & 11 & 31 & 48 & 50 & 11 \\ 17 & 34 & 14 & 39 & 39 & 20 \\ 15 & 23 & 28 & 31 & 24 & 2 \\ 9 & 22 & 48 & 48 & 35 & 24 \\ 20 & 9 & 36 & 3 & 21 & 17 \\ 35 & 10 & 34 & 14 & 20 & 49 \end{bmatrix}.$$

The transition rate matrices for each player are defined as follows

$$q_{(i,j|1)}^1 = \begin{bmatrix} -0.5371 & 0.0444 & 0.2305 & 0.0946 & 0.0705 & 0.0970 \\ 0.0208 & -0.5381 & 0.0294 & 0.0665 & 0.0471 & 0.3743 \\ 0.1179 & 0.0965 & -0.6554 & 0.0939 & 0.1042 & 0.2429 \\ 0.1871 & 0.0965 & 0.1622 & -0.5826 & 0.0285 & 0.1083 \\ 0.0825 & 0.1871 & 0.0671 & 0.0431 & -0.4624 & 0.0827 \\ 0.0831 & 0.1685 & 0.1221 & 0.3425 & 0.0432 & -0.7593 \end{bmatrix},$$

$$q_{(i,j|2)}^1 = \begin{bmatrix} -1.6112 & 0.1333 & 0.6916 & 0.2839 & 0.2114 & 0.2911 \\ 0.0624 & -1.6142 & 0.0881 & 0.1996 & 0.1412 & 1.1228 \\ 0.3538 & 0.2894 & -1.9662 & 0.2817 & 0.3127 & 0.7287 \\ 0.5614 & 0.2894 & 0.4867 & -1.7477 & 0.0855 & 0.3248 \\ 0.2474 & 0.5614 & 0.2012 & 0.1292 & -1.3873 & 0.2482 \\ 0.2492 & 0.5055 & 0.3662 & 1.0275 & 0.1295 & -2.2780 \end{bmatrix},$$

$$q_{(i,j|3)}^1 = \begin{bmatrix} -0.5371 & 0.0444 & 0.2305 & 0.0946 & 0.0705 & 0.0970 \\ 0.0208 & -0.5381 & 0.0294 & 0.0665 & 0.0471 & 0.3743 \\ 0.1179 & 0.0965 & -0.6554 & 0.0939 & 0.1042 & 0.2429 \\ 0.1871 & 0.0965 & 0.1622 & -0.5826 & 0.0285 & 0.1083 \\ 0.0825 & 0.1871 & 0.0671 & 0.0431 & -0.4624 & 0.0827 \\ 0.0831 & 0.1685 & 0.1221 & 0.3425 & 0.0432 & -0.7593 \end{bmatrix},$$

$$q_{(i,j|1)}^2 = \begin{bmatrix} -0.8499 & 0.2201 & 0.3707 & 0.1271 & 0.0374 & 0.0947 \\ 0.3467 & -0.6729 & 0.1271 & 0.0376 & 0.0970 & 0.0644 \\ 0.2831 & 0.0856 & -0.6306 & 0.0706 & 0.0376 & 0.1537 \\ 0.0703 & 0.1577 & 0.1369 & -0.8573 & 0.3673 & 0.1250 \\ 0.3727 & 0.0964 & 0.0944 & 0.1298 & -0.8026 & 0.1092 \\ 0.1627 & 0.1095 & 0.1237 & 0.0754 & 0.4537 & -0.9250 \end{bmatrix},$$

$$q_{(i,j|2)}^2 = \begin{bmatrix} -0.8499 & 0.2201 & 0.3707 & 0.1271 & 0.0374 & 0.0947 \\ 0.3467 & -0.6729 & 0.1271 & 0.0376 & 0.0970 & 0.0644 \\ 0.2831 & 0.0856 & -0.6306 & 0.0706 & 0.0376 & 0.1537 \\ 0.0703 & 0.1577 & 0.1369 & -0.8573 & 0.3673 & 0.1250 \\ 0.3727 & 0.0964 & 0.0944 & 0.1298 & -0.8026 & 0.1092 \\ 0.1627 & 0.1095 & 0.1237 & 0.0754 & 0.4537 & -0.9250 \end{bmatrix},$$

$$q_{(i,j|3)}^2 = \begin{bmatrix} -1.1332 & 0.2934 & 0.4942 & 0.1694 & 0.0498 & 0.1263 \\ 0.4623 & -0.8972 & 0.1694 & 0.0502 & 0.1294 & 0.0859 \\ 0.3774 & 0.1141 & -0.8408 & 0.0942 & 0.0501 & 0.2050 \\ 0.0938 & 0.2102 & 0.1825 & -1.1431 & 0.4898 & 0.1667 \\ 0.4970 & 0.1286 & 0.1258 & 0.1730 & -1.0701 & 0.1456 \\ 0.2169 & 0.1460 & 0.1650 & 0.1005 & 0.6049 & -1.2334 \end{bmatrix}.$$

9.8.1 Computing the Disagreement Point

Given δ and γ and applying the extraproximal method we obtain the convergence of the strategies for the disagreement point in terms of the variable $c_{(i_1, k_1)}^1$ for the player 1 (see Fig. 9.4) and the convergence of the strategies $c_{(i_2, k_2)}^2$ for the player 2 (see Fig. 9.5).

$$c^1 = \begin{bmatrix} 0.0517 & 0.0540 & 0.0523 \\ 0.0560 & 0.0605 & 0.0607 \\ 0.0542 & 0.0548 & 0.0514 \\ 0.0660 & 0.0672 & 0.0635 \\ 0.0332 & 0.0372 & 0.0385 \\ 0.0582 & 0.0679 & 0.0727 \end{bmatrix}, c^2 = \begin{bmatrix} 0.0824 & 0.0766 & 0.0830 \\ 0.0669 & 0.0449 & 0.0584 \\ 0.0840 & 0.0736 & 0.0823 \\ 0.0407 & 0.0215 & 0.0325 \\ 0.0399 & 0.0564 & 0.0503 \\ 0.0371 & 0.0329 & 0.0366 \end{bmatrix}.$$

Following (9.3.6) the mixed strategies obtained for the players are as follows

$$d^1 = \begin{bmatrix} 0.3273 & 0.3416 & 0.3311 \\ 0.3160 & 0.3416 & 0.3424 \\ 0.3378 & 0.3416 & 0.3205 \\ 0.3354 & 0.3416 & 0.3230 \\ 0.3051 & 0.3416 & 0.3533 \\ 0.2926 & 0.3416 & 0.3658 \end{bmatrix}, d^2 = \begin{bmatrix} 0.3405 & 0.3166 & 0.3429 \\ 0.3933 & 0.2637 & 0.3429 \\ 0.3503 & 0.3068 & 0.3429 \\ 0.4295 & 0.2275 & 0.3429 \\ 0.2723 & 0.3847 & 0.3429 \\ 0.3484 & 0.3087 & 0.3429 \end{bmatrix}.$$

Fig. 9.4 Convergence of the strategies for player 1 in the disagreement point

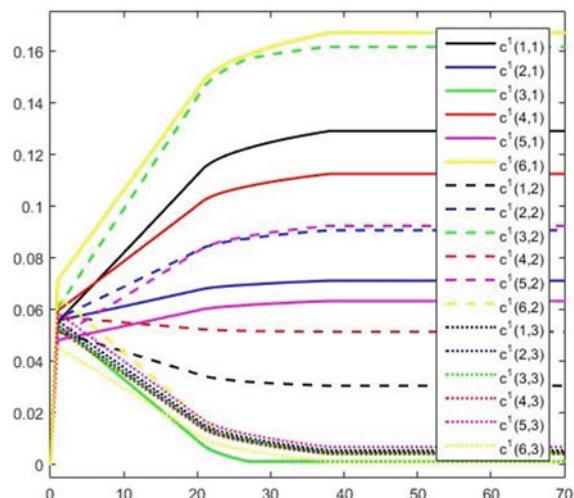
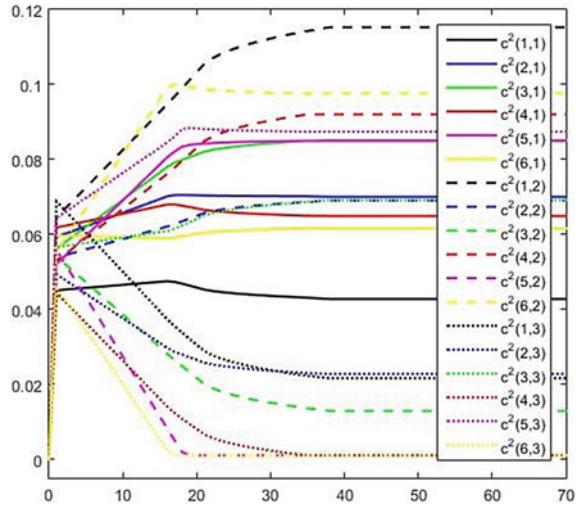


Fig. 9.5 Convergence of the strategies for player 2 in the disagreement point



With the strategies calculated, the resulting utilities following (9.3.3), in the disagreement point for each player $d^l(c^1, c^2)$, are as follows:

$$d^1(c^1, c^2) = 905.6447, d^2(c^1, c^2) = 704.2493.$$

9.8.2 Computing the Nash Bargaining Solution

Given δ , γ , α^l and applying the extraproximal method for the Nash bargaining solution, we obtain the convergence of the strategies in terms of the variable $c_{(i_1, k_1)}^1$ for the player 1 (see Fig. 9.6) and the convergence of the strategies $c_{(i_2, k_2)}^2$ for the player 2 (see Fig. 9.7).

$$c^1 = \begin{bmatrix} 0.0281 & 0.0677 & 0.0623 \\ 0.0010 & 0.0758 & 0.1003 \\ 0.0907 & 0.0686 & 0.0010 \\ 0.1115 & 0.0842 & 0.0010 \\ 0.0010 & 0.0466 & 0.0613 \\ 0.0010 & 0.0851 & 0.1127 \end{bmatrix}, c^2 = \begin{bmatrix} 0.1227 & 0.0350 & 0.0842 \\ 0.1100 & 0.0010 & 0.0592 \\ 0.1555 & 0.0010 & 0.0835 \\ 0.0607 & 0.0010 & 0.0329 \\ 0.0010 & 0.0946 & 0.0510 \\ 0.0663 & 0.0032 & 0.0371 \end{bmatrix}.$$

Fig. 9.6 Convergence of the strategies for player 1 in the Nash solution

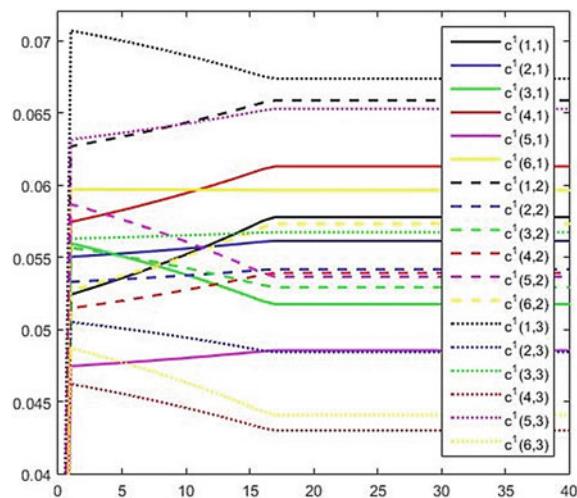
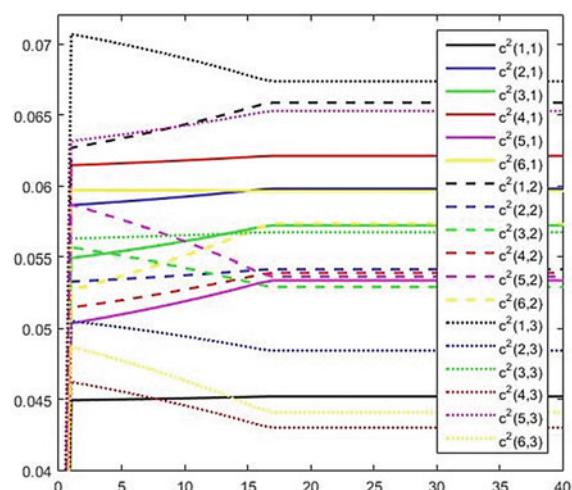


Fig. 9.7 Convergence of the strategies for player 2 in the Nash solution



The mixed strategies obtained for the players are as follows

$$d^1 = \begin{bmatrix} 0.1778 & 0.4280 & 0.3942 \\ 0.0056 & 0.4280 & 0.5663 \\ 0.5658 & 0.4280 & 0.0062 \\ 0.5669 & 0.4280 & 0.0051 \\ 0.0092 & 0.4280 & 0.5628 \\ 0.0050 & 0.4280 & 0.5670 \end{bmatrix}, d^2 = \begin{bmatrix} 0.5073 & 0.1447 & 0.3479 \\ 0.6462 & 0.0059 & 0.3479 \\ 0.6479 & 0.0042 & 0.3479 \\ 0.6415 & 0.0106 & 0.3479 \\ 0.0068 & 0.6453 & 0.3479 \\ 0.6221 & 0.0300 & 0.3479 \end{bmatrix}.$$

With the strategies calculated, the resulting utilities in the Nash bargaining solution for each player, are as follows:

$$u^1(c^1, c^2) = 958.0281, u^2(c^1, c^2) = 813.2879.$$

9.8.3 Computing the Kalai-Smorodinsky Bargaining Solution

Given δ, γ, α^l and applying the extraproximal method for the Kalai-Smorodinsky bargaining solution with the L_1 -norm, we obtain the convergence of the strategies in terms of the variable $c_{(i_1, k_1)}^1$ for the player 1 (see Fig. 9.8) and the convergence of the strategies $c_{(i_2, k_2)}^2$ for the player 2 (see Fig. 9.9).

$$c^1 = \begin{bmatrix} 0.0010 & 0.0432 & 0.1139 \\ 0.0010 & 0.0484 & 0.1278 \\ 0.1156 & 0.0438 & 0.0010 \\ 0.1420 & 0.0537 & 0.0010 \\ 0.0010 & 0.0297 & 0.0782 \\ 0.0010 & 0.0543 & 0.1435 \end{bmatrix} \cdot c^2 = \begin{bmatrix} 0.2061 & 0.0010 & 0.0349 \\ 0.1447 & 0.0010 & 0.0245 \\ 0.2044 & 0.0010 & 0.0346 \\ 0.0800 & 0.0010 & 0.0136 \\ 0.0010 & 0.1245 & 0.0211 \\ 0.0903 & 0.0010 & 0.0154 \end{bmatrix}.$$

Fig. 9.8 Convergence of the strategies for player 1 in the KS solution

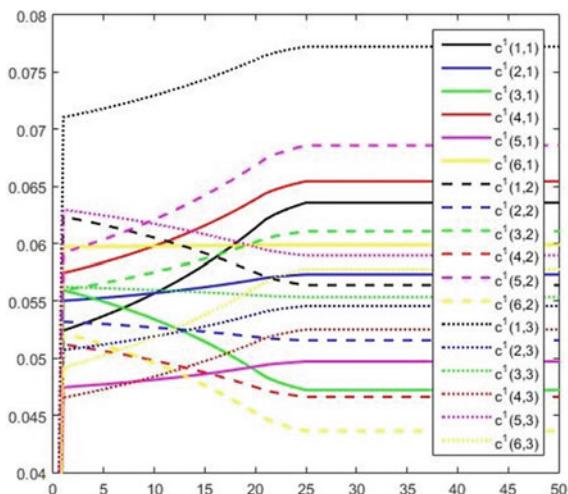
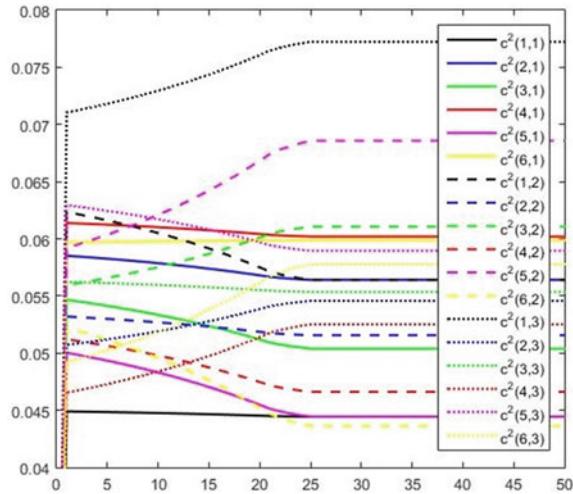


Fig. 9.9 Convergence of the strategies for payer 2 in the KS solution



The mixed strategies obtained for the players are as follows

$$d^1 = \begin{bmatrix} 0.0063 & 0.2730 & 0.7207 \\ 0.0056 & 0.2730 & 0.7213 \\ 0.7207 & 0.2730 & 0.0062 \\ 0.7219 & 0.2730 & 0.0051 \\ 0.0092 & 0.2730 & 0.7178 \\ 0.0050 & 0.2730 & 0.7219 \end{bmatrix}, d^2 = \begin{bmatrix} 0.8518 & 0.0041 & 0.1441 \\ 0.8500 & 0.0059 & 0.1441 \\ 0.8517 & 0.0042 & 0.1441 \\ 0.8454 & 0.0106 & 0.1441 \\ 0.0068 & 0.8491 & 0.1441 \\ 0.8465 & 0.0094 & 0.1441 \end{bmatrix}.$$

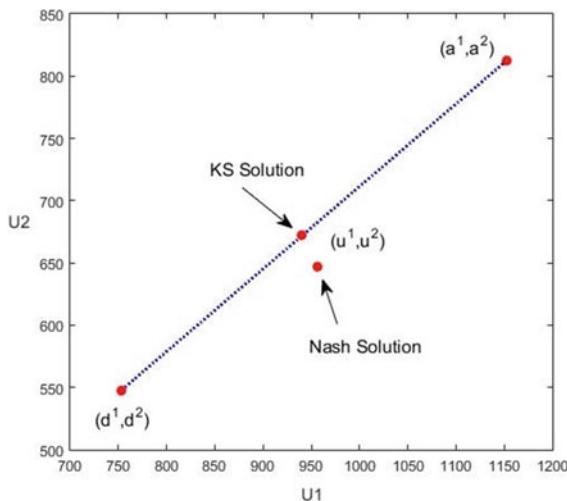
With the strategies calculated, the resulting utilities in the Kalai-Smorodinsky bargaining solution for each player are as follows:

$$u^1(c^1, c^2) = 960.5554, u^2(c^1, c^2) = 841.0831.$$

Figure 9.10 shows the straight line linking the utilities obtained at the disagreement point and those obtained at the utopia point. We can also observe that the Nash solution approaches this line while the Kalai-Smorodinsky solution is exactly on this line. The utilities on the utopia point for the bargaining problem are for each player as follows:

$$a^1(c^1, c^2) = 964.3472, a^2(c^1, c^2) = 849.8365.$$

Fig. 9.10 The bargaining Solution



References

1. Abreu, D., Manea, M.: Markov equilibria in a model of bargaining in networks. *Games Econ. Behav.* **75**(1), 1–16 (2012)
2. Agastya, M.: Adaptive play in multiplayer bargaining situations. *Games Econ. Behav.* **64**(3), 411–426 (1997)
3. Alexander, C.: The kalai-smorodinsky bargaining solution in wage negotiations. *J. Oper. Res. Soc.* **43**(8), 779–786 (1992)
4. Anant, T.C.A., Mukherji, B., Basu, K.: Bargaining without convexity: Generalizing the kalai-smorodinsky solution. *Econ. Lett.* **33**(2), 115–119 (1990)
5. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. *Comput. Math. Math. Phys.* **45**(11), 1893–1914 (2005)
6. Bolt, W., Houba, H.: Strategic bargaining in the variable threat game. *Econ. Theory* **11**(1), 57–77 (1998)
7. Cai, H.: Inefficient markov perfect equilibria in multilateral bargaining. *Econ. Theory* **22**(3), 583–606 (2003)
8. Carrillo, L., Escobar, J., Clempner, J.B., Poznyak, A.S.: Solving optimization problems in chemical reactions using continuous-time markov chains. *J. Math. Chem.* **54**, 1233–1254 (2016)
9. Clempner, J.B.: Shaping emotions in negotiation: a nash bargaining solution. *Cognit. Comput.* **12**, 720–735 (2020)
10. Clempner, J.B.: Manipulation power in bargaining games using machiavellianism. *Econ. Comput. Econ. Cybern. Stud. Res.* **55**(2), 299–313 (2021)
11. Clempner, J.B., Poznyak, A.S.: Simple computing of the customer lifetime value: a fixed local-optimal policy approach. *J. Syst. Sci. Syst. Eng.* **23**(4), 439–459 (2014)
12. Clempner, J.B., Poznyak, A.S.: Multiobjective markov chains optimization problem with strong pareto frontier: principles of decision making. *Expert Syst. Appl.* **68**, 123–135 (2017)
13. Coles, M.G., Muthoo, A.: Bargaining in a non-stationary environment. *J. Econ. Theory* **109**(1), 70–89 (2003)
14. Cripps, M.W.: Markov bargaining games. *J. Econ. Dyn. Control* **22**(3), 341–355 (1998)
15. Driesen, B., Pereira, A., Peters, H.: The kalai-smorodinsky bargaining solution with loss aversion. *Math. Soc. Sci.* **61**(1), 58–64 (2011)
16. Dubra, J.: An asymmetric kalai-smorodinsky solution. *Econ. Lett.* **73**(2), 131–136 (2001)

17. Forgó, F., Szép, J., Szidarovszky, F.: Introduction to the Theory of Games: Concepts, Methods, Applications. Kluwer Academic Publishers (1999)
18. Kalai, E.: Social Goals and Social Organization, chap. Solutions to the Bargaining Problem, pp. 75–105. Cambridge University Press, Cambridge (1985)
19. Kalai, E., Smorodinsky, M.: Other solutions to nash's bargaining problem. *Econometrica* **43**(3), 513–518 (1975)
20. Kalandrakis, A.: A three-player dynamic majoritarian bargaining game. *J. Econ. Theory* **116**(2), 294–322 (2004)
21. Kennan, J.: Repeated bargaining with persistent private information. *Rev. Econ. Stud.* **68**, 719–755 (2001)
22. Köpperling, V., Peters, H.: The effect of decision weights in bargaining problems. *J. Econ. Theory* **110**(1), 154–175 (2003)
23. Merlo, A., Wilson, C.: A stochastic model of sequential bargaining with complete information. *Econometrica* **63**(2), 371–399 (1995)
24. Moulin, H.: Implementing the kalai-smorodinsky bargaining solution. *J. Econ. Theory* **33**(1), 32–45 (1984)
25. Muthoo, A.: Bargaining Theory with Applications. Cambridge University Press (2002)
26. Naidu, S., Hwang, S., Bowles, S.: Evolutiogame bargaining with intentional idiosyncratic play. *Econ. Lett.* **109**(1), 31–33 (2010)
27. Nash, J.F.: The bargaining problem. *Econometrica* **18**(2), 155–162 (1950)
28. Nash, J.F.: Two person cooperative games. *Econometrica* **21**, 128–140 (1953)
29. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944)
30. Osborne, M., Rubinstein, A.: Bargaining and Markets. Academic Press, Inc. (1990)
31. Peters, H., Tijs, S.: Individually monotonic bargaining solutions for n-person bargaining games. *Methods Oper. Res.* **51**, 377–384 (1984)
32. Poznyak, A.S.: Advance Mathematical Tools for Automatic Control Engineers. Stochastic Techniques, vol. 2. Elsevier, Amsterdam (2009)
33. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker, New York (2000)
34. Raiffa, H.: Arbitration schemes for generalized two-person games. *Ann. Math. Stud.* **28**, 361–387 (1953)
35. Roth, A.E.: An impossibility result concerning n-person bargaining games. *Int. J. Game Theory* **8**(3), 129–132 (1979)
36. Rubinstein, A., Wolinsky, A.: Equilibrium in a market with sequential bargaining. *Econometrica* **53**(5), 1133–1150 (1985)
37. Trejo, K., Clempner, J., Poznyak, A.: Computing the bargaining approach for equalizing the ratios of maximal gains in continuous-time markov chains games. *Comput. Econ.* **54**, 933–955 (2019)
38. Trejo, K., Clempner, J., Poznyak, A.: Computing the nash bargaining solution for multiple players in discrete-time markov chains games. *Cybern. Syst.* **51**(1), 1–26 (2020)
39. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the lp-strong nash equilibrium looking for cooperative stability in multiple agents markov games. In: 12th International Conference on Electrical Engineering, Computing Science and Automatic Control, pp. 309–314. Mexico City. Mexico (2015)
40. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the stackelberg/nash equilibria using the extraproximal method: convergence analysis and implementation details for markov chains games. *Int. J. Appl. Math. Comput. Sci.* **25**(2), 337–351 (2015)
41. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the strong l_p -nash equilibrium for markov chains games: convergence and uniqueness. *Appl. Math. Modell.* **41**, 399–418 (2017)
42. Trejo, K.K., Juarez, R., Clempner, J., Poznyak, A.S.: Non-cooperative bargaining with unsophisticated agents. *Comput. Econ.* **61**, 937–974 (2023)

Chapter 10

Multi-traffic Signal-Control Synchronization



Abstract In this chapter, we provide a brand-new paradigm for combining game theory and the extraproximal approach to represent the multi-traffic signal-control synchronization problem. The intersection's goal is to reduce queuing time, and finding the best signal timing strategy, or assigning a green period to each signal phase, is a challenge for signal controllers. The game's players are referred to as signal controllers. Finding a green period at each junction seeks to reduce signal and queuing delays. The issue poses two inherent limitations: (a) the number of arriving and departing vehicles varies for each street of the intersection; and (b) for every intersection, the period of time allowed for vehicles to reach the green light is equal to the duration of the red light. A leader-follower Stackelberg game model is determined by the first restriction: streets with higher traffic demand more green time. The most recent constraint created a simultaneous game-solution as a better evaluation of the actual circumstance. The study then demonstrates that the Nash equilibrium provides the solution in order to take use of the game's structure. To solve the problem computationally and determine the best signal timing distribution, we present the c -variable technique. The extraproximal approach is a two-stage iterative process. In the first phase, an approximate equilibrium point is calculated, and in the second step, the previous step is adjusted. The game is described in terms of linked nonlinear programming issues that use the Lagrange principle. To guarantee the convergence of the cost-functions to a Nash equilibrium point, the Tikhonov regularization approach is used. The extraproximal approach is also developed using Markov chains. The three way intersection example serves as a good demonstration of the method's use. Our contributions have significant effects on the applications in the real world.

10.1 Introduction

10.1.1 Brief Review

Because of the high expenses and the fact that the space that might be used for roads already has a set structure and traffic flow, building new roads is no longer a practical option (especially in older cities). They waste several hours stuck in traffic during rush hour. Road traffic congestion and travel time are predicted to worsen as a result of population growth. Because of this, efforts are concentrated on creating intelligent technologies to reduce traffic congestion. Finding a point of balance between the supply and demand for roads is a planning conundrum for contemporary cities with the greatest levels of congestion, as noted by [7, 37].

The most significant aspect affecting the effectiveness of the road network is the *traffic-signal control* configuration. In order to lessen traffic congestion and the amount of time spent trapped in traffic, the challenge consists of creating appropriate signal patterns by managing the timing of the green/red light cycles at a junction. Many factors make this a highly difficult problem to solve:

1. The amount of traffic varies constantly depending on when most people commute,
2. An intersection affects other roads' traffic flow (i.e., the convergence of several urban blocks),
3. Schools, frequently visited restaurants, and fast food joints, where cars clog the entrances, are traditional sources of traffic congestion,
4. Unusual congestion may come from an accident, construction, or extended holiday weekends, or inclement weather,
5. Other times of congestion may be brought on by different special events, such as sporting events, festivals, or other religious ceremonies.

10.1.2 Related Work on Traffic Control

Several researchers have actively used various strategies to address the signal control setting problem in order to reduce delays and increase the intersection capacity. Allsop [1, 2] looks at the connection between signal setup and traffic assignment and suggests two programs to solve the problem of traffic equilibrium: 1) SIGSET for figuring out how much traffic signal-controlled road junctions can handle, and 2) SIGCAP for evaluating how much traffic certain junctions can handle. A computer technique called *mixed-integer traffic optimization*, as described by Gartner et al. in their citation [26], is intended to concurrently optimize all of the traffic control variables for network offsets, splits, and cycle time. The signal setting is addressed as a *hybrid optimization* issue by Tan et al. [43]. Smith [39, 40] defines equilibrium and stability in terms of link-flow and provides requirements that ensure the existence, singularity, and stability of traffic equilibria. Moreover, Smith [41, 42] suggests a traffic control technique that distributes traffic across the network's spare

capacity while just requiring local data from vehicle detectors, such as traffic flows and wait durations. As a one-level Cournot game and a Stackelberg game, Chen and Ben-Akiva [11] provide a combined dynamic traffic assignment and dynamic traffic control. According to Fisk [19], the traffic agency and network users play a Stackelberg game to determine the worldwide optimum signal setting problem. Alvarez et. al. [3], and Moya and Poznyak [33] treat the signal setup as a noncooperative game issue and a Stackelberg-Nash game theory method based on the extrapoximal approach, respectively (but without intersection avoiding). A methodology is presented for the integration of dynamic traffic assignment with real-time control by Gartner et al. [24, 25, 27]. The framework is extended to the dynamic case, which involves the incorporation of cutting-edge intelligent transportation systems, combining the model for signal control and route choice in urban traffic networks. First, discuss the static case, which involves the interaction between travelers (demand) and transportation facilities. Using evolutionary algorithms, Lee and Machemehl [30] analyze the signal setup. In their study of the combined assignment-intersection control problem, Cascetta et al. [8] demonstrate that, in the case of locally optimized control systems (such as adaptive traffic-lights), it deviates from the more general equilibrium network design problem by putting forth a stochastic user equilibrium (SUE) model with asymmetric delay functions. Castillo-Gonzalez et al. [10] considered a mathematically rigorous study of the continuous-time, discrete state, multi-traffic signal control problem using a non-cooperative game theory approach. Aragon-Gomez and Clempner [5] introduced a reinforcement learning approach for solving the Traffic-Signal Control problem for multiple intersections using Continuous-Time Markov Games.

In addition, Cascetta et al. [9] propose models and algorithms for the optimization of signal settings in urban networks, putting forward both a global approach (optimization of intersection signal settings throughout the whole network) and a local one (optimization of signal settings intersection by intersection). Models and applications pertaining to the analysis of transportation networks are presented by Cascetta [7]. Qu et al. [37] create a traffic control algorithm that takes traffic assignment into account. Dafermos [17, 18], Fisk and Nguyen [20], Florian and Spiess [21], Gartner et. al. [23], Meneguzzi [31, 32], Cantarella et. al. [6] and D'Acierno et. al. [16] deal with the local optimization of signal settings problem as a fixed-point problem, where look for an equilibrium traffic flows congruent with costs and signal settings obtained in accordance to a local-optimal control policy. The local optimization of signal settings problem, which arises when signal control parameters of an urban road network are locally optimized and have to be consistent with equilibrium traffic flows, is also studied by Gallo and D'Aciernob [22]. They compare various solution algorithms put forth in the literature.

Sheffi and Powell [38], Heydecker and Khoo [29], Yang and Yagar [49], Heydecker [28], Wong and Yang [47], Wong [48], Chiou [12], Wey [46], Ziyou and Yifan [50] and Cascetta [9], among others, approach the global optimization of signal settings as a (non-linear) constrained optimization problem where signal settings act as decision variables. Focusing on multi-objective traffic signal control, Khamis and Gomaa (Khamis) compute a consistent traffic signal configuration at each junction

that maximizes many performance indices. In order to forecast the impacts of potential control measures in a short time frame, Placzek [34] presents a self-organizing traffic light system for an urban road network.

In this chapter, we discuss a modeling approach based on game theory and the extraproximal method [4] for the multi-traffic signal-control synchronization problem [10, 15, 33]. The paper's main goal is to resolve scenarios with various degrees of demand (traffic flows) and various numbers of signalized crossings. The issue presents two natural limitations:

1. the number of entering and exiting vehicles varies for each street of the intersection; and
2. there is a traffic conflict between the two movements in the intersection area, preventing them from moving at the same time and resulting in the opposite traffic signal, where one movement has a green light and the other has a red light.

On the one hand, restriction (a) establishes a leader-follower Stackelberg game model, where streets with higher traffic volumes need more time for the green light and are therefore leaders. Contrarily, restriction (b) forces a simultaneous solution to the game as a better account of actual situations, i.e., the time allotted for cars to reach the green light is equal to the time allotted to reach the red light at a junction. The Nash equilibrium then provides the Stackelberg game's solution, which can be used to exploit the suggested game's structure [13, 14].

In order to solve the computationally challenging signal timing distribution problem, we provide the *c*-variable technique. The *c*-variable method has the significant benefit of being easily deployed in practical circumstances. The capacity of the *c*-variable approach to identify abnormal circumstances from data in the simplex is another crucial component of its introduction. A straightforward test on the *c*-variable can be used to identify a no possible solution. We use the extraproximal approach to determine the game's Nash equilibrium. We use coupled nonlinear programming issues that employ the Lagrange principle to illustrate the original game concept. In order to guarantee the convergence of the cost-functions to a Nash equilibrium point, Tikhonov's regularization method is also used. The extraproximal approach is an iterated, two-step process. A preliminary location to the equilibrium point is calculated in the first stage, and the prior forecast is adjusted in the second step. The Projector Gradient Technique is used to find the minimal condition for each equation in this system, which is an optimization issue. The value of this study comes in how well a real-world problem for designing signal settings while taking into account traffic flows and signal-controllers for several junctions was resolved.

10.2 Preliminaries

A *Controllable Markov chain* is a 4-tuple $MC = \{S, A, \Upsilon, \Pi\}$ where S is a finite set of states, $S \subset \mathbb{N}$, endowed with discrete topology; A is the set of actions, which is a metric space. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \cup_{s \in S} A(s)$; $\Upsilon = \{(s, a) | s \in S, a \in A(s)\}$ is the set of admissible state-action pairs, which is a measurable subset of $S \times A$; $\Pi(k) = [\pi_{j|ik}]$ is a stationary transition controlled matrix, where

$$\pi_{j|ik} \equiv P(s(n+1) = s_{(j)} | s(n) = s_{(i)}, a(n) = a_{(k)})$$

represents the probability associated with the transition from state $s_{(i)}$ to state $s_{(j)}$ under an action $a_{(k)} \in A$ ($k = 1, \dots, M$) at time $n = 0, 1, \dots$.

The dynamic of the game for Markov chains is described as follows. The game consists of a set $\mathcal{N} = \{1, \dots, n\}$ of players (denoted by $l = \overline{1, n}$) and begins at the initial state $s^l(0)$ which (as well as the states further realized by the process) is assumed to be completely measurable. Each player l is allowed to randomize, with distribution $d_{k|i}^l(n)$, over the pure action choices $a_{(k)}^l \in A^l$, $i = \overline{1, N_l}$ and $k = \overline{1, M_l}$. The players make the strategy selection trying to realize a Nash-equilibrium. Below we will consider only stationary strategies $d_{k|i}^l(n) = d_{k|i}^l$. These choices induce the state distribution dynamics

$$P^l(s^l(n+1) = s_{(j_l)}) = \sum_{i_l=1}^{N_l} \left(\sum_{k_l=1}^{M_l} \pi_{j_l|i_l k_l}^l d_{k_l|i_l}^l \right) P^l(s^{(l)}(n) = s_{i_l}).$$

In the ergodic case (when all Markov chains are ergodic for any stationary strategy $d_{k|i}^l$ the distributions $P^l(s^l(n+1) = s_{j_l})$ exponentially quickly converge to their limits $P^l(s = s_i)$ satisfying

$$P^l(s^{(l)} = s_{j_l}) = \sum_{i_l=1}^{N_l} \left(\sum_{k_l=1}^{M_l} \pi_{j_l|i_l k_l}^l d_{k_l|i_l}^l \right) P^l(s^{(l)} = s_{i_l}). \quad (10.2.1)$$

For any player l , his *individual rationality* is the player's cost function J^l of any fixed policy d^l is defined over all possible combinations of states and actions, and indicates the expected value when taking action a^l in state s^l and following policy d^l thereafter. The J -values can be expressed by

$$\mathbf{J}^l(c^l) := \sum_{i_l, k_l} W_{i_l k_l}^l c_{i_l k_l}^l,$$

where

$$W_{i_l k_l}^l = \sum_{j_l} J_{(i_l, j_l, k_l)}^l \pi_{j_l | i_l k_l}^l,$$

and $c^l := \|c_{i_l k_l}^l\|_{i_l=1, \dots, n; k_l=1, \dots, M_l}$ is a matrix with elements

$$c_{i_l k_l}^l = d_{k_l | i_l}^l P^l(s^{(l)} = s_{i_l}), \quad (10.2.2)$$

satisfying

$$c^{(l)} \in C_{adm}^{(l)} = \begin{cases} c^{(l)} : \sum_{i_l, k_l} c_{i_l k_l}^l = 1, c_{i_l k_l}^l \geq 0, \\ \sum_{k_l} c_{j_l k_l}^l = \sum_{i_l, k_l} \pi_{j_l | i_l k_l}^l c_{i_l k_l}^l. \end{cases} \quad (10.2.3)$$

The function $J_{(i_l, j_l, k_l)}^l$ is a constant cost at state s_{i_l} when the action $a_{k_l}^l$ is applied and the transfer to the state s_{j_l} is realized.

Then, the cost function of each player, depending on the states and actions of all participants, are given by the values $W_{i_1, k_1; \dots; i_n, k_n}^l$, so that the “average cost function” \mathbf{J}^l for each player l in the stationary regime can be expressed as

$$\mathbf{J}^l(c^0, \dots, c^n) := \sum_{i_0, k_0} \dots \sum_{i_n, k_n} W_{i_1, k_1; \dots; i_n, k_n}^l \prod_{l=0}^n c_{i_l k_l}^l. \quad (10.2.4)$$

Notice that by (10.2.2) it follows that

$$P^l(s^{(l)} = s_{i_l}) = \sum_{k_l} c_{i_l k_l}^l, \quad d_{k_l | i_l}^l = \frac{c_{i_l k_l}^l}{\sum_{k_l} c_{i_l k_l}^l}. \quad (10.2.5)$$

In the ergodic case $\sum_{k_l} c_{i_l k_l}^l > 0$ for all $l = \overline{0, n}$. The *individual aim* of each participant is

$$\mathbf{J}^l(c^l) \rightarrow \min_{c^{(l)} \in C_{adm}^{(l)}}. \quad (10.2.6)$$

10.3 Nash Equilibrium

Let us introduce the following variables

$$u^l := \text{col } c^{(l)}, \quad U^l := C_{adm}^{(l)} \quad (l = \overline{1, n}). \quad (10.3.1)$$

Consider a *non-zero sum game* with n players with strategies $u^l \in U^l$ ($l = \overline{1, n}$). Denote by

$$u = (u^1, \dots, u^n) \in U := \bigotimes_{l=1}^n U^l, \quad (10.3.2)$$

the joint strategy of the players. They are trying to reach one of the Nash equilibria, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^l \in X^l$ and any $l = \overline{1, n}$ the system of inequalities (*the Nash condition*)

$$\left. \begin{aligned} g_l(u^l, x^{\hat{l}*}) &\leq 0 \text{ for any } u^l \in U^l \text{ and all } l = \overline{1, n}, \\ g_l(u^l, x^{\hat{l}}) &:= \varphi_l(x^l, x^{\hat{l}}) - \varphi_l(u^l, x^{\hat{l}}), \end{aligned} \right\} \quad (10.3.3)$$

where $x^{\hat{l}}$ is a strategy of the rest of the players adjoint to u^l , namely,

$$x^{\hat{l}} := (u^1, \dots, u^{l-1}, u^{l+1}, \dots, u^n) \in X^{\hat{l}} := \bigotimes_{m=1, m \neq l}^n U^m,$$

and

$$x^l := \arg \min_{u^l \in U^l} \varphi_l(u^l, x^{\hat{l}}),$$

is the best-reply of the l player. Here $\varphi_l(u^l, x^{\hat{l}})$ is the cost-function of the player l which plays the strategy $u^l \in U^l$ and the rest of the players the strategy $x^{\hat{l}} \in X^{\hat{l}}$.

Lemma 10.1 *The Nash equilibrium $\bar{x} \in X$ (10.3.3) can be equivalently expressed in the joint format [44]*

$$\left. \begin{aligned} \max_{x \in X} g(u, \bar{x}) &\leq 0, \\ g(u, x) &:= \sum_{l=1}^n [\varphi_l(x^l, x^{\hat{l}}) - \varphi_l(u^l, x^{\hat{l}})], \\ \varphi_l(x^l, x^{\hat{l}}) &:= \min_{z^l \in X^l} \varphi_l(z^l, x^{\hat{l}}), \\ u^l \in X^l, x \in X &:= \bigotimes_{l=1}^n X^l. \end{aligned} \right\} \quad (10.3.4)$$

Proof Summing (10.3.3) implies (10.3.4). And inverse, taking $u^m = x^m$ for all $m \neq l$ in (10.3.4), which is valid for any admissible u^l , we obtain (10.3.3). \square

Notice that the condition $g(u, \bar{x}) \leq 0$ (10.3.4) for all admissible U , is equivalent to

$$\max_{u \in U} \{g(u, \bar{x})\} \leq 0.$$

for any fixed $u \in U$.

The functions $\varphi_l(u^l, x^l)$ ($l = \overline{1, n}$) are assumed to be convex in all their arguments.

Definition 10.1 A strategy $x^* \in X$ is said to be a **Nash equilibrium** if

$$x^* \in \operatorname{Arg} \max_{u \in U, x \in X} \{g(u, x) | g(u, x) \leq 0\} = \{u \in U, x \in X) | g(u, x) = 0\}. \quad (10.3.5)$$

10.3.1 The Regularized Lagrange Principle Application

Applying the Lagrange principle (see, for example, [36]) for Definition 10.1, we may conclude that (10.3.5) can be rewritten as

$$\left. \begin{aligned} &x^* \in \operatorname{Arg} \min_{u \in U, x \in X} \max_{\lambda \geq 0} \mathcal{L}(u, x, \lambda), \\ &\mathcal{L}(u, x, \lambda) := -g(u, x) + \lambda g(u, x) = (\lambda - 1)g(u, x). \end{aligned} \right\} \quad (10.3.6)$$

The approximative solution obtained by the Tikhonov regularization (see [36]) is given by

$$\left. \begin{aligned} &x_\delta^* = \operatorname{arg} \min_{u \in U, x \in X} \max_{\lambda \geq 0} \mathcal{L}_\delta(u, x, \lambda), \\ &\mathcal{L}_\circ(u, x, \lambda) := (\lambda - 1)g_\delta(u, x) - \frac{\delta}{2}\lambda^2, \end{aligned} \right\} \quad (10.3.7)$$

where $\delta > 0$ and

$$g_\delta(u, x) = g(u, x) - \frac{\delta}{2}(\|u\|^2 + \|x\|^2). \quad (10.3.8)$$

Notice that with $\delta > 0$ the considered functions becomes to be strictly convex providing the uniqueness of the considered conditional optimization problem (10.3.7). For $\delta = 0$ we have (10.3.6). Notice also that the Lagrange function in (10.3.7) satisfies the saddle-point [35] condition, namely, for all $x \in X$ and $\lambda \geq 0$ we have

$$\mathcal{L}_\circ(u_\delta^*, x_\delta, \lambda_\delta) \leq \mathcal{L}_\circ(u_\delta^*, x_\delta^*, \lambda_\delta^*) \leq \mathcal{L}_\circ(u_\delta, x_\delta^*, \lambda_\delta^*). \quad (10.3.9)$$

10.3.2 The Proximal Format

In the *proximal format* (see, [4]) the relation (10.3.7) can be expressed as

$$\left. \begin{aligned} \lambda_\delta^* &= \arg \max_{\lambda \geq 0} \left\{ -\frac{1}{2} \|\lambda - \lambda_\delta^*\|^2 + \gamma \mathcal{L}_o(u_\delta^*, x_\delta^*, \lambda) \right\}, \\ u_\delta^* &= \arg \min_{u \in U} \left\{ \frac{1}{2} \|u - u_\delta^*\|^2 + \gamma \mathcal{L}_o(u, x_\delta^*, \lambda_\delta^*) \right\}, \\ x_\delta^* &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_\delta^*\|^2 + \gamma \mathcal{L}_o(u_\delta^*, x, \lambda_\delta^*) \right\}, \end{aligned} \right\} \quad (10.3.10)$$

where the solutions u_δ^*, x_δ^* , and λ_δ^* depend on the small parameters $\delta, \gamma > 0$.

10.3.3 The Extraproximal Method

The *Extraproximal Method* for the conditional optimization problems (10.3.7) was suggested in [4]. We design the method for the static Nash game in a general format following [45].

The general format iterative version ($n = 0, 1, \dots$) of the extraproximal method with some fixed admissible initial values ($x_0 \in X, v_0 \in V$ and $\lambda_0 \geq 0$) is as follows

1. The *first half-step* (prediction):

$$\left. \begin{aligned} \bar{\lambda}_n &= \arg \min_{\lambda \geq 0} \left\{ \frac{1}{2} \|\lambda - \lambda_n\|^2 - \gamma \mathcal{L}_\delta(u_n, x_n, \lambda) \right\}, \\ \bar{u}_n &= \arg \min_{u \in U} \left\{ \frac{1}{2} \|u - u_n\|^2 + \gamma \mathcal{L}_o(u, x_n, \bar{\lambda}_n) \right\}, \\ \bar{x}_n &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_\delta(u_n, x, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (10.3.11)$$

2. The *second (basic) half-step*

$$\left. \begin{aligned} \lambda_{n+1} &= \arg \min_{\lambda \geq 0} \left\{ \frac{1}{2} \|\lambda - \lambda_n\|^2 - \gamma \mathcal{L}_\delta(\bar{u}_n, \bar{x}_n, \lambda) \right\}, \\ u_{n+1} &= \arg \min_{u \in U} \left\{ \frac{1}{2} \|u - u_n\|^2 + \gamma \mathcal{L}_o(u, \bar{x}_n, \bar{\lambda}_n) \right\}, \\ x_{n+1} &= \arg \min_{x \in X} \left\{ \frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{L}_\delta(\bar{u}_n, x, \bar{\lambda}_n) \right\}. \end{aligned} \right\} \quad (10.3.12)$$

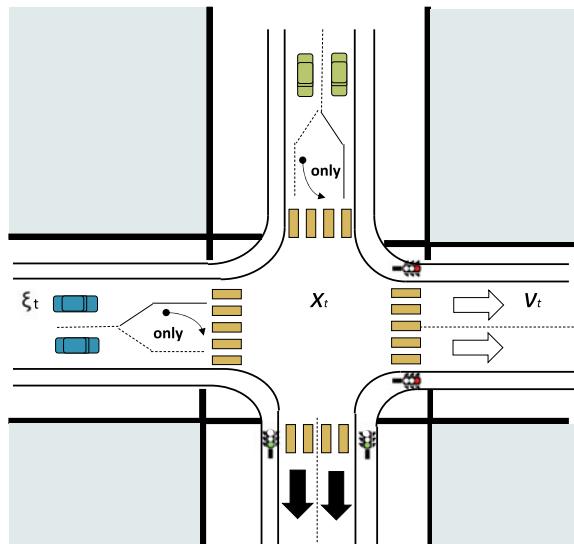
10.4 Traffic-Signal-Control Problem Formulation

The simplest game considers a two one-way-street intersection (Fig. 10.1). Let us suppose that each signal controller phase has been established. There is a traffic problem in the intersection area between the two movements, therefore they cannot travel at the same time and the traffic signal will be opposite, i.e. when one movement has a green light others movements have a red light. The objective of the intersection is to minimize the queuing delay and the problem for a signal controller is to find an optimal signal timing strategy, i.e. establishing green time to each signal phase. The dynamics of the game is as follows. Intersections are considered the players in the game. Each intersection aims at finding green time that minimizes its signal and queuing delay. Each player chooses a strategy to achieve the system optimum. So the conflict appears when each player wants to minimize its queue. The extended probability vector will be used in the cost function with the constraints on the behavior of the vehicles.

The vehicle flow is controlled by a signal controller of two color lights (actions) a_1 and a_2 representing the red and the green lights respectively. Let x and y the two controllers corresponding to the players l . At time n the total number of cars in the street is defined by $x(n)$ ($y(n)$), the number of entering cars is determined by $\xi^x(n)$ ($\xi^y(n)$), and the number of exiting cars is $v^x(n)$ ($v^y(n)$). The maximum capacity of the street (the queue) is determined by x^+ (y^+) meaning (l^+) . The dynamics of the flow of vehicles in a simple intersection of streets for the players x and y is defined as follows:

$$a(n) = a_1 : \text{red}(x)/\text{green}(y)$$

Fig. 10.1 Traffic signal control



$$x(n+1) = \begin{cases} x^+ & x(n) + \xi^x(n) > x^+ \\ x(n) + \xi^x(n) & x(n) + \xi^x(n) \leq x^+, \end{cases}$$

$$y(n+1) = \begin{cases} y^+ & y(n) + \xi^y(n) - v^y(n) > y^+ \\ [y(n) + \xi^y(n) - v^y(n)]_+ & y(n) + \xi^y(n) - v^y(n) \leq y^+, \end{cases}$$

where

$$[Z]_+ = \begin{cases} Z & Z \geq 0 \\ 0 & Z < 0, \end{cases}$$

$$a(n) = a_2 : green(x)/red(y)$$

$$x(n+1) = \begin{cases} x^+ & x(n) + \xi^x(n) - v^x(n) > x^+ \\ [x(n) + \xi^x(n) - v^x(n)]_+ & x(n) + \xi^x(n) - v^x(n) \leq x^+, \end{cases}$$

$$y(n+1) = \begin{cases} y^+ & y(n) + \xi^y(n) > y^+ \\ y(n) + \xi^y(n) & y(n) + \xi^y(n) \leq y^+. \end{cases}$$

These expressions describes the principle of a queue processing technique called FIFO, where the cars leave the queue in the order they arrive, or waiting one's turn at a traffic control signal a . We assume a Poisson distribution given by

$$f(l, \lambda t) = P(X = l) = \frac{e^{-(\lambda t)} (\lambda t)^l}{l!}$$

(10.4.1)

for a random variable X where l is the actual number of successes of an event and λ is the mean number of successes of an event. We will denote the input parameter by λ and the output parameter by μ . Note that the input parameter $\lambda^x > 0$ ($\lambda^y > 0$) and output parameter $\mu^x > 0$ ($\mu^y > 0$) are different for each player.

10.4.1 Transition Matrix

The following Theorems define a transition matrix $\pi_{j|ia}$ for each action a in the controlled Markov chain.

Red light (transition matrix $\pi_{j|ia_1}$)

Theorem 10.1 *Let $f(l, \lambda)$ be a Poisson distribution (10.4.1) with input parameter $\lambda^x > 0$ and the dynamics of the flow of the vehicles for the player x given by equation (10.4) with x^+ the maximum capacity of the street. Then, the transition matrix $\pi_{j|ik}^x$ to go from state i to state j using the action $a(n) = a_1 : red(x)$ is given by*

$$\boxed{\pi_{j|ia_1}^x = \delta_{j,x^+} \left(1 - e^{-\lambda^x} \sum_{l=0}^{x^+-i} \frac{(\lambda^x)^l}{l!} \right) + e^{-2\lambda^x} \frac{(\lambda^x)^{(j-i)}}{(j-1)!} \sum_{l=0}^{x^+-i} \frac{(\lambda^x)^l}{l!} \chi(j \geq i),}$$

where δ_{j,x^+} is the delta Kronecker,

$$\delta_{j,x^+} = \begin{cases} 1 & j = x^+ \\ 0 & j \neq x^+, \end{cases}$$

$\chi(j \geq i)$ is a characteristic function given by

$$\chi(j \geq i) = \begin{cases} 1 & j \geq i \\ 0 & j < i, \end{cases}$$

and i and j are the indexes corresponding to the states x_i and x_j .

Proof The transition matrix $\pi_{j|ia_1}^x$ is given by

$$\pi_{j|ia_1}^x = P(x(n+1) = j | x(n) = i, a(n) = a_1).$$

Using Eq. (10.4) and considering that $P(B) = \sum_q P(B|A_q)P(A_q)$ (see Poznyak [35]) we have that

$$\begin{aligned} \pi_{j|ia_1}^x &= P(x(n+1) = j | x(n) = i, a(n) = a_1, x(n) \\ &\quad + \xi^x(n) > x^+) P(x(n) + \xi^x(n) > x^+) |_{x(n)=i} + \\ &P(x(n+1) = j | x(n) = i, a(n) = a_1, x(n) \\ &\quad + \xi^x(n) \leq x^+) P(x(n) + \xi^x(n) \leq x^+) |_{x(n)=i}, \end{aligned}$$

where $x(n+1) = j$ and $x(n) = i$ are the index corresponding to the states x_j and x_i . Then, we have that

$$P(x(n) + \xi^x(n) > x^+) |_{x(n)=i} = P(\xi^x(n) > x^+ - x(n)) |_{x(n)=i} = P(\xi^x(n) > x^+ - i) =$$

$$\sum_{l=x^+-i+1}^{\infty} \frac{e^{-(\lambda^x t)} (\lambda^x t)^l}{l!} \Big|_{t=1} = e^{-\lambda^x} \sum_{l=x^+-i+1}^{\infty} \frac{(\lambda^x)^l}{l!} = 1 - e^{-\lambda^x} \sum_{l=0}^{x^+-i} \frac{(\lambda^x)^l}{l!},$$

and considering that the maximum index j of x_j is x^+ we have that

$$P(x(n+1) = j | x(n) = i, a(n) = a_1, x(n) + \xi^x(n) > x^+) = \delta_{j,x^+}.$$

As well, we have that

$$P(x(n) + \xi^x(n) \leq x^+) |_{x(n)=i} = P(\xi^x(n) \leq x^+ - x_i) = e^{-\lambda^x} \sum_{l=0}^{x^+ - x_i} \frac{(\lambda^x)^l}{l!},$$

and

$$\begin{aligned} P(x(n+1) = j | x(n) = i, a(n) = a_1, x(n) + \xi^x(n) \leq x^+) \\ P(x(n+1) = j | x(n) = i, a(n) = a_1, \xi^x(n) \leq x^+ - i) \\ = P(\xi^x(n) \leq j - i) = e^{-\lambda^x} \frac{(\lambda^x)^{(j-i)}}{(j-i)!}. \end{aligned}$$

Then, we have that

$$\pi_{j|ia_1}^x = \delta_{j,x^+} \left(1 - e^{-\lambda^x} \sum_{l=0}^{x^+ - i} \frac{(\lambda^x)^l}{l!} \right) + e^{-2\lambda^x} \frac{(\lambda^x)^{(j-i)}}{(j-1)!} \sum_{l=0}^{x^+ - i} \frac{(\lambda^x)^l}{l!} \chi(j \geq i).$$

□

Corollary 10.1 Let $f(l, \lambda)$ be a Poisson distribution (10.4.1) with input parameter $\lambda^y > 0$ and the dynamics of the flow of the vehicles for the player y given by Eq. (10.4) with y^+ the maximum capacity of the street. Then, the transition matrix $\pi_{j|ik}^y$ to go from state i to state j using the action $a(n) = a_2 : red(y)$ is given by

$$\boxed{\pi_{j|ia_2}^y = \delta_{y_j, y^+} \left(1 - e^{-\lambda^y} \sum_{l=0}^{y^+ - i} \frac{(\lambda^y)^l}{l!} \right) + e^{-2\lambda^y} \frac{(\lambda^y)^{(j-i)}}{(j-1)!} \sum_{l=0}^{y^+ - i} \frac{(\lambda^y)^l}{l!} \chi(j \geq i),}$$

where δ_{y_j, y^+} is the delta Kronecker, $\chi(j \geq i)$ is a characteristic function, and i and j are the indexes corresponding to the states y_i and y_j .

Green light (transition matrix $\pi_{j|ia_2}$)

Theorem 10.2 Let $f(l, \lambda)$ be a Poisson distribution (10.4.1) with input parameter $\lambda^x > 0$, output parameter μ^x and the dynamics of the flow of vehicles for the player x given by Eq. (10.4) with x^+ the maximum capacity of the street. Then, the transition matrix $\pi_{j|ia_2}^x$ to go from state i to state j using the action $a(n) = a_2 : green(x)$ is given by

$$\boxed{\begin{aligned} \pi_{j|ia_2}^x = \delta_{j,x^+} \left(1 - e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \frac{(\mu^x)^m}{m!} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right) + \\ \left(\sum_{m=0}^{\infty} \left[e^{-2\lambda^x} \left(\sum_{l=[m-1]_+}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \sum_{q=[m-1]_+}^{\infty} \frac{(\lambda^x)^q}{q!} \right) + \delta_{j,0} \left(e^{-\lambda^x} \sum_{l=0}^{[m-i]_+ + 1} \frac{(\lambda^x)^l}{l!} \right) \right] \left(e^{-\mu^x} \frac{(\mu^x)^m}{m!} \right) \right) \times \\ \left(e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \frac{(\mu^x)^m}{m!} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right) \end{aligned}}$$

where δ_{j,x^+} and $\delta_{j,0}$ are the delta Kronecker and $\chi(j \geq i)$ is a characteristic function, and i and j are the indexes corresponding to the states x_i and x_j .

Proof The transition matrix $\pi_{j|ia_2}^x$ is given by

$$\pi_{j|ia_2}^x = P(x(n+1) = j|x(n) = i, a(n) = a_2).$$

Using Eq. (10.4) and considering that $P(B) = \sum_l P(B|A_l)P(A_l)$ (see Poznyak [35]) we have that

$$\begin{aligned}\pi_{j|ia_2}^x &= P(x(n+1) = j|x(n) = i, a(n) = a_2, x(n) + \xi^x(n) \\ &\quad - v^x(n) > x^+)P(x(n) + \xi^x(n) - v^x(n) > x^+)|_{x(n)=i} + \\ P(x(n+1) &= j|x(n) = i, a(n) = a_2, x(n) + \xi^x(n) - v^x(n) \\ &\leq x^+)P(x(n) + \xi^x(n) - v^x(n) \leq x^+)|_{x(n)=i},\end{aligned}$$

where $x(n+1) = j$ and $x(n) = i$ are the index corresponding to the states x_j and x_i .

$$\begin{aligned}P(x(n) + \xi^x(n) - v^x(n) > x^+)|_{x(n)=i} &= P(\xi^x(n) - v^x(n) > x^+ - i) = \\ &= \sum_{m=0}^{\infty} P(\xi^x(n) - v^x(n) > x^+ - i)|_{v(n)=m} P(v^x(n) = m) \\ &= \sum_{m=0}^{\infty} \left[\left(e^{-\lambda^x} \sum_{l=x^+-i+m+1}^{\infty} \frac{(\lambda^x)^l}{l!} \right) \left(e^{-\mu^x} \frac{(\mu^x)^m}{m!} \right) \right] \\ &= 1 - \sum_{m=0}^{\infty} \left[\left(e^{-\lambda^x} \sum_{l=0}^{x^+-i+m} \frac{(\lambda^x)^l}{l!} \right) \left(e^{-\mu^x} \frac{(\mu^x)^m}{m!} \right) \right] \\ &= 1 - e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \left[\left(\frac{(\mu^x)^m}{m!} \right) \left(\sum_{l=0}^{x^+-i+m} \frac{(\lambda^x)^l}{l!} \right) \right],\end{aligned}$$

and considering that the maximum index j of x_j is x^+ we have that

$$P(x(n+1) = j|x(n) = i, a(n) = a_2, x(n) + \xi^x(n) - v^x(n) > x^+) = \delta_{j,x^+}.$$

As well, we have that

$$\begin{aligned}
P(x(n) + \xi^x(n) - v^x(n) \leq x^+) |_{x(n)=i} &= P(\xi^x(n) - v^x(n) \leq x^+ - i) \\
&= \sum_{m=0}^{\infty} P(\xi^x(n) - v^x(n) \leq x^+ - i) |_{v(n)=m} P(v^x(n) = m) \\
&= \sum_{m=0}^{\infty} P(\xi^x(n) \leq x^+ - i + m) P(v^x(n) = m) \\
&= \sum_{m=0}^{\infty} \left[\left(e^{-\lambda^x} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right) \left(e^{-\mu^x} \frac{(\mu^x)^m}{m!} \right) \right] \\
&= e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \frac{(\mu^x)^m}{m!} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!},
\end{aligned}$$

and

$$\begin{aligned}
P(x(n+1) = j | x(n) = i, a(n) = a_2, x(n) + \xi^x(n) - v^x(n) \leq x^+) |_{x(n)=i} \\
&= P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) - v^x(n) \leq x^+ - i) \\
&= \sum_{m=0}^{\infty} P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \\
&\quad - v^x(n) \leq x^+ - i \wedge v^x(n) = m) \cdot \\
P(v^x(n) = m) &= \sum_{m=0}^{\infty} P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \\
&\leq x^+ - i + m) P(v^x(n) = m) \\
&\sum_{m=0}^{\infty} [P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \\
&\leq x^+ - i + m \wedge \xi^x(n) + i - m \geq 0) \cdot \\
P(\xi^x(n) + i - m \geq 0) &+ P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \leq x^+ - i + m \wedge \xi^x(n) + i - m < 0) \cdot \\
P(\xi^x(n) + i - m < 0)] P(v^x(n) = m).
\end{aligned}$$

Now

$$\begin{aligned}
P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \leq x^+ - i + m \wedge \xi^x(n) + i - m \geq 0) \\
&= e^{-\lambda^x} \sum_{l=[m-1]_+}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!},
\end{aligned}$$

$$\begin{aligned}
P(\xi^x(n) + i - m \geq 0) &= P(\xi^x(n) \geq [m - i]_+) \\
&= e^{-\lambda^x} \sum_{q=[m-1]_+}^{\infty} \frac{(\lambda^x)^q}{q!},
\end{aligned}$$

$$P(x(n+1) = j | x(n) = i, a(n) = a_2, \xi^x(n) \leq x^+ - i + m \wedge \xi^x(n) + i - m < 0) = \delta_{j,0},$$

$$\begin{aligned} P(\xi^x(n) + i - m < 0) &= P(\xi^x(n) < m - i) = P(\xi^x(n) < [m - i]_+ + 1) \\ &= e^{-\lambda^x} \sum_{l=0}^{[m-i]_++1} \frac{(\lambda^x)^l}{l!}. \end{aligned}$$

Then, we have that

$$\begin{aligned} \pi_{j|ia_2}^x &= \delta_{j,x^+} \left(1 - e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \frac{(\mu^x)^m}{m!} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right) + \\ &\quad \left(\sum_{m=0}^{\infty} \left[e^{-2\lambda^x} \left(\sum_{l=[m-1]_+}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right) \sum_{q=[m-1]_+}^{\infty} \frac{(\lambda^x)^q}{q!} \right] + \delta_{j,0} \left(e^{-\lambda^x} \sum_{l=0}^{[m-i]_++1} \frac{(\lambda^x)^l}{l!} \right) \right) \left(e^{-\mu^x} \frac{(\mu^x)^m}{m!} \right) + \\ &\quad \left(e^{-(\lambda^x + \mu^x)} \sum_{m=0}^{\infty} \frac{(\mu^x)^m}{m!} \sum_{l=0}^{x^+ - i + m} \frac{(\lambda^x)^l}{l!} \right). \end{aligned}$$

□

Corollary 10.2 Let $f(l, \lambda)$ be a Poisson distribution (10.4.1) with input parameter $\lambda^y > 0$, output parameter μ^y and the dynamics of the flow of vehicles for the player y given by Eq. (10.4) with y^+ the maximum capacity of the street. Then, the transition matrix $\pi_{j|ia_2}^y$ to go from state i to state j using the action corresponding to $a(n) = a_2 : \text{green}(y)$ is given by

$$\begin{aligned} \pi_{j|ia_2}^y &= \delta_{j,y^+} \left(1 - e^{-(\lambda^y + \mu^y)} \sum_{m=0}^{\infty} \frac{(\mu^y)^m}{m!} \sum_{l=0}^{y^+ - i + m} \frac{(\lambda^y)^l}{l!} \right) + \\ &\quad \left(\sum_{m=0}^{\infty} \left[e^{-2\lambda^y} \left(\sum_{l=[m-1]_+}^{y^+ - i + m} \frac{(\lambda^y)^l}{l!} \right) \sum_{q=[m-1]_+}^{\infty} \frac{(\lambda^y)^q}{q!} \right] + \delta_{j,0} \left(e^{-\lambda^y} \sum_{l=0}^{[m-i]_++1} \frac{(\lambda^y)^l}{l!} \right) \right) \left(e^{-\mu^y} \frac{(\mu^y)^m}{m!} \right) \times \\ &\quad \left(e^{-(\lambda^y + \mu^y)} \sum_{m=0}^{\infty} \frac{(\mu^y)^m}{m!} \sum_{l=0}^{y^+ - i + m} \frac{(\lambda^y)^l}{l!} \right), \end{aligned}$$

where δ_{j,y^+} and $\delta_{j,0}$ is the delta Kronecker and $\chi(j \geq i)$ is a characteristic function, and i and j are the indexes corresponding to the states y_i and y_j .

10.4.2 Ergodicity

The corresponding Markov chain has the coefficients of ergodicity k'_{erg} estimated (see Chapter 1) from below as

$$k'_{erg} := \min_{n_0^{(l)}} \max_{j^{(l)}=1, \dots, N} \min_{i^{(l)}=1, \dots, N} \tilde{\pi}_{j^{(l)}|i^{(l)}}^{(l)*} (n_0^{(l)} = 1) \geq$$

$$\max_{j^{(l)}=1, \dots, N} \min_{i^{(l)}=1, \dots, N} \min \left\{ \pi_{j^{(l)}|i^{(l)}a_1^{(l)}}^{(l)}; \pi_{j^{(l)}|i^{(l)}a_2^{(l)}}^{(l)} \right\} := \rho^l.$$

So that if ρ^l is positive for every player, then the Markov chains are ergodic.

10.4.3 Cost Function

The individual aim of player (x) (under stationary strategies) can be formulated as follows:

$$\mathbf{J}^{(x)} = \sum_{i^{(y)} j^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} \sum_{i^{(x)} j^{(x)}}^{x^+} \sum_{\substack{k^{(x)} \\ k^{(y)} \neq k^{(x)}}}^{M_x} j^{(x)} \left(\pi_{j^{(x)}|i^{(x)}k^{(x)}}^{(x)} d_{k^{(x)}|i^{(x)}}^{(x)} P_{i^{(x)}}^{(x)} \right) \left(\pi_{j^{(y)}|i^{(y)}k^{(y)}}^{(y)} d_{k^{(y)}|i^{(y)}}^{(y)} P_{i^{(y)}}^{(y)} \right).$$

Since $\sum_{i^{(y)} j^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} \pi_{j^{(y)}|i^{(y)}k^{(y)}}^{(y)} = 1$ we have that

$$\mathbf{J}^{(x)} = \sum_{i^{(x)} j^{(x)}}^{x^+} \sum_{k^{(x)}}^{M_x} j^{(x)} \left(\pi_{j^{(x)}|i^{(x)}k^{(x)}}^{(x)} d_{k^{(x)}|i^{(x)}}^{(x)} P_{i^{(x)}}^{(x)} \right) \sum_{i^{(y)}}^{y^+} \sum_{\substack{k^{(y)} \\ k^{(y)} \neq k^{(x)}}}^{M_y} d_{k^{(y)}|i^{(y)}}^{(y)} P_{i^{(y)}}^{(y)},$$

and we obtain

$$\mathbf{J}^{(x)} = \sum_{i^{(x)}}^{x^+} \sum_{k^{(x)}}^{M_x} \left(\sum_{j^{(x)}}^{x^+} j^{(x)} \pi_{j^{(x)}|i^{(x)}k^{(x)}}^{(x)} \right) d_{k^{(x)}|i^{(x)}}^{(x)} P_{i^{(x)}}^{(x)} \sum_{i^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} d_{k^{(y)}|i^{(y)}}^{(y)} P_{i^{(y)}}^{(y)},$$

having

$$W_{i^{(x)}k^{(x)}}^{(x)} = \sum_{j^{(x)}}^{x^+} j^{(x)} \pi_{j^{(x)}|i^{(x)}k^{(x)}}^{(x)},$$

where $(i)^+$ is the size of the buffer (the maximum admissible number of cars for each player l). In general, for each player l we have

$$\mathbf{J}^{(l)} = \sum_{i^{(l)}}^{(i)^+} \sum_{k^{(l)}}^{M_l} W_{i^{(l)} k^{(l)}}^{(l)} c_{i^{(l)} k^{(l)}}^{(l)} \sum_{i^{\hat{l}}}^{\hat{(i)}^+} \sum_{k^{\hat{l}}}^{M_{\hat{l}}} c_{i^{\hat{l}} k^{\hat{l}}}^{(\hat{l})},$$

where

$$c_{i^{(l)} k^{(l)}}^{(l)} = d_{k^{(l)} | i^{(l)}}^{(l)} P_{i^{(l)}}^{(l)}, \quad c_{i^{\hat{l}} k^{\hat{l}}}^{(\hat{l})} = d_{k^{\hat{l}} | i^{\hat{l}}}^{(\hat{l})} P_{i^{\hat{l}}}^{(\hat{l})}.$$

The signal traffic control is based on the green light duration determined by the estimated number of vehicles entering the road during a green/red light cycle. The distance of player l is calculated in terms of the size of the buffer $((i)^+)$ multiplied by the number of entering and exiting cars represented by the transition matrix $(\pi_{j^{(l)} | i^{(l)} k^{(l)}}^{(l)})$ and, the long-run of the complementary player $(c_{i^{\hat{l}} k^{\hat{l}}}^{(\hat{l})})$.

10.5 Gradient Solver

The loss-functions (10.2.4) of $\mathbf{J}^l(n)$ for each player l in the nonstationary regime can be expressed as

$$\begin{aligned} \mathbf{J}_{(c^0, \dots, c^n)}^l(n) &:= \sum_{i_0, k_0} \dots \sum_{i_n, k_n} W_{i_1, k_1, \dots, i_n, k_n}^l(n) \prod_{l=0}^n c_{i_l k_l}^l(n) = \\ &\sum_{i_0, k_0} \dots \sum_{i_n, k_n} \left(\sum_{j_0} \dots \sum_{j_n} J_{(i_1, j_1, k_1, \dots, i_n, j_n, k_n)}^l(n) \prod_{l=0}^n \pi_{j_l | i_l k_l}^l(n) \right) \left(\prod_{l=0}^n c_{i_l k_l}^l(n) \right). \end{aligned}$$

The *individual aim* is

$$\mathbf{J}_{(c^1, \dots, c^n)}^l(n) \rightarrow \min_{c_{i_l k_l}^l(n) \in C_{adm}^{(l)}}.$$

Note that the function $\mathbf{J}_{(c^0, \dots, c^n)}^l(n)$ is polylineal in $c_{i_l k_l}^l(n) \in C_{adm}^{(l)}$, and therefore can not be solved analytically. So, we need to apply an iterative method to find a minimizing solution which, additionally, may be not unique. For solving the ill-posed problem we introduce a Tikhonov's regularizer with regularization parameter $(\delta > 0)$ which consists on

$$\tilde{\mathbf{J}}_{(c^1, \dots, c^n)}^l(n) := \mathbf{J}_{(c^1, \dots, c^n)}^l(n) + \frac{\delta}{2} \sum_{i_l=1}^{N_l} \sum_{k_l=1}^{M_l} \|c_{i_l k_l}^l(n)\|^2 \rightarrow \min_{c_{i_l k_l}^l(n) \in C_{adm}^{(l)}},$$

$(\tilde{\mathbf{J}}_{(c^1, \dots, c^n)}^l(n))$ is as in Eqs. (10.3.11) and Eqs. (10.3.12)). Obviously, when $\delta \rightarrow 0$ we obtain the initial problem. In addition, we will ask $\tilde{\mathbf{J}}_{(c^1, \dots, c^n)}^l(n)$ to satisfy the restrictions described in Eq. (10.2.3). Then, defining

$$h^l(c_{i_l k_l}^l(n)) = \sum_{k_l} c_{j_l k_l}^l - \sum_{i_l, k_l} \pi_{j_l | i_l k_l}^l c_{i_l k_l}^l,$$

we finally propose the *Projection Gradient Method* [36] with regularization parameter ($\xi > 0$) which consists in solution of the following problem:

$$\tilde{\mathbf{V}}_{(c^1, \dots, c^n, \theta_1, \dots, \theta_n)}^l(n) = \tilde{\mathbf{J}}_{(c^1, \dots, c^n)}^l(n) + \sum_l \theta_l h^l(c_{i_l k_l}^l(n)) + \frac{\xi}{2} \sum_l \theta_l^2.$$

Algorithm 10.1: Projector gradient method

Step 1: Initialization

Choose tolerances $\varepsilon > 0$, starting point $c_{i_l k_l}^l(0)$, regularization parameter ξ_0 and the initial parameter μ_0^c .

Step 2: Iterations

for a given $i_l = \overline{1, N_l}$, $k_l = \overline{1, M_l}$ and a fixed n

2.(a) compute $c_{i_l k_l}^l(n+1)$ using the following equation:

$$\left. \begin{aligned} \tilde{c}_{i_l k_l}^l(n+1) &= c_{i_l k_l}^l(n) - \mu_n^c \nabla_{c_{i_l k_l}^l(n)} \tilde{\mathbf{V}}_{(c^0, \dots, c^n, \theta_1, \dots, \theta_n)}^l(n) \\ c_{i_l k_l}^l(n+1) &= \text{Pr}_{S^{N_l M_l}} \{ \tilde{c}_{i_l k_l}^l(n+1) \}, \end{aligned} \right\} \quad (10.5.1)$$

where $\text{Pr}_{S^{N_l M_l}} \{ \cdot \}$ is the projection operator of the vector from \mathbb{R}^{M_l} into the simplex $S^{N_l M_l}$ and $\tilde{\mathbf{J}}_{(c^0, \dots, c^n)}^l(n)$ is as in Eqs. (10.3.11) and Eqs. (10.3.12).

2.(b) compute $\theta_l(n+1)$ using the following equation:

$$\theta_l(n+1) = \theta_l(n) - \mu_n^\theta \nabla_{\theta_l(n)} \tilde{\mathbf{V}}_{(c^1, \dots, c^n, \theta_1, \dots, \theta_n)}^l(n)$$

2.(c) verify descent direction

$$\| c_{i_l k_l}^l(n) - \text{Pr}_{S^{N_l M_l}} \{ c_{i_l k_l}^l(n+1) \} \| \leq \| c_{i_l k_l}^l(n) - \tilde{c}_{i_l k_l}^l(n+1) \|$$

for any $c_{i_l k_l}^l(n+1) \in \mathbb{R}^{M_l}$ and any $c_{i_l k_l}^l(n) \in S^{N_l M_l}$.

end

Step 3: Stopping criteria

Check the convergence criteria $\| c_{i_l k_l}^l(n+1) - c_{i_l k_l}^l(n) \| < \varepsilon$ then stop. Otherwise, set $n = n + 1$ and return to Step 2.

The corresponding iterative algorithm for the Projector Gradient Method is described in the Algorithm 10.1. It is shown that we have the convergence of this method

$$c_{i_l k_l}^l(n) \xrightarrow{n \rightarrow \infty} c_{i_l k_l}^{l, \ast\ast}$$

to one of the solutions $c_{i_l k_l}^{l, \ast}$ of the initial problem (10.5) with minimal norm, i.e.,

$$c_{i_l k_l}^{l, \ast\ast}(n) := \min_{c_{i_l k_l}^l \in C_{adm}^{(l)}} \sum_{i_l^t=1}^{N_l} \sum_{k_l^t=1}^{M_l} \|c_{i_l k_l}^{l, \ast}(n)\|^2,$$

if the parameters $\{\mu_n^c\}$ and $\{\delta_n^c\}$ of the procedure (10.5.1) fulfill

$$0 < \mu_n \xrightarrow{n \rightarrow \infty} 0, \quad 0 < \delta_n \xrightarrow{n \rightarrow \infty} 0,$$

$$\sum_{n=0}^{\infty} \mu_n \delta_n = \infty, \quad \frac{|\delta_{n+1} - \delta_n|}{\mu_n \delta_n} \xrightarrow{n \rightarrow \infty} 0,$$

$$\frac{\mu_n}{\delta_n} \xrightarrow{n \rightarrow \infty} \zeta \text{ which is small enough.}$$

10.6 Application Example

We consider the three-way intersection game with three one-way-street intersection (see Fig. 10.2). Let us introduce the following notation for describing the problem

N^l = The total number of states of the street,

M^l = The total number of actions = intersections,

λ^l = The input parameter of player l ,

μ^l = The output parameter of player l ,

J^l = The cost function,

c_{ik}^l = The long run fraction of the time that the system is in state i and action k (red-green) is chosen,

l^+ = The size of the buffer (the maximum capacity of the street).

For the players $l = x, y, z$ the following identities hold



Fig. 10.2 Traffic three way singnal control

$$\begin{aligned}\mathbf{J}^{(x)}(c^{(x)}, c^{(y)}, c^{(z)}) &= \left(\sum_{i^{(x)}}^{x^+} \sum_{k^{(x)}}^{M_x} W_{i^{(x)} k^{(x)}}^{(x)} c_{i^{(x)} k^{(x)}}^{(x)} \right) \left(\sum_{i^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} c_{i^{(y)} k^{(y)}}^{(y)} \right) \left(\sum_{i^{(z)}}^{z^+} \sum_{k^{(z)}}^{M_z} c_{i^{(z)} k^{(z)}}^{(z)} \right), \\ \mathbf{J}^{(y)}(c^{(x)}, c^{(y)}, c^{(z)}) &= \left(\sum_{i^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} W_{i^{(y)} k^{(y)}}^{(y)} c_{i^{(y)} k^{(y)}}^{(y)} \right) \left(\sum_{i^{(x)}}^{x^+} \sum_{k^{(x)}}^{M_x} c_{i^{(x)} k^{(x)}}^{(x)} \right) \left(\sum_{i^{(z)}}^{z^+} \sum_{k^{(z)}}^{M_z} c_{i^{(z)} k^{(z)}}^{(z)} \right), \\ \mathbf{J}^{(z)}(c^{(x)}, c^{(y)}, c^{(z)}) &= \left(\sum_{i^{(z)}}^{z^+} \sum_{k^{(z)}}^{M_z} W_{i^{(z)} k^{(z)}}^{(z)} c_{i^{(z)} k^{(z)}}^{(z)} \right) \left(\sum_{i^{(x)}}^{x^+} \sum_{k^{(x)}}^{M_x} c_{i^{(x)} k^{(x)}}^{(x)} \right) \left(\sum_{i^{(y)}}^{y^+} \sum_{k^{(y)}}^{M_y} c_{i^{(y)} k^{(y)}}^{(y)} \right).\end{aligned}$$

The dynamics of the problem in a three one-way intersection for the players x , y and z is defined as follows:

$$a(n) = a_1 : \text{red}(x)/\text{green}(y)/\text{red}(z)$$

$$x(n+1) = \begin{cases} x^+ & x(n) + \xi^x(n) > x^+ \\ x(n) + \xi^x(n) & x(n) + \xi^x(n) \leq x^+, \end{cases}$$

$$y(n+1) = \begin{cases} y^+ & y(n) + \xi^y(n) - v^y(n) > y^+ \\ [y(n) + \xi^y(n) - v^y(n)]_+ & y(n) + \xi^y(n) - v^y(n) \leq y^+, \end{cases}$$

$$z(n+1) = \begin{cases} z^+ & z(n) + \xi^z(n) > z^+ \\ z(n) + \xi^z(n) & z(n) + \xi^z(n) \leq z^+, \end{cases}$$

$$a(n) = a_2 : \text{green}(x)/\text{red}(y)/\text{red}(z)$$

$$x(n+1) = \begin{cases} x^+ & x(n) + \xi^x(n) - v^x(n) > x^+ \\ [x(n) + \xi^x(n) - v^x(n)]_+ & x(n) + \xi^x(n) - v^x(n) \leq x^+, \end{cases}$$

$$y(n+1) = \begin{cases} y^+ & y(n) + \xi^y(n) > y^+ \\ y(n) + \xi^y(n) & y(n) + \xi^y(n) \leq y^+, \end{cases}$$

$$z(n+1) = \begin{cases} z^+ & z(n) + \xi^z(n) > z^+ \\ z(n) + \xi^z(n) & z(n) + \xi^z(n) \leq z^+, \end{cases}$$

$$a(n) = a_3 : red(x)/red(y)/green(z)$$

$$x(n+1) = \begin{cases} x^+ & x(n) + \xi^x(n) > x^+ \\ x(n) + \xi^x(n) & x(n) + \xi^x(n) \leq x^+, \end{cases}$$

$$y(n+1) = \begin{cases} y^+ & y(n) + \xi^y(n) > y^+ \\ y(n) + \xi^y(n) & y(n) + \xi^y(n) \leq y^+, \end{cases}$$

$$z(n+1) = \begin{cases} z^+ & z(n) + \xi^z(n) - v^z(n) > z^+ \\ [z(n) + \xi^z(n) - v^z(n)]_+ & z(n) + \xi^z(n) - v^z(n) \leq z^+. \end{cases}$$

Associated with the three actions a_1 , a_2 and a_3 there are three control strategies defined as

$$c_{i_l k_l}^l = d_{k_l | i_l}^l P^l (s^{(l)} = s_{(i_l)})$$

($l = x, y, z$) with the following restrictions

$$c_{i_l 1}^l, c_{i_l 2}^l, c_{i_l 3}^l > 0, \quad c_{i_l 1}^l + c_{i_l 2}^l + c_{i_l 3}^l = 1,$$

$$c = \|c_{i_l k_l}\|_{k_l=\overline{1,2,3}, i_l=\overline{0, N_l}}, \quad (c_{i_l 1}^l, c_{i_l 2}^l, c_{i_l 3}^l) \in \Delta.$$

Remark 10.1 (Intersection avoiding condition) The realization of the signal control problem requires the actions to be synchronized as follows

$$c_{1|1}^x = c_{1|1}^y, \quad c_{1|2}^x = c_{1|2}^y, \quad c_{(1|3)}^x = c_{(1|3)}^y,$$

$$c_{2|1}^x = c_{2|1}^y, \quad c_{2|2}^x = c_{2|2}^y, \quad c_{(2|3)}^x = c_{(2|3)}^y,$$

$$c_{3|1}^x = c_{3|1}^y, \quad c_{3|2}^x = c_{3|2}^y, \quad c_{3|2}^x = c_{3|3}^y.$$

Fixing $N^x = N^y = N^z = 3, M^x = M^y = M^z = 3$, and $x^+ = y^+ = z^+ = 3$, then we obtain different configurations of the game:

(a) **No leader** (uniform distribution). For $\lambda^x = \lambda^y = \lambda^z = 3$, $\mu^x = \mu^y = \mu^z = 3$, $\gamma = 0.12$ and $\delta = 0.009$ we have that the corresponding transition matrices are as follows:

$$\pi_{j|i,green}^l = \begin{bmatrix} 0.3659 & 0.1446 & 0.4895 \\ 0.2160 & 0.1417 & 0.6423 \\ 0.1027 & 0.1027 & 0.7947 \end{bmatrix}, \quad \pi_{j|i,red}^l = \begin{bmatrix} 0.0514 & 0.1543 & 0.7943 \\ 0.0000 & 0.0319 & 0.9681 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}.$$

For player x we have $\{\pi_{j|i,red}^x, \pi_{j|i,green}^x, \pi_{j|i,red}^x\}$, for y we have $\{\pi_{j|i,green}^y, \pi_{j|i,red}^y, \pi_{j|i,red}^y\}$ and for player z we have $\{\pi_{j|i,red}^z, \pi_{j|i,red}^z, \pi_{j|i,green}^z\}$. The resulting c_{ik}^l ($l = x, y, z$) will be the same for all the players

$$c_{ik}^l = \begin{bmatrix} 0.0136 & 0.0136 & 0.0136 \\ 0.0134 & 0.0134 & 0.0134 \\ 0.3063 & 0.3063 & 0.3063 \end{bmatrix}.$$

Then, fixing k and adding by i we have that the long run fraction of the time for each player is as follows:

$$c_k^l = \begin{bmatrix} \text{Red} & \text{Green} & \text{Red} \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix}, \quad c_k^y = \begin{bmatrix} \text{Green} & \text{Red} & \text{Red} \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix}, \quad c_k^z = \begin{bmatrix} \text{Red} & \text{Red} & \text{Green} \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix}.$$

Then, because the players have the same λ^l and μ^l the distribution of the time is the same for Green and Red lights.

The corresponding strategies $d_{k|i}^l$ ($l = x, y, z$) are as follows:

$$d_{k|i}^l = \begin{bmatrix} 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix}.$$

(b) **One leader and two followers.** We can make player x a leader by making $\lambda^x = 4$, and $\lambda^y = \lambda^z = 3$, $\mu^x = 4$ $\mu^y = \mu^z = 3$, $\gamma = 0.14$ and $\delta = 0.009$ we have that the corresponding transition matrices are as follows:

$$\pi_{j|i,green}^x = \begin{bmatrix} 0.1620 & 0.0606 & 0.7774 \\ 0.0717 & 0.0438 & 0.8844 \\ 0.0228 & 0.0228 & 0.9543 \end{bmatrix}, \quad \pi_{j|i,red}^x = \begin{bmatrix} 0.0119 & 0.0474 & 0.9407 \\ 0.0000 & 0.0056 & 0.9944 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}.$$

For player x we have $\{\pi_{j|i,red}^x, \pi_{j|i,green}^x, \pi_{j|i,red}^x\}$. For players y and z the transition matrices are as follows:

$$\pi_{j|i,green}^{y,z} = \begin{bmatrix} 0.3659 & 0.1446 & 0.4895 \\ 0.2160 & 0.1417 & 0.6423 \\ 0.1027 & 0.1027 & 0.7947 \end{bmatrix}, \quad \pi_{j|i,red}^{y,z} = \begin{bmatrix} 0.0514 & 0.1543 & 0.7943 \\ 0.0000 & 0.0319 & 0.9681 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}.$$

For y we have $\{\pi_{j|i,green}^y, \pi_{j|i,red}^y, \pi_{j|i,red}^y\}$ and for player z we have $\{\pi_{j|i,red}^z, \pi_{j|i,red}^z, \pi_{j|i,green}^z\}$. The resulting c_{ik}^l ($l = x, y, z$) will be the same for all the players

$$c_{ik}^l = \begin{bmatrix} 0.0062 & 0.0050 & 0.0062 \\ 0.0050 & 0.0072 & 0.0050 \\ 0.1319 & 0.7013 & 0.1319 \end{bmatrix}.$$

Then, fixing k and adding by i we have that the long run fraction of the time for each player is as follows:

$$c_k^x = \begin{bmatrix} \text{Red} & \text{Green} & \text{Red} \\ 0.1431 & 0.7135 & 0.1431 \end{bmatrix}, c_k^y = \begin{bmatrix} \text{Green} & \text{Red} & \text{Red} \\ 0.1431 & 0.7135 & 0.1431 \end{bmatrix}, c_k^z = \begin{bmatrix} \text{Red} & \text{Red} & \text{Green} \\ 0.1431 & 0.7135 & 0.1431 \end{bmatrix}.$$

Then, because player x is a leader the time assigned to the Green light (0.7135) is larger than to Red light (0.1431 + 0.1431). The players y and z have the same λ and μ , as a result the distribution of the time is the same for Green (0.1431) and Red lights (0.7135 + 0.7135).

The corresponding strategies $d_{k|i}^l$ ($l = x, y, z$) are as follows:

$$d_{k|i}^l = \begin{bmatrix} 0.3571 & 0.2858 & 0.3571 \\ 0.2909 & 0.4183 & 0.2909 \\ 0.1367 & 0.7266 & 0.1367 \end{bmatrix}.$$

(c) **Two leaders and one follower.** We can make player x and y leaders by making $\lambda^x = \lambda^y = 4$, and $\mu^x = \mu^y = 3$ and player z follower by making $\lambda^z = 3$, $\mu^z = 3$, $\gamma = 0.18$ and $\delta = 0.009$ we have that the corresponding transition matrices for players x and y are as follows:

$$\pi_{j|i,green}^{x,y} = \begin{bmatrix} 0.2715 & 0.1027 & 0.6258 \\ 0.1233 & 0.0779 & 0.7988 \\ 0.0401 & 0.0401 & 0.9198 \end{bmatrix}, \quad \pi_{j|i,red}^{x,y} = \begin{bmatrix} 0.0119 & 0.0474 & 0.9407 \\ 0.0000 & 0.0056 & 0.9944 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}.$$

For player x we have $\{\pi_{j|i,red}^x, \pi_{j|i,green}^x, \pi_{j|i,red}^x\}$ and for player y we have $\{\pi_{j|i,green}^y, \pi_{j|i,red}^y, \pi_{j|i,red}^y\}$. For player z the transition matrices are as follows:

$$\pi_{j|i,green}^z = \begin{bmatrix} 0.3659 & 0.1446 & 0.4895 \\ 0.2160 & 0.1417 & 0.6423 \\ 0.1027 & 0.1027 & 0.7947 \end{bmatrix}, \quad \pi_{j|i,red}^z = \begin{bmatrix} 0.0514 & 0.1543 & 0.7943 \\ 0.0000 & 0.0319 & 0.9681 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}.$$

For player z we have $\{\pi_{j|i,red}^z, \pi_{j|i,red}^z, \pi_{j|i,green}^z\}$. The resulting c_{ik}^l ($l = x, y, z$) will be the same for all the players

$$c_{ik}^l = \begin{bmatrix} 0.0054 & 0.0054 & 0.0082 \\ 0.0066 & 0.0066 & 0.0050 \\ 0.4116 & 0.4116 & 0.1395 \end{bmatrix}.$$

Then, fixing k and adding by i we have that the long run fraction of the time for each player is as follows:

$$c_k^x = \begin{bmatrix} \text{Red} & \text{Green} & \text{Red} \\ 0.4236 & 0.4236 & 0.1527 \end{bmatrix}, c_k^y = \begin{bmatrix} \text{Green} & \text{Red} & \text{Red} \\ 0.4236 & 0.4236 & 0.1527 \end{bmatrix}, c_k^z = \begin{bmatrix} \text{Red} & \text{Red} & \text{Green} \\ 0.4236 & 0.4236 & 0.1527 \end{bmatrix}.$$

The players x and y have the same λ and μ , as a result the distribution of the time is the same for Green (0.4236) and Red lights (0.4236 + 0.1527). Player z , as a follower, has assigned (0.1527) to the Green light and (0.4236 + 0.4236) Red light.

The corresponding strategies $d_{k|i}^l$ ($l = x, y, z$) are as follows:

$$d_{k|i}^l = \begin{bmatrix} 0.2833 & 0.2833 & 0.4334 \\ 0.3633 & 0.3633 & 0.2733 \\ 0.4276 & 0.4276 & 0.1449 \end{bmatrix}.$$

Remark 10.2 Different results can be obtained by fixing specific values for λ and μ .

References

1. Allsop, R.: Sigset: a computer program for calculating traffic capacity of signal-controlled road junctions. *Traffic Eng. Control* **12**, 58–60 (1971)
2. Allsop, R.: Sigeap: A computer program for assessing the traffic capacity of signal-controlled road junctions. *Traffic Eng. Control* **17**, 338–341 (1976)
3. Alvarez, I., Poznyak, A.S., Malo Tamayo, A.: Urban traffic control problem: a game theory approach. In: Proceedings of the 17th World Congress IFAC, The International Federation of Automatic Control Seoul, Korea, pp. 7154–7159 (2008)
4. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. *Comput. Math. Math. Phys.* **45**(11), 1893–1914 (2005)
5. Aragon-Gomez, R., Clemptner, J.B.: Traffic-signal control reinforcement learning approach for continuous-time markov games. *Eng. Appl. Artif. Intell.* **89**, 103415 (2020)
6. Cantarella, G.E., Improta, G., Sforza, A.: Road network signal setting: equilibrium conditions. In: Concise Encyclopedia of Traffic and Transportation Systems. Pergamon Press, Amsterdam (1991)
7. Casccetta, E.: Transportation Systems Analysis: Models and Applications. Springer, New York (2009)
8. Casccetta, E., Gallo, M., Montella, B.: An asymmetric sue model for the combined assignment-control problem. In: Selected Proceedings of 8th WCTR. Pergamon Press, Amsterdam (1999)
9. Casccetta, E., Gallo, M., Montella, B.: Models and algorithms for the optimization of signal settings on urban networks with stochastic assignment. *Ann. Oper. Res.* **144**, 301–328 (2006)
10. Castillo-Gonzalez, R., Clemptner, J.B., Poznyak, A.S.: Solving traffic queues at controlled-signalized intersections in continuous-time markov games. *Math. Comput. Simul.* **166**, 283–297 (2019)

11. Chen, O.J., Ben-Akiva, M.E.: Game-theoretic formulations of interaction between dynamic traffic control and dynamic traffic assignment. *Transp. Res. Rec.* **1617**, 179–188 (1998)
12. Chiou, S.W.: Optimization of area traffic control for equilibrium network flows. *Transp. Sci.* **33**, 279–289 (1999)
13. Clemptner, J.B., Poznyak, A.S.: Convergence method, properties and computational complexity for lyapunov games. *Int. J. Appl. Math. Comput. Sci.* **21**(2), 349–361 (2011)
14. Clemptner, J.B., Poznyak, A.S.: Analysis of best-reply strategies in repeated finite markov chains games. In: 52nd IEEE Conference on Decision and Control, pp. 568–573. Florence, Italy (2013)
15. Clemptner, J.B., Poznyak, A.S.: Modeling the multi-traffic signal-control synchronization: a markov chains game theory approach. *Eng. Appl. Artif. Intell.* **43**, 147–156 (2015)
16. D’Acierno, L., Gallo, M., Montella, B.: n ant colony optimisation algorithm for solving the asymmetric traffic assignment problem. *Eur. J. Oper. Res.* **217**, 459–469 (2012)
17. Dafermos, S.: Traffic equilibrium and variational inequalities. *Transp. Sci.* **14**, 42–54 (1980)
18. Dafermos, S.: Relaxation algorithms for the general asymmetric traffic equilibrium problem. *Transp. Sci.* **16**, 231–240 (1982)
19. Fisk, C.S.: Game theory and transportation systems modelling. *Transp. Res. B* **18**(4–5), 301–313 (1984)
20. Fisk, C.S., Nguyen, S.: Solution algorithms for network equilibrium models with asymmetric user costs. *Transp. Sci.* **16**, 361–381 (1982)
21. Florian, M., Spiess, H.: The convergence of diagonalization algorithms for asymmetric network equilibrium problems. *Transp. Res. B* **16**, 477–483 (1982)
22. Gallo, M., D’Aciernob, L.: Comparing algorithms for solving the local optimisation of the signal settings (loss) problem under different supply and demand configurations. *Proc. Soc. Behav. Sci.* **87**, 147–162 (2013)
23. Gartner, N.: Opac: a demand responsive strategy for traffic signal control. *Transp. Res. Rec.* **906**, 75–81 (1983)
24. Gartner, N., Al-Malik, M.: Combined model for signal control and route choice in urban traffic networks. *Transp. Res. Rec.* **1554**, 27–35 (1996)
25. Gartner, N., Gershwin, S.B., Little, J.D.C., Ross, P.: Pilot study of computer-based urban traffic management. *Transp. Res. B* **14**(1–2), 203–217 (1980)
26. Gartner, N., Little, J., Gabby, H.: Simultaneous optimization of offsets, splits and cycle time. *Transp. Res. Rec.* **596**, 6–15 (1976)
27. Gartner, N., Stamatiadis, C.: Framework for the integration of dynamic traffic assignment with real-time control. In: Proceedings of the 3rd Annual World Congress on Intelligent Transportation Systems, Orlando (1996)
28. Heydecker, B.: A decomposition approach for signal optimisation in road networks. *Transp. Res. B* **30**, 99–114 (1996)
29. Heydecker, B., Khoo, T.: The equilibrium network design problem. In: Proceedings of AIRO ’90, Conference on Models and Methods for Decision Support, Sorrento, pp. 587–602 (1990)
30. Lee, C., Machemehl, R.B.: Genetic algorithm, local and iterative searches for combining traffic assignment and signal control. In: Proceedings of the Conference on Traffic and Transportation Studies (ICTTS ’98), Beijing, China, pp. 27–29, (1998)
31. Meneguzzi, C.: Implementation and evaluation of an asymmetric equilibrium route choice model incorporating intersection-related travel times. Master’s thesis, Ph.D. Dissertation, Department of Civil Engineering, University of Illinois, Urbana (1990)
32. Meneguzzi, C.: An equilibrium route choice model with explicit treatment of the effect of intersections. *Transp. Res. B* **29**, 329–356 (1995)
33. Moya, S., Poznyak, A.S.: Extraproximal method application for a stackelberg-nash equilibrium calculation in static hierarchical games. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **39**(6), 1493–1504 (2009)
34. Placzek, B.: A self-organizing system for urban traffic control based on predictive interval microscopic model. *Eng. Appl. Artif. Intell.* **34**, 75–84 (2014)

35. Poznyak, A.S.: Advance Mathematical Tools for Automatic Control Engineers, vol. 2. Stochastic Techniques. Elsevier, Amsterdam (2009)
36. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: Self-learning Control of Finite Markov Chains. Marcel Dekker, Inc., New York, (2000)
37. Qu, Z., Xing, Y., Song, X., Duan, Y., Wei, F.: A study on the coordination of urban traffic control and traffic assignment. *Discrete Dyn. Nat. Soc.* (2012). <https://doi.org/10.1155/2012/367468>
38. Sheffi, Y., Powell, W.: Optimal signal settings over transportation networks. *J. Transp. Eng.* **109**, 824–839 (1983)
39. Smith, M.J.: The existence, uniqueness and stability of traffic equilibria. *Transp. Res. B* **13**, 295–304 (1979)
40. Smith, M.J.: Traffic control and route-choice; a simple example. *Transp. Res. B* **13**, 289–294 (1979)
41. Smith, M.J.: A local traffic control policy which automatically maximizes the overall travel capacity of an urban road network. *Traffic Eng. Control.* **21**, 298–302 (1980)
42. Smith, M.J.: Properties of a traffic control policy which ensure the existence of a traffic equilibrium consistent with the policy. *Transp. Res. B* **15**, 453–462 (1981)
43. Tan, H.N., Gershowin, S.B., Athans, M.: Hybrid Optimization in Urban Traffic Networks. Technical Report, MIT Press, Cambridge (1979)
44. Tanaka, K., Yokoyama, K.: On ϵ -equilibrium point in a noncooperative n-person game. *J. Math. Anal.* **160**, 413–423 (1991)
45. Trejo, K.K., Clemmpner, J.B., Poznyak, A.S.: A stackelberg security game with random strategies based on the extraproximal theoretic approach. *Eng. Appl. Artif. Intell.* **37**, 145–153 (2015)
46. Wey, W.M.: Model formulation and solution algorithm of traffic signal control in an urban network. *Comput. Environ. Urban Syst.* **24**, 355–377 (2000)
47. Wong, S., Yang, H.: Reserve capacity of a signal-controlled road network. *Transp. Res. B* **31**, 397–402 (1997)
48. Wong, S.C.: Group-based optimization of signal timings using parallel computing. *Transp. Res. C* **5**, 123–139 (1997)
49. Yang, H., Yagar, S.: Traffic assignment and signal control in saturated road networks. *Transp. Res. A* **29**, 125–139 (1995)
50. Ziyou, G., Yifan, S.: A reserve capacity model of optimal signal control with user-equilibrium route choice. *Transp. Res. B* **36**, 313–323 (2002)

Chapter 11

Non-cooperative Bargaining with Unsophisticated Agents



Abstract In a conventional non-cooperative negotiating scenario, two or more forward-thinking participants make offers and counteroffers alternately until an agreement is achieved, with a penalty taking into account the length of time it takes players to make a decision. We provide a game that helps myopic participants achieve equilibrium as if they were forward-thinking agents. One of the game's main mechanics is that players are penalized for deviating from their prior best reply plan as well as for the amount of time they spend to make decisions at each stage of play. Our chapter adds to existing research on typical myopic agent bargaining while also broadening the class of processes and functions that may be used to define and apply Rubinstein's non-cooperative bargaining solutions.

11.1 Introduction

There is a substantial and expanding body of research on *non-cooperative bargaining*. A bilateral non-cooperative bargaining process was represented by Rubinstein [38] as an alternating offers game with a cost for each round of negotiations. In several articles and scenarios, this model has been examined and expanded for three or more players (e.g., [3, 10, 11, 20, 23, 31, 35, 41]). The non-cooperative bargaining model and its game-theoretic solution have also been used in a variety of significant contexts, including market games [9, 39], networks [1, 2, 18, 25], apex games [27], union formation [22] and water management [13].

Despite its broad applicability, the traditional bargaining model makes several key assumptions, including that players are sophisticated in their behavior (for example, when agents are forward-looking or in the presence of externalities) and that players have complete information about the characteristics of other agents (for example, their discount factor or their utility). It has been demonstrated that the classic equilibrium idea fails when agents lack sophistication, such as when they are myopic (egocentric) [21, 24, 32–34, 40]. Therefore, it is necessary to create a general theory of bargaining that is strong enough to function in the absence of knowledgeable individuals.

Consider the Rubinstein-proposed bargaining scenario, where two players must agree on how to divide a pie of size 1 and alternately make proposals to the other player at intervals of $0, \Delta, 2\Delta, \dots$ where $\Delta > 0$, to serve as an example. Each suggestion outlines how the pie should be divided in $(x^1, x^2) \in \mathbb{R}_+^2$ (the case where $x^1 + x^2 = 1$). Also take into account the fixed discounting factors for each player $\beta^1 = e^{(-r^1\Delta)}$ and $\beta^2 = e^{(-r^2\Delta)}$, which are dependent on the individual player's discount rate r .

Keep in mind that for this approach to work, agents must be aware of each other's utility function and discount factor. Since agent 1's ability to provide x^{*1} depends on β^2 , this equilibrium does not hold when participants are unaware of the discount factor of others.

Consider a similar naïve (unsophisticated) agent, who would just perform a best-response equilibrium using their own knowledge, maybe uninformed or dismissing the utilities and actions of other participants. Consider, for example, that if a response rejects an offer, the game always continues, but the proposer never sees the game continue. The conventional Rubinstein model does not converge in such a case. In fact, take into account the scenario in which there is a set item that is divided between two players, for simplicity, use 1 as the value of the good. In each round of the negotiating process, each player's optimal move is to obtain all for himself while making no offers to the other player, i.e., in the first round, player 1 offers $x^2 = 0$ to player 2 and keeps $x^1 = 1$ for himself, then player 2 rejects and the best response is to offer $x^1 = 0$ to player 1 and keep $x^2 = 1$ for himself. The conduct of both players will be the same in the following round, trapping both players in the optimal response situation, and this straightforward game will never have a solution.

We suggest a different strategy to the conventional bargaining literature, where a planner has the ability to set up a game to help the agents reach an equilibrium. This strategy is intended to help implement bargaining solutions in the presence of inexperienced agents or those lacking knowledge about the characteristics of other agents. Consider the traditional bargaining game described above to better understand the characteristics of our game. In this game, the planner can penalize the agents based on two different criteria: first, they can be punished for deviating from the previous proposal, and second, they can be punished for the amount of time it takes to make a decision at each stage of the game. Our major finding demonstrates that, even for extremely modest penalties, the tatonnement process converges (in exponential time) over a very broad class of games and functions (Theorem 11.1).

Another benefit of our game is that, for two players, it converges to a compromise solution that is equivalent to Rubinstein's answer. It is obvious that when players are inexperienced, just imposing a severe penalty for deviating from the prior approach may not result in a compromise solution. In fact, the first agent can propose to obtain all resources for a sufficiently high penalty, and the second agent would lack the motivation to deviate since he would be required to pay a high penalty depending on his deviation. Similar to how agents could cycle in their offerings as mentioned above, a tiny penalty might not result in the implementation of a compromise solution. Our key contribution demonstrates that even in the presence of unsophisticated (and ignorant) players who just play the best-response, a very limited collection of

Table 11.1 Characteristics of the game and equilibrium proposed in comparison to the original Rubinstein's model and solution

	Rubinstein's equilibrium	Proposed equilibrium
Two-player game	✓	✓
3 or more player game		✓
Complete information	✓	✓
Incomplete information		✓
Alternating offers	✓	✓
Discount factor	✓	✓
Exists for rational players	✓	✓
Exists for unsophisticated agents		✓
Markov processes		✓

penalties may ensure the convergence of the aforementioned technique. Additionally, this convergence holds true for a broad and extremely complete collection of stochastic processes (like Markov processes) and utility functions (Sect. 11.4), in addition to the conventional processes and utilities utilized in classic bargaining literature. As a result, our study not only adds to existing research on conventional bargaining for inexperienced, ignorant actors but also broadens the range of processes and functions that may be used to define and apply the classic Rubinstein's non-cooperative bargaining problem (Table 11.1).

In order to further illustrate our method and the solution obtained in comparison to Rubinstein's, consider a game where two players have to reach an agreement on the partition of a fixed divisible good (for simplicity with value 1). Following the alternating offers model, we consider that players makes offers at times $0, \Delta, 2\Delta, \dots$ where $\Delta = 0.5$. Both players have a discount rate $r = 0.5$, then the fixed discounting factor for each player is the same and is given by $\beta = e^{(-r\Delta)} = e^{(-0.25)} = 0.7788$. The traditional Rubinstein's solution computed $x^{1*} = x^{2*} = 0.5622$, and implements the equilibrium as follows:

- Player 1 always offers $(0.5622, 0.4378)$ and always accepts at least $(0.4378, 0.5622)$,
- Player 2 always offers $(0.4378, 0.5622)$ and always accepts at least $(0.5622, 0.4378)$.

Thus, when player 1 makes the first offer, the agreement is reached in the first offer, at time 0, and the final payoffs are $(0.5622, 0.4378)$.

In comparison, while the solution employing our method considering a time function equal to the discount factor β proposed by Rubinstein is as Table 11.2. Our method also penalizes players for deviating from the previous offer made by the players, $\delta|x_n - x_{n-1}|$. In our method, the agreement reached by players is $(0.5653, 0.4347)$ when player 1 makes the first offer, and the final utilities after

Table 11.2 Behavior of the offers and utilities from our method. The offers column (x^1 and x^2) represent the best response made by players in response to their utilities and the penalties imposed by the planner. The total penalty at step n equals $\delta|x_n - x_{n-1}|$. The Accum. penalty column represent the accumulated penalty over all steps, and the utilities equal the offers made minus the penalty. Even with ten steps, both agents' best response is equivalent at a small penalty/cost to the players

	Step	Offers		$\delta \times 1e^4$	$\ x_n - x_{n-1}\ ^2$	Accum. penalty	Utilities	
		x^1	x^2				$\psi^1(x)$	$\psi^2(x)$
	0	0.5	0.5				0.5	0.5
1 →	1	0.7472	0.2528	0.7788	0.1222	0.00001	0.74719	0.25279
2 →	2	0.4973	0.5027	0.6065	0.1249	0.00002	0.49728	0.50268
1 →	3	0.6874	0.3126	0.3749	0.0723	0.00002	0.68738	0.31258
2 →	4	0.5145	0.4855	0.2551	0.0598	0.00002	0.51448	0.48548
1 →	5	0.67	0.33	0.1305	0.0484	0.00002	0.66998	0.32998
2 →	6	0.539	0.461	0.0956	0.0343	0.00002	0.53898	0.46098
1 →	7	0.6727	0.3273	0.0707	0.0358	0.00002	0.67268	0.32728
2 →	8	0.5412	0.4588	0.0527	0.0346	0.00002	0.54118	0.45878
1 →	9	0.5655	0.4345	0.0395	0.0012	0.00002	0.56548	0.43448
2 →	10	0.5653	0.4347	0.0297	0.0000	0.00002	0.56528	0.43468

the penalization equals $\psi^1(x) = 0.56528$ and $\psi^2(x) = 0.43468$, which is similar to Rubinstein's solution.

11.1.1 Related Literature

The Rubinstein's Alternate Bargaining Game [38] demonstrated the existence of equilibrium under complete knowledge through a non-cooperative bargaining procedure for two players splitting a fixed resource equal to 1. Fudenberg and Tirole [15], who uses the idea of perfect Bayesian equilibrium to analyze and solve a two-player, two-period non-cooperative bargaining game with imperfect information. The modified evolutionary stable strategies (MESS) in Rubinstein's alternating-offers, infinite-horizon bargaining game were described by Binmore et al. [10]. They demonstrated that the presence of a MESS results in agreement being reached right away, with neither player preferring to postpone the agreement by one period in order to claim the other player's portion of the surplus. In the particular subgame-perfect equilibrium of Rubinstein's game, each player's share of the surplus is then limited to the shares that they each got from the other player.

By taking into account a n -player bargaining issue where the utility possibility set is compact, convex, and strictly comprehensive, Kultti and Vartiainen [23] enhanced Rubinstein's approach. They demonstrated the existence of a stationary subgame perfect Nash equilibrium and showed that all such equilibria converge to the Nash

bargaining solution if the Pareto surface is differentiable as the length of a time interval between offers approaches zero, i.e., they demonstrated that the unique subgame perfect equilibrium result of the two-player alternating offers bargaining game proposed by Rubinstein converges to the Nash bargaining solution [30] when the length of time interval approaches zero.

It can be interpreted as a weighted majority game in which the major player has $n - 2$ votes, each minor player has one vote, and $n - 1$ votes are necessary for a majority. Montero [27] studied non-cooperative bargaining with random proposers in apex games, a n -player game with one major player and $n - 1 \geq 3$ minor players. The egalitarian protocol picks each player as the proposer with an equal likelihood, while the proportional protocol selects each player with a probability proportionate to the amount of votes he receives. These are the two procedures that Montero took into consideration.

The negotiating problem has been extensively researched for situations in which the participants are shortsighted, such in our scenario. A model of myopic tastes was introduced and described by Brown and Lewis [12], both in the context of intertemporal decision-making and choosing under uncertainty, including an uncountable number of eras or global tastes. By establishing topologies on the space of consumption plans that discount unlikely or future occurrences, they formalized myopic (short-sighted) behavior. With myopic agents and static equilibrium concepts similar to the Nash equilibrium, Marden [26] proposed a concept he called “state based potential games.” This introduces an underlying state space into the framework of potential games and shows how state based potential games can be applied to two cooperative control problems, distributed resource allocation and local control design. These examples were presented in terms of Markov chains.

The benefits and drawbacks of non-monotonic offers in alternating-offer bargaining methods were discussed by Winoto et al. [52]. He imagined a negotiation between a buyer and a seller on a single characteristic item. The justification in this case is that certain agents could be risk-averse toward collapse in the future and myopic at a different level. Therefore, even if a superior offer later on becomes available, people can choose a safe but still viewed as ideal one. Rubinstein’s negotiation between potentially time-inconsistent participants was investigated by Akin [4]. He looked at how learning and time inconsistency affected the choices made by various types of agents in a framework for bargaining. He took into account two players—one each of naive, sophisticated, partially naive, or exponential-playing in an infinite-horizon alternating-offer negotiation under the presumption that each player is aware of the type of the other player and that the naive players can learn about their own preferences. He demonstrated that the more ignorant a player is, the bigger the share obtained in a game between them and a time-consistent player. He also demonstrated that two naive players who never learn from their mistakes will always disagree. Players remain ignorant regardless of how the game changes since there is no learning; they do not update their beliefs or, as a result, their strategy. The current worth of what the naïve and somewhat naive agents believe their opponent expects to receive in exchange for rejecting is always what they give. But it turns out that from the standpoint of the opponent, rejection based on their ideas is always the best

course of action since their views are different from what their opponent believes. Thus, each of them is so optimistic in their upcoming shares and obstinate that they insist on giving the other the identical rejected share.

In contrast to previous work, our game is specified, and there are equilibria for a sizable class of games with utility functions and both complete and imperfect information.

The main contributions of this chapter are as follows:

- Introduces a new non-cooperative bargaining game between two or more myopic players that induce them to behave as if they were forward-looking players.
- Consideres a time penalization related with the time spent for each player for the decision-making at each step of the negotiation process as well as their deviation from the previous best response strategy. Techniques from linear optimization, especially the proximal algorithm, were used to show the existence and feasible computability of the equilibrium.
- Complements the study of bargaining for unsophisticated agents but also enlarges the class of processes and functions where non-cooperative bargaining model might be defined and applied.

11.2 The Rubinstein's Alternating-Offers Model

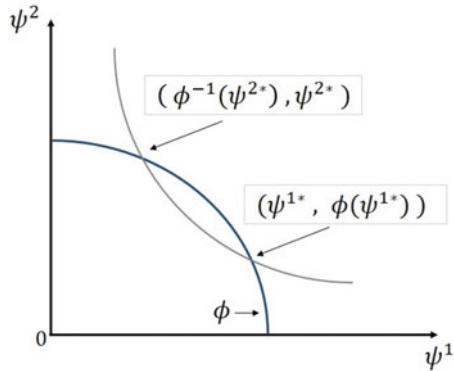
Rubinstein [38] defined seminal bargaining situation for two players ($n = 2$) who have to reach an agreement on the partition of a pie of size 1. Each player takes turns to make an offer to the other agents on how the pie should be divided between them. After player 1 has made such an offer, player 2 must decide whether to accept it, in this case the bargaining game ends and the players divide the cake according to the accepted offer, or to reject it and continue with the bargaining process. If player 2 rejected, then this player has to make a counteroffer which player 1 would accept or reject it and continue with the negotiation process. The bargaining game continues until an offer is accepted. Offers are made at discrete points in time, $0, \Delta, 2\Delta, \dots, t\Delta, \dots$, where $\Delta > 0$, and players experience an exponential discount factor which might be different across agents. For a player ι , and an offer x^* accepted at time t , his final utility is $x_\iota^* e^{(-r^\iota t \Delta)}$, where $\beta^\iota = e^{(-r^\iota \Delta)}$ is the discount factor associated to player ι .¹

Rubinstein's [38] main results shows the existence of a subgame perfect equilibrium. Indeed, for

$$x^{1*} = \frac{1 - \beta^2}{1 - \beta^1 \beta^2} \quad x^{2*} = \frac{1 - \beta^1}{1 - \beta^1 \beta^2}$$

¹ Reference [38] also studies the case of a fixed linear cost $c^\iota t \Delta$, instead of exponential, associated at every step. Our work focuses on exponential discounting rather than linear since it produces richer solutions.

Fig. 11.1 The Pareto solution of the bargaining problem at time 0



a *subgame perfect equilibrium* can be found when player 1 always offers x^{1*} and always accepts an offer x^2 if and only if $x^2 \leq x^{2*}$; and player 2 always offers x^{2*} and always accepts an offer x^1 if and only if $x^1 \leq x^{1*}$. Such an equilibrium is unique and reached at time zero when $r^\iota > 0$ for all ι .

Such a game can be extended to a general set X of possible agreements. Indeed, consider two players bargaining over X according to the alternating-offers as above, where player ι has utility function $\psi^\iota : X \rightarrow \mathbb{R}$ and exponential discount factor $e^{(-r^\iota t\Delta)}$. We denote by $\Phi(X) = \{(\psi^1(x), \psi^2(x)) | x \in X\}$ the set of possible utility pairs attainable at time 0, and Φ^e denote the Pareto frontier² of the set Φ .

When X is a compact and convex set, and the utility functions are continuous and concave, the Pareto frontier Φ^e can be represented by a graph function of a strictly decreasing and concave function, denoted by ϕ , whose domain is an interval $I^1 \subseteq \mathbb{R}$ and range an interval $I^2 \subseteq \mathbb{R}$. For simplicity, assume that $0 \in I^1$, $0 \in I^2$ and $\phi(0) > 0$ (see, Fig. 11.1). Then,

$$\Phi^e = \{(\psi^1, \psi^2) : \psi^1 \in I^1, \psi^2 \in \phi(\psi^1)\}$$

Consider ϕ^{-1} the inverse of ϕ , a strictly decreasing and concave function from I^2 to I^1 , with $\phi^{-1}(0) > 0$. Then, for any $\psi^1 \in I^1$, $\phi(\psi^1)$ is the maximum utility that player 2 receives subject to player 1 receiving a utility ψ^1 ; in the same way, for any $\psi^2 \in I^2$, $\phi^{-1}(\psi^2)$ is the maximum utility that player 1 receives subject to player 2 receiving a utility ψ^2 .

Let Z^ι , a non-empty subset of X , defined as follows

$$Z^\iota = \left\{ x^\iota := \arg \max_{x \in X} \psi^\iota(x) : \psi^m(x^\iota) = \beta^m \psi^m(x^m), (m \neq \iota) \right\}. \quad (11.2.1)$$

² A utility pair $(\psi^1, \psi^2) \in \Phi^e$ if and only if $(\psi^1, \psi^2) \in \Phi$ and there does not exist another utility pair $(\varphi^1, \varphi^2) \in \Phi$ such that $\varphi^1 \geq \psi^1, \varphi^2 \geq \psi^2$.

Proposition 11.1 *For any $x^{\iota*} \in Z^\iota$, $\iota = 1, 2$, the following pair of strategies is a subgame perfect equilibrium of the general Rubinstein model:*

- *Player 1 always offers x^{1*} and always accepts an offer x^2 if and only if $\psi^1(x^2) \geq \beta^1\psi^{1*}$*
- *Player 2 always offers x^{2*} and always accepts an offer x^1 if and only if $\psi^2(x^1) \geq \beta^2\psi^{2*}$*

where $\psi^{1*} = \phi^{-1}(\beta^2\psi^{2*})$ and $\psi^{2*} = \phi(\beta^1\psi^{1*})$.

The generality of this Bargaining model and the proof of this result can be found in [29]. Note that if Z^ι contains more than one element, then there exist more than one subgame perfect equilibrium in the general Rubinstein model. In any subgame perfect equilibrium, if agreement is reached at time 0 and it is player 1 who makes the offer, then the equilibrium payoff for player 1 is ψ^{1*} and for player 2 is $\phi(\psi^{1*})$; similarly, if it is player 2 who makes the offer at time 0, then the equilibrium payoff for player 1 is $\phi^{-1}(\psi^{2*})$ and for player 2 is ψ^{2*} . This equilibrium pair is Pareto efficient (See Fig. 11.1).

11.3 Bargaining with Unsophisticated Players

We present a variation of [38]’s model that works when agents use a myopic best-response behavior rather than the forward-looking behavior used by Rubinstein. The game presented below works for any time discount function $T : \mathbb{R}_+^n \rightarrow \mathbb{R}_+^n$ that adds a set of penalties to agents that deviate from the previous proposed strategy.

Definition 11.1 (*Bargaining game with disagreement penalties*) A bargaining game with disagreement penalties consists of:

- A set of players $\mathcal{N} = \{1, \dots, n\}$ making offers alternatively following the order $1, 2, \dots, n, 1, 2, \dots$
- The common space of offers X that is a convex and compact set. Once an offer is proposed, it needs to be accepted by all players for it to be final.
- For a given path of offers (x_1, x_2, \dots, x_t) , where $x_l \in X$ and x_t the finally accepted offer, the payoff to agent ι is given by $T^\iota(t)(\psi^\iota(x_t) - D^*(x_1, x_2, \dots, x_t))$, where $\psi : X \rightarrow \mathbb{R}_+^n$ is a concave twice-differentiable real-valued function that represents the utility of agents from accepting x_t , and

$$D^*(x_1, x_2, \dots, x_t) = \sum_{l=1}^t \delta_l \| (x_l - x_{l-1}) \|^2 \quad (11.3.1)$$

is the disagreement cost to the agents at the path (x_1, x_2, \dots, x_t) for some arbitrary initial point $x_0 \in X$ and a sequence of penalties $\delta_1, \delta_2, \delta_3, \dots$.

Like the general Bargaining model discussed in Sect. 11.2, we only assume that offers are made from an arbitrary set X that is convex and compact and the utility function ψ_ι of every agent is a concave twice-differentiable real-valued function. In contrast with such a model, our time discount function $T^\iota(t)$ is an arbitrary decreasing function that is not necessarily exponential.

Our game introduces penalties (costs) imposed to agents when deviating from a previously proposed offer. This is captured by the function D^* that is increasing on both time to reach the offer, as well as the distance from the previously rejected offer. In particular, the cost at time t equals the cost at time $t - 1$ plus a penalty δ_t of the distance from the previous offer, $\delta_t \|x_t - x_{t-1}\|^2$.³

Note that when the sequence of penalties are equal to zero, $\delta_t = 0$ for all t , and the time function equals $T^\iota(l) = e^{(-r^\iota l \Delta)}$, then this resembles the general Rubinstein's model discussed in Sect. 11.2.

Definition 11.2 (*Tatonnement equilibrium*) A path of offers (x_1, x_2, \dots, x_t) is a Tatonnement (unsophisticated best-response) equilibrium if at every step l , where $1 \leq l \leq t$ the offer x_l proposed by agent ι satisfies

$$x_l \in \arg \max_{x \in X} T^\iota(t)(\psi^\iota(x) - D^*(x_1, x_2, \dots, x_{l-1}, x)). \quad (11.3.2)$$

Under a tatonnement equilibrium, agents choose their best response disregarding the behavior of the other agents. This is a standard equilibrium concept formulated from the early origins of Nash equilibrium. Benefits of such an equilibrium includes that agents only need their own information to make a decision. Furthermore, such an equilibrium works whether agents are sophisticated or not.

Theorem 11.1 Consider an arbitrary time discount function $T : \mathbb{R}_+^n \rightarrow \mathbb{R}_+^n$, and strategy space X be a convex and compact polytope. Then, for any initial point x_0 , there exists a sequence of penalties $\delta_1, \delta_2, \dots$ such that the bargaining game with disagreement penalties for utility ψ , time function T and penalties $\delta_1, \delta_2, \dots$ have a tatonnement equilibrium that converges. Furthermore, if the derivative of the utility function ψ is Lipschitz continuous with constant K , the sequence $\{x_t\}$ generated by the procedure, monotonically converges with exponential rate $q(\delta, K)$ to one of the equilibria point x^* , i.e.,

$$\|x_{t+1} - x^*\|^2 \leq q(\delta, K)^{t+1} \|x_0 - x^*\|^2 \quad \text{as } t \rightarrow \infty. \quad (11.3.3)$$

The main implication of this theorem is that even when agents lack sophistication in their behavior, achieving a compromise equilibrium is possible by imposing monetary penalties in their utility. This result is remarkably strong, both on the time

³ While we currently assume that all the agents receive the same penalty $D^*(x_1, x_2, \dots, x_t)$, our work can be extended to asymmetric penalties, for instance, when only the proposing agents is penalized. We note that penalizing all the agents symmetrically guarantees a faster convergence than an asymmetric penalty.

Table 11.3 Number of steps needed for convergence of the strategies of the space X as a function of the minimal distance d between the strategy at time t and an equilibrium

Step	$(9/13)^{t+1}$
1	0.479289941
2	0.331816113
3	0.229718847
4	0.159036125
5	0.110101933
6	0.076224415
7	0.052770749
8	0.036533595
9	0.025292489
10	0.017510185
11	0.012122436
:	:
18	0.000924026
19	6.3971×10^{-4}
20	4.42876×10^{-4}

discount function used, and regardless of the initial point x_0 that is used. Furthermore, the class of problems that it covers is very general, as minimal concavity conditions on the utility functions are assumed. This will be illustrated in the following section, where we apply this to very general problems that include, for instance, continuous-time Markov chains.

Important to note is that unlike Rubinstein's equilibrium, our equilibrium will not converge at time zero, as agents are not fully rational and take some time for the penalties to incentivize them reach a rational equilibrium outcome (*the price of unsophistication*). As such, there is an implicit loss in the efficiency, which can be expressed either as the speed of convergence that it takes for agents to reach a rational equilibrium or in terms of the size of penalties D^* that need to be imposed to agents for them to achieve the equilibrium. Indeed, Remark 11.1 shows that best speed of convergence of the process, which can be achieved by implementing the appropriate penalties to agents. Example 11.1 illustrates the *price of unsophistication* for the division of one unit of a good.

Remark 11.1 The non-cooperative bargaining process converges with exponential rate $q_{\min} = 9/13$, which means that the best rate of convergence at each step of the bargaining process is $(9/13)^{t+1}$ (see, Table 11.3).

Example 11.1 (*The price of unsophistication for the simplex*) In the following example, we measure the cost of having players that are unsophisticated in relation to Rubinstein's main result. As shown in the main result, we say that the cost D at the equilibrium is small. Consider two players dividing a good with value 1, the

Table 11.4 Behavior of the offers and utilities

	Step	Offers		Utilities	
		x^1	x^2	$\psi^l(x)$	$\psi^2(x)$
	0	1	0	1	0
1 →	1	0.9997	0.0003	0.9996	0.0002
2 →	2	0.7956	0.2044	0.7955	0.2043
1 →	3	0.9363	0.0637	0.9362	0.0636
2 →	4	0.7322	0.2678	0.7321	0.2677

solution of Rubinstein when the discount rate $r = 1$ and $\Delta = 1$ is the same for each player $(0.7311, 0.2689)$. Now, for some parameter δ for both players and an initial point $x_0 = (1, 0)$, the solution obtained with our method is $(0.7322, 0.2678)$ and the utilities reached at each step of the process are as follows (Table 11.4).

Our work generalizes the above example for any time function and a very general class of utility functions that are differentiable and satisfy a general Lipschitz condition. Surprisingly, the unsophisticated best-response strategies converge in exponential time.

11.4 An Extension to Continuous-Time Markov Chains

In this section, we extend the convergence results of the previous section to the case of bargaining in the presence of continuous-time Markov chains. In particular, our main results illustrate the computability of the solutions [46, 49, 50].

Definition 11.3 ([19]) A controllable continuous-time Markov chain is a 4-tuple

$$CTMC = (S, A, \mathbb{K}, Q), \quad (11.4.1)$$

where:

- S is the state space, which is a finite set of states $\{s_{(1)}, \dots, s_{(N)}\}$, $N \in \mathbb{N}$, endowed with a discrete topology;
- A is the set of actions, a finite space endowed with the corresponding Borel σ -algebra $\mathcal{B}(A)$. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$;
- $\mathbb{K} = \{(s, a) | s \in S, a \in A(s)\}$ is the class of admissible state-action pairs, which is considered as a topological subspace of $S \times A$;
- Q is the matrix of the transition rates $[q_{j|ik}]$, the transition from state $s_{(i)}$ to state $s_{(j)}$ under an action $a_{(k)} \in A(s_{(i)})$, $k = 1, \dots, M$; satisfying $q_{j|ik} \geq 0$ for all $(s, a) \in \mathbb{K}$ and $i \neq j$ such that

$$[q_{j|ik}] = \begin{cases} -\sum_{i \neq j}^N \rho_{j|i}(a_k), & \text{if } i = j, \\ \rho_{j|i}(a_k), & \text{if } i \neq j, \end{cases}$$

where $\rho_{(j|i)}$ is a transition rate between state $s_{(i)}$ and s_j , $\rho_i = \sum_{i \neq j}^N \rho_{j|i}$. This matrix

is assumed to be conservative, i.e., $\sum_{j=1}^N q_{j|ik} = 0$ and stable, which means that

$$q_i^* := \sup_{a_{(k)} \in A(s_{(i)})} q_{i,k} < \infty \quad \forall s_{(i)} \in S,$$

where $q_{i,k} := -q_{i|i,k} \geq 0$.

Definition 11.4 A continuous-time Markov Decision Process is a pair

$$CTMDP = \{CTMC, U\}, \quad (11.4.2)$$

where:

- CTMC is a controllable continuous-time Markov chain (11.4.1);
- $U : S \times \mathbb{K} \rightarrow \mathbb{R}$ is a utility function, associating to each state a real value.

A strategy (policy) is defined as a sequence $d = \{d(t), t \geq 0\}$ of stochastic kernels $d(t)$ such that: for each time $t \geq 0$, $d_{k|i}(t)$ is a probability measure on A such that $d_{(A(s_i)|i)}(t) = 1$ and for every $E \in \mathcal{B}(A)$ $d_{(E|i)}(t)$ is a Borel measurable function in $t \geq 0$. Let us denote the collection $\{d_{k|i}(t)\}$ by D .

Definition 11.5 A continuous-time Markov game is a pair

$$\mathcal{G} = \{\mathcal{N}, CTMDP\}, \quad (11.4.3)$$

where:

- CTMDP is a continuous-time Markov decision process (11.4.2);
- $\mathcal{N} = \{1, \dots, n\}$ is the set of players, each player is indexed by $\iota = \overline{1, n}$.

From now on, we will consider only stationary strategies $d_{k|i}^\iota(t) = d_{k|i}^\iota$. For each strategy $d_{k|i}^\iota$ the associated transition rate matrix is defined as:

$$Q^\iota(d^\iota) := [q_{j|i}^\iota(d^\iota)] = \sum_{k=1}^M q_{j|ik}^\iota d_{k|i}^\iota,$$

such that on a stationary state distribution for all $d_{k|i}^\ell$ and $t \geq 0$ we have that $\Pi^{\ell*}(d) = \lim_{t \rightarrow \infty} e^{\mathcal{Q}^\ell(d^\ell)t}$ (see [19]), where $\Pi^{\ell*}(d^\ell)$ is a stationary transition controlled matrix.

11.4.1 Solution Method

Let introduce the joint strategy $c^\ell := [c_{ik}^\ell]_{i=1, \dots, N; k=1, \dots, M}$ which is a matrix with elements

$$c_{ik}^\ell = d_{k|i}^\ell P^\ell(s^\ell = s_{(i)}), \quad (11.4.4)$$

and satisfies the following restrictions:

1. Each vector from the matrix $c^\ell := [c_{ik}^\ell]$ represents a stationary mixed-strategy that belongs to the simplex

$$\mathcal{S}^{N \times M} := \left\{ c_{ik}^\ell \in \mathbb{R}^{N \times M} : c_{ik}^\ell \geq 0, \sum_{i=1}^N \sum_{k=1}^M c_{ik}^\ell = 1 \right\}.$$

2. The variable c_{ik}^ℓ satisfies the continuous time and the ergodicity constraints, and belongs to the convex, closed and bounded set defined as follows:

$$c^\ell \in C_{\text{adm}}^\ell = \left\{ h_{(j)}^\ell(c) = \sum_{i=1}^N \sum_{k=1}^M \pi_{j|ik}^\ell c_{ik}^\ell - \sum_{k=1}^M c_{(j,k)}^\ell = 0, \sum_{i=1}^N \sum_{k=1}^M q_{j|ik}^\ell c_{ik}^\ell = 0 \right\}.$$

Notice that by (11.4.4) it follows that

$$P^\ell(s^\ell = s_{(i)}) = \sum_{k=1}^M c_{ik}^\ell, \quad d_{k|i}^\ell = \frac{c_{ik}^\ell}{\sum_{k=1}^M c_{ik}^\ell}. \quad (11.4.5)$$

Considering a utility matrix $U_{(j,i,k)}^\ell$ and the transition matrix $\pi_{j|ik}^\ell$ of each player, let us define a utility function of each player that represents the behavior of each player given by

$$W_{ik}^\ell = \sum_{j=1}^N U_{(j,i,k)}^\ell \pi_{j|ik}^\ell, \quad (11.4.6)$$

so that the “average utility function” in the stationary regime can be expressed as

$$\psi^\ell(c^1, \dots, c^n) := \sum_{i=1}^N \sum_{k=1}^M W_{ik}^\ell \prod_{\ell=1}^n c_{ik}^\ell. \quad (11.4.7)$$

Let us consider a game with players' strategies denoted by $x^\iota \in X^\iota$ ($\iota = \overline{1, n}$) where $X := \bigotimes_{\iota=1}^n X^\iota$ is a convex and compact set,

$$x^\iota := \text{col}(c^\iota), \quad X^\iota := C_{\text{adm}}^\iota,$$

where col is the column operator.

Denote by $x = (x^1, \dots, x^n)^\top \in X$, the joint strategy of the players and $x^{\hat{\iota}}$ is a strategy of the rest of the players adjoint to x^ι , namely,

$$x^{\hat{\iota}} := (x^1, \dots, x^{\iota-1}, x^{\iota+1}, \dots, x^n)^\top \in X^{\hat{\iota}} := \bigotimes_{h=1, h \neq \iota}^n X^h,$$

such that $x = (x^\iota, x^{\hat{\iota}})$, $\iota = \overline{1, n}$.

The process to solve the non-cooperative bargaining game consists of two main steps: firstly to find the initial point of the negotiation (an ideal agreement that players can reach if they negotiate cooperatively, this point is the Pareto optimal solution of the bargaining game), the formulation and solution for this problem is called the strong Nash equilibrium (for the complete formulation, solution and convergence analysis see [47, 48]); finally, for the solution of the non-cooperative bargaining process we follow the model presented in Sect. 11.3.

11.4.2 The Pareto Optimal Solution of the Bargaining Problem

The Pareto set can be defined as [16, 17]

$$\mathcal{P} := \left\{ x^*(\lambda) := \arg \max_{x \in X} \left[\sum_{\iota=1}^n \lambda^\iota \psi^\iota(x) \right], \lambda \in \mathcal{S}^n \right\}, \quad (11.4.8)$$

such that

$$\mathcal{S}^n := \left\{ \lambda \in \mathbb{R}^n : \lambda \in [0, 1], \sum_{\iota=1}^n \lambda^\iota = 1 \right\}$$

for

$$\psi(x^*(\lambda)) = (\psi^1(x^*(\lambda)), \psi^2(x^*(\lambda)), \dots, \psi^n(x^*(\lambda))).$$

The vector x^* is called a Pareto optimal solution for \mathcal{P} . Then, a strong Nash equilibrium is a strategy $x^* = (x^{1*}, \dots, x^{n*})$ such that

$$\psi(x^{1*}, \dots, x^{n*}) > \psi(x^{1*}, \dots, x^\iota, \dots, x^{n*})$$

for any $x^\iota \in X$, $x^\iota \neq x^{\iota*}$.

Consider that players try to reach the strong Nash equilibrium, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^\iota \in X^\iota$ and any $\iota = \overline{1, n}$

$$G_{L_p}(x(\lambda), \hat{x}(x, \lambda)) := \left[\sum_{\iota=1}^n \left| \lambda^\iota \left[\psi^\iota(x^\iota, x^{\hat{\iota}}) - \psi^\iota(\bar{x}^\iota, x^{\hat{\iota}}) \right] \right|^p \right]^{1/p}, \quad (11.4.9)$$

where $\hat{x}(x, \lambda) = (x^{\hat{1}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}$, $p \geq 1$ ([42, 43]) and \bar{x}^ι is the utopia point defined as follows,

$$\bar{x}^\iota := \arg \max_{x^\iota \in X^\iota} \psi^\iota(x^\iota, x^{\hat{\iota}}). \quad (11.4.10)$$

Here $\psi^\iota(x^\iota, x^{\hat{\iota}})$ is the cost-function of player ι which plays the strategy $x^\iota \in X^\iota$ and the rest of players the strategy $x^{\hat{\iota}} \in X^{\hat{\iota}}$. The functions $\psi^\iota(x^\iota, x^{\hat{\iota}})$, $\iota = \overline{1, n}$, are assumed to be concave in all their arguments.

Remark 11.2 The function $G_{L_p}(x(\lambda), \hat{x}(x, \lambda))$ satisfies the Nash condition

$$\psi^\iota(x^\iota, x^{\hat{\iota}}) - \psi^\iota(\bar{x}^\iota, x^{\hat{\iota}}) \leq 0$$

for any $x^\iota \in X^\iota$ and all $\iota = \overline{1, n}$

Definition 11.6 A strategy $x^* \in X$ is said to be a Strong L_p -Nash equilibrium if

$$x_{L_p}^* \in \operatorname{Arg} \max_{x \in X, \lambda \in S^n} \{G_{L_p}(x(\lambda), \hat{x}(x, \lambda))\}.$$

Remark 11.3 If $G_{L_p}(x(\lambda), \hat{x}(x, \lambda))$ is strictly concave then

$$x_{L_p}^* = \arg \max_{x \in X, \lambda \in S^n} \{G_{L_p}(x(\lambda), \hat{x}(x, \lambda))\}.$$

Applying the Lagrange principle (see, for example, [37]) we may conclude

$$x_{L_p}^* = \arg \max_{x \in X, \hat{x}(x) \in \hat{X}, \lambda \in S^n} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} \mathcal{L}_\delta(x, \hat{x}(x), \lambda, \mu, \xi, \eta), \quad (11.4.11)$$

where

$$\begin{aligned} \mathcal{L}_\delta(x, \hat{x}(x), \lambda, \mu, \xi, \eta) &:= G_{L_p, \delta}(x(\lambda), \hat{x}(x, \lambda)) - \sum_{i=1}^n \sum_{j=1}^N \mu_{(j)}^\iota h_{(j)}^\iota(x^\iota) - \\ &\sum_{i=1}^n \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^\iota q_{j|ik}^\iota x_{ik}^\iota - \sum_{i=1}^n \sum_{j=1}^N \sum_{k=1}^M \eta^\iota (x_{ik}^\iota - 1) + \frac{\delta}{2} (\|\mu\|^2 + \|\xi\|^2 + \|\eta\|^2) \end{aligned}$$

and

$$\begin{aligned} G_{L_p, \delta}(x(\lambda), \hat{x}(x, \lambda)) &= \\ \left[\sum_{i=1}^n \left| \lambda^\iota [\psi^\iota(x^\iota, x^\hat{i}) - \psi^\iota(\bar{x}^\iota, x^\hat{i})] \right|^p \right]^{1/p} - \frac{\delta}{2} (\|x\|^2 + \|\hat{x}(x)\|^2 + \|\lambda\|^2). \end{aligned}$$

In order to find the Pareto optimal solution, the relation (11.4.11) can be expressed in the proximal format (see [6]) as

$$\left. \begin{aligned} \mu_\delta^* &= \arg \min_{\mu \geq 0} \left\{ \frac{1}{2} \|\mu - \mu_\delta^*\|^2 + \gamma \mathcal{L}_o(x_\delta^*, \hat{x}_\delta^*(x), \lambda_\delta^*, \mu, \xi_\delta^*, \eta_\delta^*) \right\}, \\ \xi_\delta^* &= \arg \min_{\xi \geq 0} \left\{ \frac{1}{2} \|\xi - \xi_\delta^*\|^2 + \gamma \mathcal{L}_o(x_\delta^*, \hat{x}_\delta^*(x), \lambda_\delta^*, \mu_\delta^*, \xi, \eta_\delta^*) \right\}, \\ \eta_\delta^* &= \arg \min_{\eta \geq 0} \left\{ \frac{1}{2} \|\eta - \eta_\delta^*\|^2 + \gamma \mathcal{L}_o(x_\delta^*, \hat{x}_\delta^*(x), \lambda_\delta^*, \mu_\delta^*, \xi_\delta^*, \eta) \right\}, \\ x_\delta^* &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_\delta^*\|^2 + \gamma \mathcal{L}_o(x, \hat{x}_\delta^*(x), \lambda_\delta^*, \xi_\delta^*) \right\}, \\ \hat{x}_\delta^*(x) &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\frac{1}{2} \|\hat{x}(x) - \hat{x}_\delta^*(x)\|^2 + \gamma \mathcal{L}_o(x_\delta^*, \hat{x}(x), \lambda_\delta^*, \xi_\delta^*) \right\}, \\ \lambda_\delta^* &= \arg \max_{\lambda \in S^N} \left\{ -\frac{1}{2} \|\lambda - \lambda_\delta^*\|^2 + \gamma \mathcal{L}_o(x_\delta^*, \hat{x}_\delta^*(x), \lambda, \xi_\delta^*) \right\}, \end{aligned} \right\} \quad (11.4.12)$$

where the solutions x_δ^* , $\hat{x}_\delta^*(x)$, λ_δ^* , μ_δ^* , ξ_δ^* and η_δ^* depend on the small parameters $\delta, \gamma > 0$.

11.4.3 The Non-cooperative Bargaining Solution

In order to find the non-cooperative bargaining solution, let us define a time function that depends of the transition rates between states of each player as follows [14, 51]

$$\boxed{\tau'_{j|ik} := \begin{cases} \frac{1}{\sum\limits_{i \neq j}^N q'_{j|ik}} & \text{if } i = j \\ \frac{1}{q'_{j|ik}}, & \text{if } i \neq j. \end{cases}} \quad (11.4.13)$$

Also, let us redefined the utility function in Eq. (11.4.6) to involves the previous time function (11.4.13)

$$W_{ik}^\iota = \sum_{j=1}^N (\tau'_{j|ik})^{-1} U_{(j,i,k)}^\iota \pi_{j|ik}^\iota, \quad (11.4.14)$$

so that the average utility function in the stationary regime can be expressed as

$$\psi^\ell(x) := \sum_{i=1}^N \sum_{k=1}^M W_{ik}^\ell \prod_{t=1}^n c_{ik}^\ell. \quad (11.4.15)$$

Then, let us define the norm of the strategies x that depends on the transition time cost of each player as follows

$$\|(x - x^*)\|_\Lambda^2 = \sum_{\ell=1}^n \sum_{k=1}^M \left\| \left(x_{(k)}^\ell - x_{(k)}^{\ell*} \right) \right\|^2 = \sum_{\ell=1}^n \sum_{k=1}^M \left(x_{(k)}^\ell - x_{(k)}^{\ell*} \right)^T \Lambda_{(k)}^\ell \left(x_{(k)}^\ell - x_{(k)}^{\ell*} \right), \quad (11.4.16)$$

where

$$x_{(k)}^\ell = (c_{(1,k)}^\ell, \dots, c_{(N,k)}^\ell)^T \in \mathbb{R}^N, \quad k = \overline{1, M}$$

and

$$\Lambda_{(k)}^\ell := \frac{1}{2} \left[\tilde{\Lambda}_{(k)}^\ell + \tilde{\Lambda}_{(k)}^{\ell T} \right], \quad \tilde{\Lambda}_{(k)}^\ell := [\tau_{j|ik}^\ell], \quad \tilde{\Lambda}_{(k)}^\ell \in \mathbb{R}^{N \times N}.$$

Considering the utility function that depends on the average utility function $\psi^\ell(x)$ defined as follows

$$\begin{aligned} F^\ell(x, \mu, \xi, \eta) &:= \psi^\ell(x) - \psi^\ell(x^*) - \frac{1}{2} \sum_{\ell=1}^n \sum_{j=1}^N \mu_{(j)}^\ell h_{(j)}^\ell(x^\ell) - \\ &\quad \frac{1}{2} \sum_{\ell=1}^n \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^\ell q_{j|ik}^\ell x_{ik}^\ell - \frac{1}{2} \sum_{\ell=1}^n \sum_{i=1}^N \sum_{k=1}^M \eta^\ell (x_{ik}^\ell - 1), \end{aligned}$$

we may conclude that

$$x^* = \arg \max_{x \in X} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} F^\ell(x, \mu, \xi, \eta). \quad (11.4.17)$$

Finally we have that the player in turn has to fix the strategies according to the solution of the non-cooperative bargaining problem in proximal format defined as follows

$$\left. \begin{aligned} \mu^* &= \arg \min_{\mu \geq 0} \{ \delta^\ell \|\mu - \mu^*\|^2 + \alpha^\ell F^\ell(x^*, \mu, \xi^*, \eta^*) \}, \\ \xi^* &= \arg \min_{\xi \geq 0} \{ \delta^\ell \|\xi - \xi^*\|^2 + \alpha^\ell F^\ell(x^*, \mu^*, \xi, \eta^*) \}, \\ \eta^* &= \arg \min_{\eta \geq 0} \{ \delta^\ell \|\eta - \eta^*\|^2 + \alpha^\ell F^\ell(x^*, \mu^*, \xi^*, \eta) \}, \\ x^* &= \arg \max_{x \in X} \{ -\delta^\ell \|(x - x^*)\|_\Lambda^2 + \alpha^\ell F^\ell(x, \mu^*, \xi^*, \eta^*) \}. \end{aligned} \right\} \quad (11.4.18)$$

11.5 Numeric Simulations

11.5.1 Division of a Fix Resource

Consider a bargaining situation where two players have to reach an agreement about the partition of certain amount of money. In the process, each player has to make in turn an offer, i.e., a proposal as to how it should be divided. Considering the bargaining model 1, we have that the bargaining problem is as follows:

$$x^* = \arg \max_{x \in X} \left\{ -\delta_t^\ell t^\ell(x) \| (x - x^*) \|^2 + \alpha_t^\ell t^\ell(x) [\psi^\ell(x) - \psi^\ell(x^*)] \right\},$$

where $x = [x^1, x^2]$, $x^* = [x^{1*}, x^{2*}]$ and the vector x belongs to the simplex

$$\mathcal{S}^n := \left\{ x \in \mathbb{R}^n : x \in [0, 1], \sum_{i=1}^n x^i = 1 \right\}.$$

The utility functions for the problem described above are as follows

$$\psi^1(x^1, x^2) = x^1,$$

$$\psi^2(x^1, x^2) = x^2.$$

Then, we have the bargaining problem for each player as follows

$$x_{t+1} = \arg \max_{x \in X} \left\{ -\delta_t^\ell t^\ell(x) \| (x - x_t) \|^2 + \alpha_t^\ell t^\ell(x) [\psi^\ell(x) - \psi^\ell(x_t)] \right\}.$$

Once the player in turn makes a new offer according to equation above, the next player must decide either to accept or to reject the offer. If the player rejects the offer, then now it is his turn to calculate the strategies that benefit his utility and to make a new offer.

For this example let's consider a basic time function of the form $t = \exp(-n)$, where n is each step of the negotiation process, and both players with the same parameters values α, δ . Considering Theorem 11.4, we fix initial values $n_0 = 5$, $\alpha_0 = 0.1$ and $\delta_0 = 0.05$, and to obtain the maximal rate of convergence we take $\alpha = 2/3$ and $\delta = 1/3$, i.e., at each step of the process after $n = 5$ we compute the optimal parameters to ensure the maximal rate of convergence. Then, fixing an initial point $x_0 = [0.7, 0.3]$, that means 70% for player 1 and 30% for player 2, and solving the bargaining problem presented above we obtain the behavior of the proposals during the process, see Fig. 11.2. The utilities obtained at each step of the process are shown in Table 11.5.

For this example, the dynamics of the game is as follows. The first player to make a proposal is player 1, who keep the 100% for himself and 0% for the other player. In the next step $n = 2$, player 2 rejects the offer and makes a new one, offering 50%

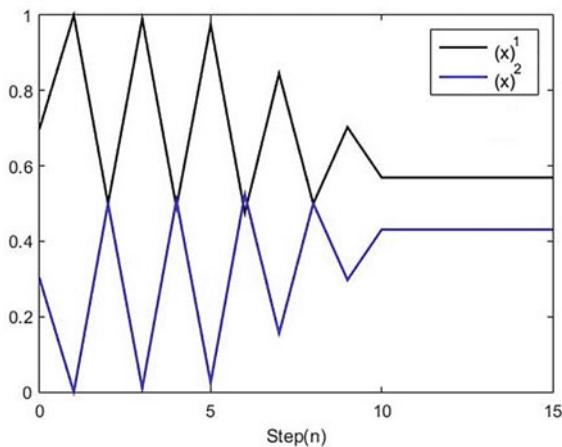


Fig. 11.2 Convergence of the utilities

Table 11.5 Utilities of each player

	x^1 (%)	x^2 (%)
	70	30
1 →	100	0
2 →	50	50
1 →	99.03	0.97
2 →	49.03	50.97
1 →	97.50	2.5
2 →	47.54	52.46
1 →	84.51	15.49
2 →	49.96	50.04
1 →	70.27	29.73
2 →	56.90	43.10
1 →	56.90	43.10

to player 1, but in the next step $n = 3$ player 1 rejects and offers 0.97% to player 2. At $n = 4$ player 2 offers 49.03% which player 1 rejects and offers 2.5% to player 2. In the next step player 2 decreases his offer to 47.54% for player 1, this offer is rejected and player 1 makes a new one, offering 15.49% to player 2. At $n = 8$ player 2 offers 49.96% which player 1 rejects and offers 29.73% to player 2. Finally at step 10 player 2 offers 56.90% to player 1, this offer is accepted and the negotiation ends.

Consider other scenarios of this same example. What happen if we set a different initial point? For example if we fix $x_0 = [0.9, 0.1]$, $x_0 = [0.5, 0.5]$ or $x_0 = [0.3, 0.7]$ we obtain the same final utilities for both players, but if the initial point is $x_0 = [0.1, 0.9]$ then the utilities are different and player 2 gets a greater utility than player 1. Now, what happen if player 2 starts the process? Basically we have

a symmetrical behavior, for some initial points like $x_0 = [0.1, 0.9]$, $x_0 = [0.3, 0.7]$, $x_0 = [0.5, 0.5]$ or $x_0 = [0.7, 0.3]$ both players obtain the same final utilities as in the previous scenario $\psi^1(x) = 43.10\%$ and $\psi^2(x) = 56.90\%$, and for the initial point $x_0 = [0.9, 0.1]$ the final utilities are different, i.e., player 1 gets a greater utility than player 2.

11.6 Extensions

We now present two extensions of the bargaining game provided above that include the case when agents have different discount factors, and another where agents might coordinate on their demands. The convergence of results follows trivially from our general analysis presented in the appendix above.

11.6.1 Bargaining Under Different Discounting

In this approach we present a solution where at each step of the negotiation process players calculate the Nash equilibrium considering the utility functions of all players but with the particularity that internally each player reaches this equilibrium point in a different time. Following the description of the model presented previously, we redefine the advantage of propose a new offer that depends on the utility function

$$f(x_t, x_{t+1}) := \sum_{\ell=1}^n [\psi^\ell(x_{t+1}) - \psi^\ell(x_t)] \geq 0$$

for all players to reject the offer x_t and making a new offer x_{t+1} given the time spent to benefit of this advantage $T(x_{t+1}) > 0$, and $\alpha^\ell(x_t)$ be the weight that players put on their advantages to reject the offer x_t . Thus, the advantages to reject the offer x_t and to propose a new offer x_{t+1} are given by $A(x_t, x_{t+1}) = \alpha(x_t)T(x_{t+1})f(x_t, x_{t+1})$.

Remark 11.4 The function $f(x_t, x_{t+1})$ satisfies the Nash condition

$$\psi^\ell(x_{t+1}) - \psi^\ell(x_t) \geq 0$$

for any $x_{t+1} \in X$ and all players.

Definition 11.7 A strategy $x^* \in X$ is said to be a Nash equilibrium if

$$x^* \in \operatorname{Argmax}_{x_t \in X} \{f(x_t, x_{t+1})\}$$

for any $x_{t+1} \in X$ and all players.

Then, at each step of the bargaining game we have in proximal format that the players must select their strategies according to

$$\boxed{x^* = \arg \max_{x \in X} \left\{ -\delta_t T(x) \| (x - x^*) \|^2 + \alpha_t T(x) f(x, x^*) \right\},} \quad (11.6.1)$$

where

$$f(x, x^*) := \sum_{\iota=1}^n [\psi^\iota(x) - \psi^\iota(x^*)].$$

At each step of the bargaining process, players calculate simultaneously the Nash equilibrium but considering that each player reach the equilibrium in a different time.

11.6.1.1 Markov Chains Interpretation

Let us to define the Nash equilibrium as a strategy $x^* = (x^{1*}, \dots, x^n)$ such that

$$\psi(x^{1*}, \dots, x^{n*}) \geq \psi(x^{1*}, \dots, x^\iota, \dots, x^{n*})$$

for any $x^\iota \in X$.

Consider that players try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^\iota \in X^\iota$ and any $\iota = \overline{1, n}$

$$f(x, \hat{x}(x)) := \sum_{\iota=1}^n [\psi^\iota(x^\iota, x^{\hat{\iota}}) - \psi^\iota(\bar{x}^\iota, x^{\hat{\iota}})] \quad (11.6.2)$$

where $\hat{x} = (x^{\hat{1}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}$ [42, 43], \bar{x}^ι is the utopia point defined as Eq. (11.4.10) and $\psi^\iota(x^\iota, x^{\hat{\iota}})$ is the concave cost-function of player ι which plays the strategy $x^\iota \in X^\iota$ and the rest of players the strategy $x^{\hat{\iota}} \in X^{\hat{\iota}}$ defined as Eq. (11.4.15) considering the time function.

Remark 11.5 The property $f(x, \hat{x}(x))$ less or equal to 0 is equivalent to the Nash condition

$$\psi^\iota(x^\iota, x^{\hat{\iota}}) - \psi^\iota(\bar{x}^\iota, x^{\hat{\iota}}) \leq 0 \quad (11.6.3)$$

for any $x^\iota \in X^\iota$ and all $\iota = \overline{1, n}$.

Definition 11.8 A strategy $x^* \in X$ is said to be a Nash equilibrium if

$$x^* \in \text{Arg} \max_{x \in X_{\text{adm}}} \{f(x, \hat{x}(x))\}.$$

Remark 11.6 If $f(x, \hat{x}(x))$ is strictly concave then

$$x^* = \arg \max_{x \in X_{\text{adm}}} \{f(x, \hat{x}(x))\}.$$

We redefine the utility function that depends of the average utility function of all players as follows

$$\begin{aligned} F(x, \hat{x}(x)) &:= f(x, \hat{x}(x)) - \frac{1}{2} \sum_{\iota=1}^n \sum_{j=1}^N \mu_{(j)}^\iota h_{(j)}^\iota(x^\iota) - \\ &\quad \frac{1}{2} \sum_{\iota=1}^n \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^M \xi_{(j)}^\iota q_{j|ik}^\iota x_{ik}^\iota - \frac{1}{2} \sum_{\iota=1}^n \sum_{i=1}^N \sum_{k=1}^M \eta^\iota (x_{ik}^\iota - 1), \end{aligned}$$

then, we may conclude that

$$x^* = \arg \max_{x \in X, \hat{x} \in \hat{X}} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} F(x, \hat{x}(x), \mu, \xi, \eta). \quad (11.6.4)$$

Finally we have that at each step of the bargaining process, players calculate the Nash equilibrium (but they reach the equilibrium at different time) according to the solution of the non-cooperative bargaining problem in proximal format defined as follows

$$\left. \begin{aligned} \mu^* &= \arg \min_{\mu \geq 0} \{-\delta \|\mu - \mu^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu, \xi^*, \eta^*)\}, \\ \xi^* &= \arg \min_{\xi \geq 0} \{-\delta \|\xi - \xi^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu^*, \xi, \eta^*)\}, \\ \eta^* &= \arg \min_{\eta \geq 0} \{-\delta \|\eta - \eta^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu^*, \xi^*, \eta)\}, \\ x^* &= \arg \max_{x \in X} \{-\delta \|x - x^*\|_\Lambda^2 + \alpha F(x, \hat{x}^*(x), \mu^*, \xi^*, \eta^*)\}, \\ \hat{x}^* &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\delta \|\hat{x} - \hat{x}^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}(x), \mu^*, \xi^*, \eta^*) \right\}. \end{aligned} \right\} \quad (11.6.5)$$

11.6.2 Bargaining with Collusive Behavior

In this approach we analyze a bargaining situation where players make groups and alternately each group makes an offer to the others until they reach an equilibrium point (agreement). We describe a bargaining model with two teams of players as follows. Let us consider a bargaining game with $n + m$ players. Let $\mathcal{N} = \{1, \dots, n\}$ denote the set of players called team A and let's define the behavior of all players $\iota = \overline{1, n}$ as $x_\iota = (x_\iota^1, \dots, x_\iota^n) \in X$, where X is a convex and compact set. In the same way, the rest $\mathcal{M} = \{1, \dots, m\}$ players are the team B and let the set of the strategy profiles of all player $m = \overline{1, m}$ be defined by $y_t = (y_t^1, \dots, y_t^m) \in Y$, where Y is a convex and compact set. Then, $X \times Y$ in the set of full strategy profiles. In this model the function $\psi(x, y)$ represents the utility function of team A which determines the

decision of accept or reject the offer; similarly, team B makes the decision according to its utility function $\varphi(x, y)$.

Following the description of the model presented above, we redefine the advantage of propose a new offer considering the utility function for team A as follows

$$f(x_t, y_t, x_{t+1}, y_{t+1}) := \sum_{\iota=1}^n [\psi^\iota(x_{t+1}, y_t) - \psi^\iota(x_t, y_t)] \geq 0,$$

and, similarly the utility function for team B is as follows

$$g(x_t, y_t, x_{t+1}, y_{t+1}) := \sum_{m=1}^m [\varphi^\iota(x_t, y_{t+1}) - \varphi^\iota(x_t, y_t)] \geq 0.$$

Thus, the advantages for team A to reject the offer x_t and to propose a new offer x_{t+1} are given by $A(x_t, y_t, x_{t+1}, y_{t+1}) = \alpha(x_t)T(x_{t+1})f(x_t, y_t, x_{t+1}, y_{t+1})$; in the same way, the advantages for team B to reject the offer y_t and to propose a new offer y_{t+1} are given by

$$A(x_t, y_t, x_{t+1}, y_{t+1}) = \alpha(y_t)T(y_{t+1})g(x_t, y_t, x_{t+1}, y_{t+1}).$$

Remark 11.7 The nonpositivity of the function $f(x_t, y_t, x_{t+1}, y_{t+1})$ is equivalent to the Nash condition

$$\psi^\iota(x_{t+1}, y_t) - \psi^\iota(x_t, y_t) \geq 0$$

for any $x \in X$, $y \in Y$ and $\iota = \overline{1, n}$ players.

Remark 11.8 The nonpositivity of the function $g(x_t, y_t, x_{t+1}, y_{t+1})$ is also equivalent to the Nash condition

$$\varphi^\iota(x_t, y_{t+1}) - \varphi^\iota(x_t, y_t) \geq 0$$

for any $x \in X$, $y \in Y$ and $m = \overline{1, m}$ players.

The dynamics of the bargaining game is as follows: at each step of the negotiation process the team A chooses a strategy $x \in X$ considering the utility function $f(x_t, y_t, x_{t+1}, y_{t+1})$, then team B must decide between to accept or reject the offer calculating a new offer (strategies) $y \in Y$ considering the utility function of the group $g(x_t, y_t, x_{t+1}, y_{t+1})$. Following the description of the model 1, now we have that teams solve the problem in proximal format as follows:

$$\boxed{\begin{aligned} x^* &= \arg \max_{x \in X} \left\{ -\delta_t T(x) \| (x - x^*) \|^2 + \alpha_t T(x) f(x, y, x^*, y^*) \right\}, \\ y^* &= \arg \max_{y \in Y} \left\{ -\delta_t T(y) \| (y - y^*) \|^2 + \alpha_t T(y) g(x, y, x^*, y^*) \right\}, \end{aligned}} \quad (11.6.6)$$

where

$$f(x, y, x^*, y^*) := \sum_{\iota=1}^n [\psi^\iota(x, y^*) - \psi^\iota(x^*, y^*)],$$

$$g(x, y, x^*, y^*) := \sum_{m=1}^m [\varphi^m(x^*, y) - \varphi^m(x^*, y^*)].$$

At each step, teams make a new offer according to equation (11.6.6), both teams solve the bargaining problem together but they reach the equilibrium at different time, the bargaining game continues until the offers (strategies) of all player show convergence.

11.6.2.1 Markov Chains Interpretation

For this model, in the same way that we define the strategies $x \in X$, let us consider a set of strategies denoted by $y^m \in Y^m$ ($m = \overline{1, m}$) where $Y := \bigotimes_{m=1}^m Y^m$ is a convex and compact set,

$$y^m := \text{col}(c^m), \quad Y^m := C_{\text{adm}}^m,$$

where *col* is the column operator.

Denote by $y = (y^1, \dots, y^m)^\top \in Y$, the joint strategy of the players and \hat{y}^m is a strategy of the rest of the players adjoint to y^m , namely,

$$\hat{y}^m := (y^1, \dots, y^{m-1}, y^{m+1}, \dots, y^m)^\top \in Y^{\hat{m}} := \bigotimes_{h=1, h \neq m}^m Y^h,$$

such that $y = (y^m, \hat{y}^m)$, $m = \overline{1, m}$.

Consider that players of team A try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^\iota \in X^\iota$ and any $\iota = \overline{1, n}$

$$f(x, \hat{x}(x)|y) := \sum_{\iota=1}^n \left[\psi^\iota(x^\iota, x^{\hat{\iota}}|y) - \psi^\iota(\bar{x}^\iota, x^{\hat{\iota}}|y) \right] \leq 0, \quad (11.6.7)$$

where $\hat{x} = (x^{\hat{1}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}$ [42, 43], \bar{x}^ι is the utopia point defined as Eq. (11.4.10) and $\psi^\iota(x^\iota, x^{\hat{\iota}}|y)$ is the concave cost-function of player ι which plays

the strategy $x' \in X'$ and the rest of players the strategy $x^i \in X^i$ fixing the strategies $y \in Y$ of team B, and it is defined as Eq. (11.4.15) considering the time function.

Similarly, consider that players of team B also try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $y^* = (y^{1*}, \dots, y^{m*}) \in Y$ satisfying for any admissible $y^m \in Y^m$ and any $m = \overline{1, m}$

$$g(y, \hat{y}(y)|x) := \sum_{m=1}^m \left[\psi^m \left(y^m, y^{\hat{m}} | x \right) - \psi^m \left(\bar{y}^m, y^{\hat{m}} | x \right) \right] \leq 0, \quad (11.6.8)$$

where $\hat{y} = (y^{\hat{1}\top}, \dots, y^{\hat{m}\top})^\top \in \hat{Y} \subseteq \mathbb{R}^{m(m-1)}$, \bar{y}^m is the utopia point defined as Eq. (11.4.10) and $\psi^m \left(y^m, y^{\hat{m}} | x \right)$ is the concave cost-function of player m which plays the strategy $y^m \in Y^m$ and the rest of players the strategy $y^{\hat{m}} \in Y^{\hat{m}}$ fixing the strategies $x \in X$ of team A, and it is defined as Eq. (11.4.15) considering the time function.

Then, we have that a strategy $x^* \in X$ of team A together with the collection $y^* \in Y$ of team B are defined as the equilibrium of a strictly concave bargaining problem if

$$(x^*, y^*) = \arg \max_{x \in X_{\text{adm}}, y \in Y_{\text{adm}}} \{ f(x, \hat{x}(x)|y) \leq 0, g(y, \hat{y}(y)|x) \leq 0 \}$$

We redefine the utility function that depends of the average utility function of all players as follows

$$\begin{aligned} F(x, \hat{x}(x), y, \hat{y}(y)) &:= f(x, \hat{x}(x)|y) + g(y, \hat{y}(y)|x) - \frac{1}{2} \sum_{\iota=1}^n \sum_{j=1}^N \mu_{(j)}^\iota h_{(j)}^\iota(x^\iota) - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{j=1}^N \mu_{(j)}^m h_{(j)}^m(y^m) - \frac{1}{2} \sum_{\iota=1}^n \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^\iota q_{j|ik}^\iota x_{ik}^\iota - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^m q_{j|ik}^m y_{ik}^m - \frac{1}{2} \sum_{\iota=1}^n \sum_{i=1}^N \sum_{k=1}^M \eta^\iota (x_{ik}^\iota - 1) - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{i=1}^N \sum_{k=1}^M \eta^m (y_{ik}^m - 1), \end{aligned}$$

then, we may conclude that

$$(x^*, y^*) = \arg \max_{x \in X, \hat{x} \in \hat{X}, y \in Y, \hat{y} \in \hat{Y}} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} F(x, \hat{x}(x), y, \hat{y}(y), \mu, \xi, \eta). \quad (11.6.9)$$

Finally we have that at each step of the bargaining process, players calculate their equilibrium according to the solution of the non-cooperative bargaining problem in proximal format defined as follows

$$\left. \begin{array}{l} \mu^* = \arg \min_{\mu \geq 0} \left\{ -\delta \|\mu - \mu^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu, \xi^*, \eta^*) \right\} \\ \xi^* = \arg \min_{\xi \geq 0} \left\{ -\delta \|\xi - \xi^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi, \eta^*) \right\} \\ \eta^* = \arg \min_{\eta \geq 0} \left\{ -\delta \|\eta - \eta^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta) \right\} \\ x^* = \arg \max_{x \in X} \left\{ -\delta \|x - x^*\|_\Lambda^2 + \alpha F(x, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\} \\ \hat{x}^* = \arg \max_{\hat{x} \in \hat{X}} \left\{ -\delta \|\hat{x} - \hat{x}^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\} \\ y^* = \arg \max_{y \in Y} \left\{ -\delta \|y - y^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}^*(x), y, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\} \\ \hat{y}^* = \arg \max_{\hat{y} \in \hat{Y}} \left\{ -\delta \|\hat{y} - \hat{y}^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}(y), \mu^*, \xi^*, \eta^*) \right\} \end{array} \right\} \quad (11.6.10)$$

11.7 Appendix: Proofs

11.7.1 The Non-cooperative Bargaining Game

In this section, we present a general version of the model presented in Sect. 11.3. Consider the game theory problem of a concave twice-differentiable real-valued function ψ defined on X , which is a compact and convex subset of \mathbb{R}^N

$$\max_{x \in X} \psi(x).$$

Following the proximal point algorithm for solving game theory problems presented by [5], the unique solution is a sequence (x_t) $y \in \mathbb{N}$ with a initial value $x_0 \in X$,

$$\max_{x \in X} [\psi(x) - \delta_t \|x - x_t\|^2], \quad (11.7.1)$$

where $\delta_t > 0$, $\delta_t \downarrow 0$ and the term $\|x - x_t\|^2$ ensures that the objective function (11.7.1) is strictly positive definite and that some iterative method presents convergence [44, 45]. The result obtained is not affected by the quadratic term for $\delta_t > 0$ and $\delta_t \downarrow 0$.

The bargaining game model considered in this paper involves game theory problems with an additional penalization, a time cost related with the time spent for each player to move from one position to another one [7, 8, 28], i.e., to decide either to accept an offer or to reject it and choose another.

The Bargaining Model

In this approach, we consider the model presented by [38], and we provide a solution to a bargaining situation where players are individual-rational and alternately

make offers and counteroffers thinking only of their own interests, i.e., they compute independently the strategies that maximize only their own utility.

In general terms, the dynamic of the multilateral non-cooperative bargaining game is as follows. The game consists of a set $\mathcal{N} = \{1, \dots, n\}$ of players bargaining a certain transaction according to the alternating-offers procedure. Define the behavior of each player $\iota = \overline{1, n}$ as a sequence $x_t^\iota \in X^\iota$, $n \in \mathbb{N}$, where X^ι is the decision space (strategies) of each player. Then, we can define the strategies set of all players as $x_t = (x_t^1, \dots, x_t^n) \in X$ where X is a convex and compact set. Players take turns to analyze and present their position in the negotiation process, i.e., at each step n the player ι in turn must decide between to stay in the same strategy $x_{t+1} = x_t$, that is that player ι accepts the offer, or to choose a new strategy $x_{t+1} \neq x_t$, that means that player rejects the offer and makes a new one. The function $\psi^\iota(x)$ represents the utility function of each player which determines the decision of to accept or to reject the offer.

At turn $n = 0$, the first player to make an offer chooses a strategy set x_0 considering the utility function $\psi^\iota(x)$, then, the rest of the players must decided either to accept the offer and finish the game or to reject it and continue with the process, in this case, at step $n = 1$ the next player makes a counteroffer by choosing a strategy set x_1 that benefits him more or in equal measure than the offer proposed by the first player according to his utility function, if this counteroffer is accepted then agreement is struck, otherwise, the player in turn makes a new offer at step $n = 2$, and the process continues.

The time cost between offers is defined for each player as a function $\Lambda^\iota : X \times X \rightarrow \mathbb{R}$ which can be interpreted as a distance function of each player where $\Lambda^\iota(x_t, x_{t+1}) = \kappa^\iota(x_t, x_{t+1})$, we have that $\kappa^\iota(x_t, x_{t+1}) = 0$ if $x_{t+1} = x_t$ (accepts the offer) or $\kappa^\iota(x_t, x_{t+1}) > 0$ if $x_t \neq x_{t+1}$ (rejects and makes a new one). In general, the time cost function can be reexpressed as $\Lambda^\iota(x_t, x_{t+1}) := T^\iota(x_t, x_{t+1})\kappa^\iota(x_t, x_{t+1})$ where $T^\iota(x_t, x_{t+1}) \geq 0$ is the time spent for each player to reject an offer x_t and to make a new one x_{t+1} and $\kappa^\iota(x_t, x_{t+1})$ is the offer cost function associated to each player.

In the simplest case, each player makes a new offer trying to obtain the highest possible payoff according to the utility function, $\psi^\iota(x_{t+1}) - \psi^\iota(x_t) \geq 0$ given the time spent $T^\iota(x_{t+1}) > 0$ to analyze the advantage of to reject the offer x_t and make a new offer x_{t+1} , and $\alpha^\iota(x_t)$ be the weight that players put on their advantages of to reject the offer x_t . Thus, the advantages of to reject the offer x_t and to propose a new offer x_{t+1} are given by $A^\iota(x_t, x_{t+1}) = \alpha^\iota(x_t)T^\iota(x_{t+1})[\psi^\iota(x_{t+1}) - \psi^\iota(x_t)]$.

The dynamics of the bargaining game with alternating-offers considering the time cost is as follows. At each step $n \in \mathbb{N}$, the player in turn considers to reject the offer x_t and propose a new offer x_{t+1} . For each player, to make a new proposal is acceptable if the advantages $A^\iota(x_t, x_{t+1})$ are determined by $\delta^\iota(x_t) \in [0, 1]$ (degree of acceptability) of the time cost $\Lambda^\iota(x_t, x_{t+1})$. Then, the set of strategies that maximizes the utility of each player is defined by

$$F^\iota(x_t) = \{x_{t+1} \in X : \alpha^\iota(x_t)T^\iota(x_{t+1})[\psi^\iota(x_{t+1}) - \psi^\iota(x_t)] \geq \delta^\iota(x_t)T^\iota(x_{t+1})\kappa^\iota(x_t, x_{t+1})\}.$$

We define a utility function $\psi^\ell : X \rightarrow \mathbb{R}$ such that the impact of experience on cost is constant and limited to the most recent element x_t on the trajectory (x_t) . In addition, the advantages to change $A^\ell(x_t, x_{t+1})$ are determined by the degree of acceptability $\delta_t^\ell(x_t) \in [0, 1]$ of the costs to move $\Lambda^\ell(x_t, x_{t+1})$.

Thus, the acceptance criterion to propose a new offer satisfies the condition

$$\alpha_t^\ell(x_t) T^\ell(x_{t+1}) [\psi^\ell(x_{t+1}) - \psi^\ell(x_t)] \geq \delta_t^\ell(x_t) T^\ell(x_{t+1}) \kappa^\ell(x_t, x_{t+1}).$$

This algorithms are naturally linked with several classical proximal algorithms given in Eq. (11.7.1). That is, by taking $\delta_t^\ell T^\ell(x) \kappa^\ell(x^*, x) = \delta_t^\ell T^\ell(x) \| (x - x^*) \|^2$ and $A^\ell(x, x^*) := \alpha_t^\ell t^\ell(x) [\psi^\ell(x) - \psi^\ell(x^*)]$, the point x^* solves the maximization problem if remains a fixed point of the proximal mapping, that is,

$$x^* = \arg \max_{x \in X} \left\{ -\delta_t^\ell T^\ell(x) \| (x - x^*) \|^2 + \alpha_t^\ell T^\ell(x) f(x, x^*) \right\}, \quad (11.7.2)$$

where

$$f(x, x^*) := \psi^\ell(x) - \psi^\ell(x^*).$$

Once the player in turn makes a new offer according to equation (11.7.2), the next player must decide either to accept or to reject the offer. If the player rejects the offer, then now it is his turn to calculate the strategies that benefit his utility and to make a new offer. This process continues until an agreement is reached, i.e. the proposals (strategies) of the players do not change (convergence).

11.7.2 Formulation of the Problem

Consider the following constrained programming problem

$$\left. \begin{array}{l} \max_{x \in X_{\text{adm}}} f(x, x_t), \\ X_{\text{adm}} := \{x \in \mathbb{R}^n : x \geq 0, A_0 x = b_0 \in \mathbb{R}^{M_0}, A_1 x \leq b_1 \in \mathbb{R}^{M_1}\}, \end{array} \right\} \quad (11.7.3)$$

where X_{adm} is a bounded set. Introducing the vector $u \in \mathbb{R}^{M_1}$ with components $u_i \geq 0$ for all $i = 1, \dots, M_1$, the original problem (11.7.3) can be rewritten as

$$\left. \begin{array}{l} \max_{x \in X_{\text{adm}}, u \geq 0} f(x, x_t), \\ X_{\text{adm}} := \{x \in \mathbb{R}^n : x \geq 0, A_0 x = b_0, A_1 x - b_1 + u = 0\}, \end{array} \right\} \quad (11.7.4)$$

Notice that this problem may have non-unique solution and $\det(A_0^\top A_0) = 0$. Define by $X^* \subseteq X_{\text{adm}}$ the set of all solutions of the problem (11.7.4) and consider the objective function

$$\begin{aligned} \mathbb{P}_{\alpha,\delta}(x, u|x_t) := & -\frac{\delta}{2} T(x) \|x - x_t\|^2 + \alpha T(x) f(x, x_t) - \\ & \frac{1}{2} \|A_0 x - b_0\|^2 - \frac{1}{2} \|A_1 x - b_1 + u\|^2 - \frac{\delta}{2} \|u\|^2, \end{aligned} \quad (11.7.5)$$

where the parameters α, δ . Then, the game theory problem is as follows

$$\max_{x \in X_{\text{adm}}, u \geq 0} \mathbb{P}_{\alpha,\delta}(x, u|x_t). \quad (11.7.6)$$

11.7.3 Convergence Analysis

The game consists of a set $\mathcal{N} = \{1, \dots, n\}$ of players. Let $x^\ell \in X^\ell$ be the strategy of each player $\ell = \overline{1, n}$ where X^ℓ is the decision space (strategies) of each player. Then, we can define the strategies set of all players as

$$x = (x^1, \dots, x^n) \in X, \quad X := \bigotimes_{\ell=1}^n X^\ell,$$

where X is a convex and compact set. Then, in order to prove Theorem 11.1, let's present the following lemma and its proof.

Lemma 11.1 *The bounded set X^* of all solutions of the original game theory problem (11.7.4) is not empty and the Slater's condition holds, that is, there exists a point $\hat{x} \in X_{\text{adm}}$ such that*

$$A_1 \hat{x} < b_1. \quad (11.7.7)$$

Moreover, the parameters α and δ are time-varying, i.e.,

$$\alpha = \alpha_t, \quad \delta = \delta_t \quad (n = 0, 1, 2, \dots),$$

such that

$$0 < \alpha_t \downarrow 0, \quad \frac{\alpha_t}{\delta_t} \downarrow 0 \quad \text{when } n \rightarrow \infty. \quad (11.7.8)$$

Then

$$\begin{aligned} x_t^* &:= x^*(\alpha_t, \delta_t) \xrightarrow{n \rightarrow \infty} x^{**}, \\ u_t^* &:= u^*(\alpha_t, \delta_t) \xrightarrow{n \rightarrow \infty} u^{**}, \end{aligned}$$

where $x^{**} \in X^*$ and $u^{**} \in \mathbb{R}^{M_1}$ define the solution of the original problem (11.7.4) with the minimal weighted norm,

$$\|x^{**}\|^2 + \|u^{**}\|^2 \leq \|x^*\|^2 + \|u^*\|^2,$$

for all $x^* \in X^*$, $u^* \in \mathbb{R}^{M_1}$ and

$$u^{**} = b_1 - A_1 x^{**}.$$

Proof 1. First, let us prove that the Hessian matrix \mathbb{H} associated with the objective function (11.7.5) is strictly negative definite for any positive α and δ , to show that the objective function (11.7.5) is strictly concave. If the set of solutions of problem (11.7.4) is non-empty then the objective function (11.7.5) is strictly concave.

It should be proven that for all $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^{M_1}$

$$\mathbb{H} = \begin{bmatrix} \frac{\partial^2}{\partial x^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) & \frac{\partial^2}{\partial u \partial x} \mathbb{P}_{\alpha,\delta}(x, u | x_t) \\ \frac{\partial^2}{\partial x \partial u} \mathbb{P}_{\alpha,\delta}(x, u | x_t) & \frac{\partial^2}{\partial u^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) \end{bmatrix} < 0,$$

Employing Schur's lemma [36] it is necessary and sufficient to prove that

$$\begin{aligned} 1. \quad & \frac{\partial^2}{\partial x^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) < 0, & 2. \quad & \frac{\partial^2}{\partial u^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) < 0, \\ 3. \quad & \frac{\partial^2}{\partial x^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) < \frac{\partial^2}{\partial u \partial x} \mathbb{P}_{\alpha,\delta}(x, u | x_t) \left[\frac{\partial^2}{\partial u^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) \right]^{-1} \frac{\partial^2}{\partial x \partial u} \mathbb{P}_{\alpha,\delta}(x, u | x_t). \end{aligned}$$

Applying the Schur's lemma over the objective function (11.7.5) it follows for condition 1

$$\begin{aligned} \frac{\partial^2}{\partial x^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) &= -\delta T(x_t) I_{n \times n} + \alpha T(x_t) \frac{\partial^2}{\partial x^2} f(x, x_t) - A_0^\top A_0 - A_1^\top A_1 \leq \\ &\alpha T(x_t) \frac{\partial^2}{\partial x^2} f(x, x_t) - \delta T(x_t) I_{n \times n} \leq \delta T(x_t) \left(\frac{\alpha}{\delta} \lambda^+ - 1 \right) I_{n \times n} < 0, \end{aligned}$$

for all $\delta > 0$, where

$$\lambda^+ := \max_{x \in X_{\text{adm}}} \left[\lambda_{\max} \left(\frac{\partial^2}{\partial x^2} f(x, x_t) \right) \right] < 0.$$

Then, for condition 2 we have

$$\frac{\partial^2}{\partial u^2} \mathbb{P}_{\alpha,\delta}(x, u | x_t) = -(1 + \delta) I_{M_1 \times M_1} < 0.$$

By condition 3, it is necessary to satisfy that

$$\begin{aligned} \frac{\partial^2}{\partial x^2} \mathbb{P}_{\alpha, \delta}(x, u|x_t) &= -\delta T(x_t) I_{n \times n} + \alpha T(x_t) \frac{\partial^2}{\partial x^2} f(x, x_t) - A_0^\top A_0 - A_1^\top A_1 < \\ \frac{\partial^2}{\partial u \partial x} \mathbb{P}_{\alpha, \delta}(x, u|x_t) \left[\frac{\partial^2}{\partial u^2} \mathbb{P}_{\alpha, \delta}(x, u|x_t) \right]^{-1} \frac{\partial^2}{\partial x \partial u} \mathbb{P}_{\alpha, \delta}(x, u|x_t) &= -(1+\delta)^{-1} A_1^\top A_1, \end{aligned}$$

or equivalently,

$$\alpha t(x_t) \frac{\partial^2}{\partial x^2} f(x, x_t) - \delta T(x_t) I_{n \times n} - A_0^\top A_0 - \frac{\delta}{1+\delta} A_1^\top A_1 < 0,$$

which holds for any $\delta > 0$ having

$$\begin{aligned} T(x_t) (\alpha \lambda^+ - \delta) I_{n \times n} - A_0^\top A_0 - \frac{\delta}{1+\delta} A_1^\top A_1 &\leq \\ \delta T(x_t) \left(\frac{\alpha}{\delta} \lambda^+ - 1 \right) I_{n \times n} &= \delta T(x_t) (o(1) - 1) I_{n \times n} < 0. \end{aligned}$$

As a result, the Hessian is $\mathbb{H} < 0$ which means that proximal function (11.7.5) is strictly concave and, hence, has a unique maximal point defined below as $x^*(\alpha, \delta)$ and $u^*(\alpha, \delta)$.

2. If the proximal function (11.7.5) is strictly concave then the sequence $\{x_t\}$ of the proximal function (11.7.5) converges when $n \rightarrow \infty$, i.e. the proximal function has a maximal point defined by $x^*(\alpha, \delta)$ and $u^*(\alpha, \delta)$.

Following the strictly concavity property (Theorem 11.1) for any $y := \begin{pmatrix} x \\ u \end{pmatrix}$ and any vector $y_t^* := \begin{pmatrix} x_t^* = x^*(\alpha_t, \delta_t) \\ u_t^* = u^*(\alpha_t, \delta_t) \end{pmatrix}$ for the function $\mathbb{P}_{\alpha, \delta}(x, u|x_t) = \mathbb{P}_{\alpha, \delta}(y|x_t)$ we have

$$\begin{aligned} 0 &\leq (y_t^* - y)^\top \frac{\partial}{\partial y} \mathbb{P}_{\alpha_t, \delta_t}(y_t^*|x_t) \\ &= (x_t^* - x)^\top \frac{\partial}{\partial x} \mathbb{P}_{\alpha_t, \delta_t}(x_t^*, u_t^*|x_t) + (u_t^* - u)^\top \frac{\partial}{\partial u} \mathbb{P}_{\alpha_t, \delta_t}(x_t^*, u_t^*|x_t) \\ &= (x_t^* - x)^\top \left(-\delta_t T(x_t)(x_t^* - x_t) + \alpha_t T(x_t) \frac{\partial}{\partial x} f(x_t^*, x_t) - A_0^\top [A_0 x_t^* - b_0] \right. \\ &\quad \left. - A_1^\top [A_1 x_t^* - b_1 + u_t^*] \right) + (u_t^* - u)^\top (-A_1 x_t^* + b_1 - (1 + \delta_t) u_t^*) \\ &= \alpha_t T(x_t) (x_t^* - x)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - [A_0 (x_t^* - x)]^\top [A_0 x_t^* - b_0] \\ &\quad - [A_1 (x_t^* - x)]^\top [A_1 x_t^* - b_1 + u_t^*] - \delta_t T(x_t) (x_t^* - x)^\top (x_t^* - x_t) \\ &\quad - (u_t^* - u)^\top [A_1 x_t^* - b_1 + (1 + \delta_t) u_t^*]. \end{aligned} \tag{11.7.9}$$

Now, selecting $x := x^* \in X^*$ (x^* is one of the admissible solutions such that $A_0x^* = b_0$ and $A_1x^* = b_1 - u^*$) and $u := (1 + \delta_t)^{-1} (b_1 - A_1x_t^*)$ we obtain

$$\begin{aligned} 0 &\leq \alpha_t T(x_t) (x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - [A_0(x_t^* - x^*)]^\top [A_0x_t^* - b_0] - \\ &\quad [A_1(x_t^* - x^*)]^\top [A_1x_t^* - b_1 + u_t^*] - \delta_t T(x_t) (x_t^* - x^*)^\top (x_t^* - x_t) - \\ &\quad (1 + \delta_t)^{-1} [u_t^*(1 + \delta_t) - b_1 + A_1x_t^*]^\top [A_1x_t^* - b_1 + (1 + \delta_t)u_t^*] - \\ &\quad \delta_t (u_t^* - b_1 - A_1x_t^*)^\top u_t^*. \end{aligned} \tag{11.7.10}$$

Simplifying Eq. (11.7.10) we have

$$\begin{aligned} 0 &\leq \alpha_t T(x_t) (x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - \|A_0(x_t^* - x^*)\|^2 - \|A_1(x_t^* - x^*)\|^2 - \\ &\quad \delta_t T(x_t) (x_t^* - x_t)^\top (x_t^* - x_t) - (1 + \delta_t)^{-1} \|A_1x_t^* - b_1 + u_t^*(1 + \delta_t)\|^2 - \\ &\quad \delta_t (u_t^* - b_1 - A_1x_t^*)^\top u_t^*. \end{aligned}$$

Dividing both sides of this inequality by δ_t we obtain

$$\begin{aligned} 0 &\leq \frac{\alpha_t}{\delta_t} T(x_t) (x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{1}{\delta_t} \|A_0(x_t^* - x^*)\|^2 - \\ &\quad \frac{1}{\delta_t} \|A_1(x_t^* - x^*)\|^2 - \frac{1}{\delta_t} (1 + \delta_t)^{-1} \|A_1x_t^* - b_1 + u_t^*(1 + \delta_t)\|^2 - \\ &\quad T(x_t) (x_t^* - x_t)^\top (x_t^* - x_t) - (u_t^* - b_1 - A_1x_t^*)^\top u_t^*. \end{aligned} \tag{11.7.11}$$

Now, taking $x = x_t^*$ and $u = 0$ from Eq. (11.7.9) one has

$$\begin{aligned} 0 &\leq -(u_t^*)^\top [A_1x_t^* - b_1 + (1 + \delta_t)u_t^*] \\ &= -(u_t^*)^\top (A_1x_t^* - b_1) - (1 + \delta_t) \|u_t^*\|^2 \\ &= - \left(\left\| \sqrt{1 + \delta_t} u_t^* \right\|^2 + 2 \left(\sqrt{1 + \delta_t} u_t^* \right)^\top \left[\frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right] + \left\| \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2 \right. \\ &\quad \left. - \left\| \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2 \right) \\ &= - \left\| \sqrt{1 + \delta_t} u_t^* + \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2 + \left\| \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2, \end{aligned}$$

implying

$$\left\| \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2 \geq \left\| \sqrt{1 + \delta_t} u_t^* + \frac{(A_1x_t^* - b_1)}{2\sqrt{1 + \delta_t}} \right\|^2,$$

and

$$1 \geq \left\| e + 2(1 + \delta_t) u_t^* \right\| \left(A_1 x_t^* - b_1 \right)^{-1}^2,$$

where $\|e\| = 1$. Which means that the sequence $\{u_t^*\}$ is bounded. In view of this and taking into account that by the supposition that $\frac{\alpha_t}{\delta_t} \xrightarrow{n \rightarrow \infty} 0$, from Eq. (11.7.11) it follows

$$\begin{aligned} \text{Const} &= \limsup_{n \rightarrow \infty} (|(x_t^* - x^*)^\top (x_t^* - x_t)| + |(u_t^* - b_1 - A_1 x_t^*)^\top u_t^*|) \geq \limsup_{n \rightarrow \infty} \frac{1}{\delta_t} \times \\ &\quad \left(\|A_0(x_t^* - x^*)\|^2 + \|A_1(x_t^* - x^*)\|^2 + (1 + \delta_t)^{-1} \|A_1 x_t^* - b_1 + (1 + \delta_t) u_t^*\|^2 \right). \end{aligned} \quad (11.7.12)$$

From Eq. (11.7.12) we may conclude that

$$(1 + \delta_t)^{-1} \|A_1 x_t^* - b_1 + (1 + \delta_t) u_t^*\|^2 = O(\delta_t), \quad (11.7.13)$$

and

$$\begin{aligned} A_0 x_\infty^* - A_0 x^* &= A_0 x_\infty^* - b_0 = 0, \\ A_1 v_\infty^* - A_1 x^* &= A_1 x_\infty^* - b_1 + u_\infty^* = 0, \end{aligned}$$

where $x_\infty^* \in X^*$ is a partial limit of the sequence $\{x_t^*\}$ which, obviously, may be not unique. The vector u_∞^* is also a partial limit of the sequence $\{u_t^*\}$.

3. Now, denote by \hat{x}_t the projection of x_t^* to the set X_{adm} , namely,

$$\hat{x}_t = \Pr_{X_{\text{adm}}} (x_t^*),$$

where \Pr is the projection operator. And show that

$$\|x_t^* - \hat{x}_t\| \leq C\sqrt{\delta_t}, \quad C = \text{const} > 0. \quad (11.7.14)$$

From Eq. (11.7.13) we have that

$$\|A_1 x_t^* - b_1 + u_t^*\| \leq C_1 \sqrt{\delta_t}, \quad C_1 = \text{const} > 0,$$

implying

$$A_1 x_t^* - b_1 \leq C_1 \sqrt{\delta_t} e - u_t^* \leq C_1 \sqrt{\delta_t} e, \quad \|e\| = 1,$$

where the vector inequality is treated in component-wise sense. Hence,

$$\|x_t^* - \hat{x}_t\|^2 \leq \max_{y \in X_{\text{adm}}} \min_{A_1 x - b_1 \leq C_1 \sqrt{\delta_t} e, x \in X_{\text{adm}}} \|x - y\|^2 := d(\delta_t).$$

Introduce the new variable

$$\tilde{x} := (1 - v_t)x + v_t \dot{x} \in X_{\text{adm}},$$

where by Slater's condition given in Eq. (11.7.7)

$$0 < v_t := \frac{C_1 \sqrt{\delta_t}}{C_1 \sqrt{\delta_t} + \max_{j=1,\dots,M_1} |(A_1 \dot{x} - b_1)_j|} < 1.$$

For the new variable $x = \frac{\tilde{x} - v_t \dot{x}}{1 - v_t}$ we have

$$\begin{aligned} A_1 \tilde{x} - b_1 &= (1 - v_t) A_1 x + v_t A_1 \dot{x} - b_1 \\ &= (1 - v_t)(A_1 x - b_1) + (1 - v_t)b_1 + v_t(A_1 \dot{x} - b_1) + v_t b_1 - b_1 \\ &= (1 - v_t)(A_1 x - b_1) + v_t(A_1 \dot{x} - b_1) \\ &\leq (1 - v_t) C_1 \sqrt{\delta_t} e + v_t(A_1 \dot{x} - b_1) \\ &= \frac{C_1 \sqrt{\delta_t}}{C_1 \sqrt{\delta_t} + \max_{j=1,\dots,M_1} |(A_1 \dot{x} - b_1)_j|} \times \\ &\quad \left(\max_{j=1,\dots,M_1} |(A_1 \dot{x} - b_1)_j| e + (A_1 \dot{x} - b_1) \right) \leq 0, \end{aligned}$$

and therefore

$$\begin{aligned} d(\delta_t) &= \max_{y \in X_{\text{adm}}} \min_{A_1 x - b_1 \leq C_1 \sqrt{\delta_t} e, x \in X_{\text{adm}}} \|x - y\|^2 \\ &\leq \max_{A_1 \tilde{x} - b_1 \leq 0, \tilde{x} \in X_{\text{adm}}} \left\| \frac{\tilde{x} - v_t \dot{x}}{1 - v_t} - \tilde{x} \right\|^2 \\ &= \frac{v_t^2}{(1 - v_t)^2} \min_{A_1 \tilde{x} - b_1 \leq 0, \tilde{x} \in X_{\text{adm}}} \|\tilde{x} - \dot{x}\|^2 \\ &\leq C_2 \delta_t, \quad C_2 > 0. \end{aligned}$$

Given that $\|x_t^* - \hat{x}_t\| \leq \sqrt{d(\delta_t)} \leq \sqrt{C_2} \sqrt{\delta_t}$ which proves Eq. (11.7.14).

4. If the proximal function (11.7.5) is strictly concave and the sequence $\{x_t\}$ of the proximal function (11.7.5) converges, then, the necessary and sufficient condition for the point x^* to be the maximum point of the function $\|x_\infty^*\|^2$ on the set X^* is given by

$$0 \geq (x_\infty^* - x^*)^\top (x_\infty^* - x_t) \text{ for any } x_\infty^* \leq X^*. \quad (11.7.15)$$

In addition, this point is unique and it has a minimal norm among all possible partial limits x_∞^* .

From Eq. (11.7.11) one obtains

$$\begin{aligned}
0 &\leq T(x_t)(x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{1}{\alpha_t} \|A_0(x_t^* - x^*)\|^2 - \frac{1}{\alpha_t} \|A_1(x_t^* - x^*)\|^2 \\
&\quad - \frac{1}{\alpha_t} (1 + \delta_t)^{-1} \|A_1 x_t^* - b_1 + u_t^*(1 + \delta_t)\|^2 - \frac{\delta_t}{\alpha_t} T(x_t)(x_t^* - x^*)^\top (x_t^* - x_t) \\
&\leq T(x_t)(x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\delta_t}{\alpha_t} T(x_t)(x_t^* - x^*)^\top (x_t^* - x_t).
\end{aligned} \tag{11.7.16}$$

By the strong concavity property

$$(y - z)^\top \left(\frac{\partial}{\partial y} f(y) - \frac{\partial}{\partial y} f(z) \right) \leq 0 \text{ for any } y, z \in \mathbb{R}^N,$$

which, in view of the property (11.7.14), implies

$$\begin{aligned}
T(x_t)(x_t^* - \hat{x}_t)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) &= O(\sqrt{\delta_t}), \\
T(x_t)(\hat{x}_t - x^*)^\top \frac{\partial}{\partial x} f(\hat{x}_t, x_t) &\leq T(x_t)(\hat{x}_t - x^*)^\top \frac{\partial}{\partial x} f(x^*, x_t) \leq 0,
\end{aligned}$$

then, we have

$$\begin{aligned}
&T(x_t)(x_t^* - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) \\
&= T(x_t)(x_t^* - \hat{x}_t)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) + T(x_t)(\hat{x}_t - x^*)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) \\
&= O(\sqrt{\delta_t}) + T(x_t)(\hat{x}_t - x^*)^\top \left(\frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\partial}{\partial x} f(\hat{x}_t, x_t) \right) \\
&\quad + T(x_t)(\hat{x}_t - x^*)^\top \frac{\partial}{\partial x} f(\hat{x}_t, x_t) \\
&\leq O(\sqrt{\delta_t}) + T(x_t)(\hat{x}_t - x^*)^\top \left(\frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\partial}{\partial x} f(\hat{x}_t, x_t) \right) \\
&\quad + T(x_t)(\hat{x}_t - x^*)^\top \frac{\partial}{\partial x} f(x^*, x_t) \\
&\leq O(\sqrt{\delta_t}) + T(x_t) \|\hat{x}_t - x^*\| \left\| \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\partial}{\partial x} f(\hat{x}_t, x_t) \right\|.
\end{aligned}$$

Since any function is Lipschitz-continuous on any bounded compact set, we can conclude that

$$\left\| \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\partial}{\partial x} f(\hat{x}_t, x_t) \right\| \leq \text{Const} \|x_t^* - \hat{x}_t\| = O(\sqrt{\delta_t}),$$

which gives

$$T(x_t)(x_t^* - \hat{x}_t)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) = O(\sqrt{\delta_t}),$$

that by Eq. (11.7.16) leads to

$$\begin{aligned} 0 &\leq T(x_t)(x_t^* - \hat{x}_t)^\top \frac{\partial}{\partial x} f(x_t^*, x_t) - \frac{\delta_t}{\alpha_t} T(x_t)(x_t^* - x^*)^\top (x_t^* - x_t) \\ &= O(\sqrt{\delta_t}) - \frac{\delta_t}{\alpha_t} T(x_t)(x_t^* - x^*)^\top (x_t^* - x_t). \end{aligned} \quad (11.7.17)$$

Dividing both sides of the inequality (11.7.17) by $\frac{\alpha_t}{\delta_t}$, taking $T(x_t) = 1$, and given that $\|x_t^* - \hat{x}_t\| \leq \kappa\sqrt{\delta_t}$ by Eq. (11.7.14) we obtain that

$$0 \leq O\left(\frac{\alpha_t}{\sqrt{\delta_t}}\right) - (x_t^* - x^*)^\top (x_t^* - x_t) = o(1)\sqrt{\delta_t} - (x_t^* - x^*)^\top (x_t^* - x_t),$$

which, by Eq. (11.7.8), for $n \rightarrow \infty$ leads to Eq. (11.7.15). Finally, for any $x^* \leq X^*$ it implies

$$\begin{aligned} 0 &\geq (x_\infty^* - x^*)^\top (x_\infty^* - x_t) = \\ \|x_\infty^* - x^*\|^2 + (x_\infty^* - x^*)^\top (x^* - x_t) &\geq (x_\infty^* - x^*)^\top (x^* - x_t). \end{aligned}$$

□

Theorem 11.2 (Antipin [5]) *If the set of solutions X is non-empty for $\delta > 0$ and the objective function $f_\delta(x, x^*)$ is differentiable in x , whose partial derivative with respect to x satisfies the Lipschitz condition with positive constant K . Then, there exists a small-enough parameter*

$$\alpha < \frac{1}{\sqrt{2} K}$$

such that, the sequence $\{x_n\}$ generated by the proximal procedure, monotonically converges with exponential rate q to one of the equilibria point x^ , i.e.,*

$$\|x_{t+1} - x^*\|^2 \leq q^{t+1} \|x_0 - x^*\|^2 \quad (11.7.18)$$

as $t \rightarrow \infty$, where

$$q = 1 + \frac{4(\alpha\delta)^2}{1 + 2\alpha\delta - 2\alpha^2 K^2} - 2\alpha\delta < 1,$$

and q_{\min} is given by

$$q_{\min} = 1 - \frac{2\alpha\delta}{1 + 2\alpha\delta} = \frac{1}{1 + 2\alpha\delta}.$$

Proof See [5]. □

11.7.4 Convergence Conditions of δ and α

Then, we present the convergence conditions and compute the estimate rate of convergence of the variables α and δ .

Theorem 11.3 *Within the class of numerical sequences*

$$\alpha_t = \begin{cases} \frac{\alpha_0}{(t+t_0)^\alpha} & \text{if } t < t_0 \\ \frac{\alpha_0}{(t+t_0)^\alpha} & \text{if } t \geq t_0 \end{cases} \quad \alpha_0, t_0, \alpha > 0,$$

$$\delta_t = \begin{cases} \frac{\delta_0}{(t+t_0)^\delta} & \text{if } t \leq t_0 \\ \frac{\delta_0}{(t+t_0)^\delta} & \text{if } t > t_0 \end{cases} \quad \delta_0, t_0, \delta > 0,$$

the parameters α_t and δ_t satisfy the following conditions:

$$0 < \alpha_t \downarrow 0, \quad \frac{\alpha_t}{\delta_t} \downarrow 0 \quad \text{when } t \rightarrow \infty,$$

$$\sum_{t=0}^{\infty} \alpha_t \delta_t = \infty \quad \text{and} \quad \frac{|\delta_{t+1} - \delta_t|}{\alpha_t \delta_t} \rightarrow 0 \quad \text{when } t \rightarrow \infty$$

for $\alpha + \delta \leq 1$, $\alpha \geq \delta$, $\alpha < 1$.

Proof It follows from the estimates that

$$\alpha_t \delta_t = O\left(\frac{1}{t^{\alpha+\delta}}\right)$$

we have that

$$\begin{aligned} |\delta_{t+1} - \delta_t| &= O\left(\frac{1}{t^\delta} - \frac{1}{(t+1)^\delta}\right) = O\left(\frac{1}{(t+1)^\delta} \left[\left(1 + \frac{1}{t}\right)^\delta - 1 \right]\right) \\ &= O\left(\frac{1}{(t+1)^\delta} \left[\left(\frac{1}{t}\right)^\delta + o(1)\right]\right) = O\left(\frac{1}{t^\delta + 1}\right) \end{aligned}$$

and

$$\frac{|\delta_{t+1} - \delta_t|}{\alpha_t \delta_t} = O\left(\frac{1}{t^{1-\alpha}}\right).$$

□

Theorem 11.4 Let x and y two variables with non-negative components for the players. Then, within the class of numerical sequences we have that

$$\alpha_t = \begin{cases} \frac{\alpha_0}{(t+t_0)^\alpha} & \text{if } t < t_0 \\ \frac{\alpha_0}{(t+t_0)^\alpha} & \text{if } t \geq t_0 \end{cases} \quad \alpha_0, t_0, \alpha > 0,$$

$$\delta_t = \begin{cases} \frac{\delta_0}{(t+t_0)^\delta} & \text{if } t \leq t_0 \\ \frac{\delta_0}{(t+t_0)^\delta} & \text{if } t > t_0 \end{cases} \quad \delta_0, t_0, \delta > 0,$$

of the procedure (11.7.2) the rate of convergence for the players is given by the parameter α_t and δ_t

$$\|x_t - x^{**}\| + \|y_t - y^{**}\| = O\left(\frac{1}{t^\kappa}\right),$$

where κ is equal to

$$\kappa = \min\{\alpha - \delta; 1 - \alpha; \delta\}. \quad (11.7.19)$$

Then, the maximal rate κ^* of convergence is attained for

$$\alpha = \alpha^* = 2/3, \quad \delta = \delta^* = 1/3. \quad (11.7.20)$$

Proof It follows that for κ_0 , the rate of convergence is given by

$$r_t = \|x_t - x^*(\delta_t)\| + \|y_t - y^*(\delta_t)\| = O\left(\frac{1}{t^{\kappa_0}}\right),$$

where $\kappa_0 = \min\{\alpha - \delta; 1 - \alpha; \delta\}$. It follows that

$$\|x_t - x^{**}\| + \|y_t - y^{**}\| = r_t + O(\delta_t) = O\left(\frac{1}{t^{\kappa_0}}\right) + O\left(\frac{1}{t^\delta}\right) = O\left(\frac{1}{t^{\min\{\kappa_0, \delta\}}}\right),$$

which implies Eq. (11.7.19). The maximal value κ of κ^* is attained when $\alpha - \delta = 1 - \alpha = \delta$, i.e., when condition (11.7.20) holds. □

Remark 11.1 is a result of Theorem 11.2 and the convergence conditions of the parameters δ and α .

References

1. Abreu, D., Manea, M.: Bargaining and efficiency in networks. *J. Econ. Theory* **147**(1), 43–70 (2012)
2. Abreu, D., Manea, M.: Markov equilibria in a model of bargaining in networks. *Games Econ. Behav.* **75**, 1–16 (2012)
3. Admati, A., Perry, M.: Strategic delay in bargaining. *Rev. Econ. Stud.* **54**(3) (1987)
4. Akin, Z.: Time inconsistency and learning in bargaining games. *Int. J. Game Theory* **36**, 275–299 (2007)
5. Antipin, A.S.: The convergence of proximal methods to fixed points of extremal mappings and estimates of their rate of convergence. *Comput. Math. Math. Phys.* **35**(5), 539–551 (1995)
6. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. *Comput. Math. Math. Phys.* **45**(11), 1893–1914 (2005)
7. Attouch, H., Soubeyran, A.: Local search proximal algorithms as decision dynamics with costs to move. *Set-Valued Anal.* **19**, 157–177 (2011)
8. Bao, T.Q., Mordukhovich, B.S., Soubeyran, A.: Variational analysis in psychological modeling. *J. Optim. Theory Appl.* **164**(1), 290–315 (2015)
9. Binmore, K., Osborne, M., Rubinstein, A.: Handbook of Game Theory, chapter. In: Noncooperative Models of Bargaining, pp. 179–225. Elsevier Science Publishers (1992)
10. Binmore, K., Piccione, M., Samuelson, L.: Evolutionary stability in alternating-offers bargaining games. *J. Econ. Theory* **80**, 257–291 (1998)
11. Binmore, K., Shaked, A., Sutton, J.: Testing noncooperative bargaining theory: a preliminary study. *Am. Econ. Assoc. Q.* **75**(5), 1178–1180 (1985)
12. Brown, D., Lewis, L.: Mgames economic agents. *Econometrica* **49**(2), 359–368 (1981)
13. Carraro, C., Marchiori, C., Sgobbi, A.: Negotiating on water: insights from non-cooperative bargaining theory. *Environ. Dev. Econ.* **12**, 329–349 (2007)
14. Clemptner, J.B.: Solving the cost to go with time penalization using the lagrange optimization approach. *Soft. Comput.* **25**(6), 4191–4199 (2021)
15. Fudenberg, D., Tirole, J.: Sequential bargaining with incomplete information. *Rev. Econ. Stud.* **50**(2), 221–247 (1983)
16. Germeyer, Y.B.: Introduction to the Theory of Operations Research. Nauka, Moscow (1971)
17. Germeyer, Y.B.: Games with Nonantagonistic Interests. Nauka, Moscow (1976)
18. Ghosh, P., Roy, N., Das, S., Basu, K.: A pricing strategy for job allocation in mobile grids using a non-cooperative bargaining theory framework. *J. Parallel Distrib. Comput.* **65**, 1366–1383 (2005)
19. Guo, X., Hernández-Lerma, O.: Continuos-Time Markov Decision Processes: Theory and Applications. Springer (2009)
20. Haller, H.: Non-cooperative bargaining of $n \geq 3$ players. *Econ. Lett.* **22**, 11–13 (1986)
21. Howard, R.: Paradoxes of Rationality: Theory of Metagames and Political Behaviour. MIT Press (1971)
22. Jun, B.: Non-cooperative bargaining and union formation. *Rev. Econ. Stud.* **56**, 59–76 (1989)
23. Kultti, K., Vartiainen, H.: Multilateral non-cooperative bargaining in a general utility space. *Int. J. Game Theory* **39**, 677–689 (2010)
24. Madani, K., Hipel, K.: Non-cooperative stability definitions for strategic analysis of generic water resources conflicts. *Water Resour.* **25**(8), 1949–1977 (2011)
25. Manea, M.: Bargaining in stationary networks. *Am. Econ. Rev.* **101**, 2042–2080 (2011)
26. Marden, J.: State based potential games. *Automatica* **48**, 3075–3088 (2012)
27. Montero, M.: Non-cooperative bargaining in apex games and the kernel. *Games Econ. Behav.* **41**, 309–321 (2002)
28. Moreno, F.G., Oliveira, P.R., Soubeyran, A.: A proximal algorithm with quasidistance. application to habit's formation. *Optimization* **61**, 1383–1403 (2011)
29. Muthoo, A.: Bargaining Theory with Applications. Cambridge University Press (2002)
30. Nash, J.F.: The bargaining problem. *Econometrica* **18**(2), 155–162 (1950)

31. Okada, A.: A non-cooperative bargaining theory with incomplete information: Verifiable types. *J. Econ. Theory* **163**, 318–341 (2016)
32. Ostrom, E.: *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press (1990)
33. Ostrom, E.: A behavioral approach to the rational choice theory of collective action. *Am. Polit. Sci. Rev.* **92**(1), 1–22 (1998)
34. Ostrom, E., Gardner, R., Walker, J.: *Rules, Games, and Common-Pool Resources*. The University of Michigan Press (1994)
35. Perry, M., Reny, P.: A non-cooperative bargaining model with straategic timed offers. *J. Econ. Theory* **59**, 50–77 (1993)
36. Poznyak, A.S.: Advanced mathematical tools for automatic control engineers. In: *Deterministic Technique*, vol. 1. Elsevier, Amsterdam (2008)
37. Poznyak, A.S., Najim, K., Gomez-Ramirez, E.: *Self-learning Control of Finite Markov Chains*. Marcel Dekker, New York (2000)
38. Rubinstein, A.: Perfect equilibrium in a bargaining model. *Econometrica* **50**(1), 97–109 (1982)
39. Rubinstein, A., Wolinsky, A.: Equilibrium in a market with sequential bargaining. *Econometrica* **53**(5), 1133–1150 (1985)
40. Selbirak, T.: Some concepts of non-myopic equilibria in games with finite strategy sets and their properties. *Ann. Oper. Res.* **51**(2), 73–82 (1994)
41. Sutton, J.: Non-cooperative bargaining theory: an introduction. *Rev. Econ. Stud.* **53**(5), 709–724 (1986)
42. Tanaka, K.: The closest solution to the shadow minimum of a cooperative dynamic game. *Comput. Math. with Appl.* **18**(1–3), 181–188 (1989)
43. Tanaka, K., Yokoyama, K.: On ϵ -equilibrium point in a noncooperative n-person game. *J. Math. Anal.* **160**, 413–423 (1991)
44. Tikhonov, A.N., Arsenin, V.Y.: *Solution of Ill-Posed Problems*. Winston & sons, Washington (1977)
45. Tikhonov, A.N.N., Goncharsky, A.V., Stepanov, V.V., G., Y.A.: *Numerical Methods for the Solution of Ill-Posed Problems*. Kluwer Academic Publishers (1995)
46. Trejo, K.K., Clempner, J.B.: New perspectives and applications of modern control theory, chapter. In: *Continuous Time Bargaining Model in Controllable Markov Games: Nash versus Kalai-Smorodinsky*, pp. 335–369. Springer International Publishing (2018)
47. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: An optimal strong equilibrium solution for cooperative multi-leader-follower Stackelberg Markov chains games. *Kybernetika* **52**(2), 258–279 (2016)
48. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the strong L_p -Nash equilibrium for Markov chains games: convergence and uniqueness. *Appl. Math. Modell.* **41**, 399–418 (2017)
49. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Nash bargaining equilibria for controllable markov chains games. In: *The 20th World Congress of The International Federation of Automatic Control (IFAC)*, pp. 12772–12777. Toulouse, France (2017)
50. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the bargaining approach for equalizing the ratios of maximal gains in continuous-time markov chains games. *Comput. Econ.* **54**, 933–955 (2019)
51. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Proximal constrained optimization approach with time penalization. *Eng. Optim.* **51**(7), 1207–1228 (2019)
52. Winoto, P., McCalla, G., Vassileva, J.: Non-monotonic-offers bargaining protocol. *Auton. Agent Multi Agent Syst.* **11**(1), 45–67 (2005)

Chapter 12

Transfer Pricing as Bargaining



Abstract The Nash bargaining game theory technique is used in this chapter to examine and provide a solution to the transfer pricing problem. We analyze a company with sequential transfers among its several divisions, where central management decides on the transfer price to maximize operational profitability. Throughout the negotiation process, the price shifting between divisions is a tool for bargaining. First, we take into account a point of contention (status quo) between the firm's divisions, which serves as a deterrent. We offer a methodology and framework for calculating the disagreement point that are based on the Nash equilibrium approach. Then, we introduce the bargaining solution, which is a single-valued function that chooses an outcome from each bargaining problem's feasible pay-offs. This solution is the result of cooperation between the company divisions involved in the transfer pricing problem. The agreement achieved by the divisions in the game is the most desirable option within the range of plausible outcomes that results in a distribution of the transfer price between divisions that maximizes profit. We provide an optimization approach for computing the negotiating solution. The method's usefulness is demonstrated through a number of examples, including Markov models in both continuous and discrete time.

12.1 Introduction

12.1.1 Transfer Pricing Process

Transfer pricing refers to the rules and methods for pricing transactions within and between enterprises under common ownership or control. Because of the potential for cross-border controlled transactions to distort taxable income, tax authorities in many countries can adjust intragroup transfer prices that differ from what would have been charged by unrelated enterprises dealing at *arm's length* (the arm's-length principle) [44]. The OECD and World Bank recommend intragroup pricing rules based on the arm's-length principle, and 19 of the 20 members of the G20 have adopted similar measures through bilateral treaties and domestic legislation,

regulations, or administrative practice. Countries with transfer pricing legislation generally follow the OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations in most respects, although their rules can differ on some important details.

Where adopted, transfer pricing rules allow tax authorities to adjust prices for most cross-border intragroup transactions, including transfers of tangible or intangible property, services, and loans. For example, a tax authority may increase a company's taxable income by reducing the price of goods purchased from an affiliated foreign manufacturer or raising the royalty the company must charge its foreign subsidiaries for rights to use a proprietary technology or brand name. These adjustments are generally calculated using one or more of the transfer pricing methods specified in the OECD guidelines and are subject to judicial review or other dispute resolution mechanisms.

12.1.2 Brief Review

The term “transfer pricing” refers to the price at which a business transfers goods and services among its cooperating (or not) divisions. It is used as a profit allocation approach when a multinational firm crosses international borders to assign its net profit before tax. The technique of distributing goods and services among several divisions in various places is used by many companies. The product travels through these several locations during the product’s production and maintenance stages as these divisions belong to more than one company. (see Fig. 12.1) [27, 28, 37, 40, 56]. This dispersed method seeks a variety of advantages, including cheaper costs, less restrictions on labor facilities, and lower taxes. However, there are a number of issues with international regulations when calculating the maximum firm-wide profit surplus among divisions. The arm-length rule mandates that businesses base their price decisions on comparable, unrelated-but-at-arm’s-length transactions between its divisions [44]. The ‘fair’ agreement about the negotiation of the transfer price among divisions is determined by transfer pricing regulations. According to the official definition of the arm’s length standard, a controlled transaction satisfies the requirement if the outcomes are comparable to those that would have been obtained if uncontrolled taxpayers had participated in the same transaction under the same conditions.

The transfer price is used to calculate expenses in a multidivisional firm when divisions are required to conduct business with one another. Due to the fact that one of the divisions involved in a transfer transaction loses money, transfer prices frequently do not deviate considerably from the market price. The issue is that this will impair the performance of the businesses if they start either purchasing for more than the current performance market price or selling below the market price. Therefore, calculating the optimal transfer price is a very intriguing challenge that has drawn the attention of researchers from a variety of fields, but it is still up for discussion among both academics and practitioners [1, 2, 8, 23, 29, 41, 57].

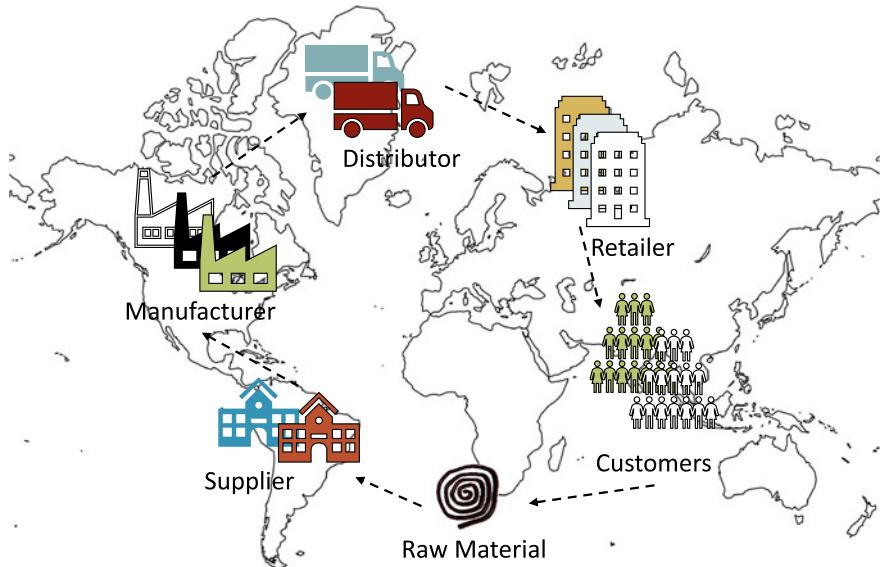


Fig. 12.1 Multidivisional firm

In determining a pricing structure for a mono-product company with a production and sales division, Hirshleifer was the first to provide analytical approaches to the transfer pricing problem and identify an efficient amount of internal trade [32, 33]. The Hirshleifer model was expanded by Arrow [4] and Baumol and Fabian [7] to settings with several products and divisions. Kanodia [36] made a suggestion on how central management may create transfer pricing that allow for risk sharing and benefit all managers. Blois [10] enhanced the Hirshleifer model [32, 33], arguing that even if central management permits descentralization, a major customer will be able to compel its suppliers to adhere to the transfer price rule of marginal cost. There is prior relevant research on transfer pricing that takes into account temporal stability [11] and a single period horizon [25]. Enzer [25] used a linear programming approach to solve the transfer pricing puzzle and arrive at an average price. Jennergren [34] suggested a method that consolidated the decision-making for the divisions, building on Enzer's work. A summary of the literature that took into account psychological and empirical evidence supporting transfer pricing was offered by Thomas in [51]. He proposed that central management enables descentralization while preserving the coordination brought on by centralization. The Dearden and Henderson concept and Thomas' notion were shown to be compatible [22, 31]. Amershi and Cheng [3], Besanko and Sibley [9] and, Ronen and Balachandran [46] found solutions to the transfer pricing conundrum by assuming that divisions are either implicitly or overtly subcontractors and by having divisions pay a transfer price and incur production expenses. In a vertically integrated supply chain, Rosenthal [47] created a cooperative game that offers transfer prices for the intermediate items (when valuation is known and when

their valuations differ), offering a solution that is just and agreeable to all divisions. Leng and Parlarb [38] extended Rosenthal's work [47] by creating a cooperative game based on calculating the Shapley value-based transfer prices for a vertically integrated supply chain firm with an upstream division and numerous downstream divisions. These divisions can independently determine their retail prices and choose whether or not to buy from the upstream division at negotiated transfer prices.

The literature has described related work that resolves transfer-price negotiating difficulties. The examination of negotiated transfer-pricing outcomes between a purchasing and selling division where each division had access to private profit information was proposed by Chalos and Haka [12]. According to Edlin and Reichelstein [24], transfer price discussions for resolving a bilateral holdup issue in a multinational corporation were explored. They demonstrated that knowledge asymmetry will lead to an unjust and ineffective transfer pricing outcome. Vaysman [55] created a negotiating model for negotiated transfer pricing that takes into account private divisional information. The business then constructs a pay scheme that makes use of both negotiated transfer pricing and divisional performance evaluation. In their study of two transfer-pricing schemes and the accompanying two-division bargaining difficulties, Haake and Martini [30] investigated two transfer-pricing systems. The literature has generally offered both cooperative and non-cooperative game theory solutions that center on bargaining [6, 13, 35, 58].

This chapter analyzes and proposes a solution for computing the transfer pricing problem considering a firm consisting of several divisions. The main results are the following:

- Proposes a solution for computing the transfer pricing problem from a point of view of the Nash bargaining game theory approach.
- Suggests a solution for computing the transfer pricing problem from a point of view of the continuous-time bargaining game theory approach.
- In this negotiation process divisions cooperate and all necessarily improve their position at the end of the process.
- Divisions operate over sequential transfers in which central management provides the transfer price decision which enables maximization of operating profits.
- The transfer pricing model involves costs and taxes.
- The division's unit production cost is dependent on the production quantity.
- The negotiation starts at the time that a division considers a disagreement point (status quo) which plays a role of a deterrent.
- Proposes a framework and method based on the Nash equilibrium approach for computing the disagreement point.
- The bargaining solution which is a single-valued function is the result of cooperation by the divisions.
- The final agreement is the most preferred alternative within the set of feasible outcomes which produces a profit-maximizing allocation of the transfer price between divisions.

- Proposes an optimization for computing the bargaining solution method.
- The result of the optimization method is a simultaneous adjustment of quantity and transfer price

12.2 Preliminaries

12.2.1 Nash's Bargaining

Detail description of Nash's bargaining game problem is given in Chap. 9.

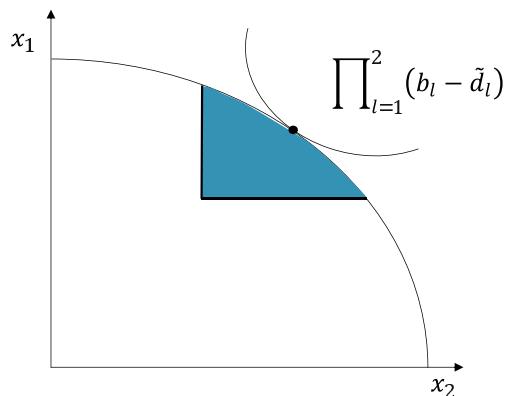
Nash's bargaining game (see Fig. 12.2) is based on a model in which players are assumed to negotiate on a set of feasible pay-offs [52, 54]. A fundamental element of the game is the disagreement point (status quo) which plays a role of a deterrent. A bargaining solution is a single-valued function that selects an outcome from the feasible pay-offs for each bargaining problem, which is the result of cooperation by the players involved in the game. The agreement reached in the game is the most preferred alternative within the set of feasible outcomes.

The bargaining problem is described by the pair (L, \tilde{d}) , where $L \subset \mathbb{R}^n$ is set of feasible payoffs, $\tilde{d} \in \mathbb{R}^n$ is a fixed disagreement vector and $l = 1, \dots, n$ is the number of players. We will call this form the condensed form of the bargaining problem (see [26, 43]). It can be derived from the normal form of an n -person game $G = \{X^1, \dots, X^n; f_1, \dots, f_n\}$ in a natural way. The set of all feasible payoffs (outcomes) is defined as

$$F = \{f : f = (f_1(x), \dots, f_n(x))\}, x \in X, \quad (12.2.1)$$

where $X = X^1 \times \dots \times X^n$. Given a disagreement vector $\tilde{d} \in \mathbb{R}^n$, the pair (F, \tilde{d}) is a bargaining problem in condensed form. We can derive another bargaining problem

Fig. 12.2 Nash bargaining



(L, \tilde{d}) from G by extending the set of feasible outcomes F to its convex hull L . Notice that any element $f \in L$ can be represented as

$$\boxed{f(\lambda) = \sum_{l=1}^n \lambda_l f_l(x(\lambda)) = (\lambda, f),} \quad (12.2.2)$$

where $f = (f_1(x(\lambda)), \dots, f_n(x(\lambda)))$, $x \in X$, $\lambda_l \geq 0$ for all l , and $\sum_{l=1}^n \lambda_l = 1$.

The payoff vector f can be realized by playing the strategies x^l with probability λ^l , and so f is the expected payoff of the players. Thus, when the players face the bargaining problem the question is, which point of L should be selected taking into account the different position and strength of the players that is reflected in the set L of extended payoffs and the disagreement point \tilde{d} .

Let B denote the set of all pairs (L, \tilde{d}) such that:

- (a) $L \subset \mathbb{R}^n$ is compact, convex;
- (b) there exists at least one $f \in L$ such that $f > \tilde{d}$ in component-wise sense.

A *Nash solution* to the bargaining problem is a vector function $b : B \rightarrow \mathbb{R}^n$, characterizing optimal distribution of payoff functions in the considered transfer bargaining problem (L, \tilde{d}) , such that $b(L, \tilde{d}) \in L$.

Proposition 12.1 *There is a unique function b such that for all $(L, \tilde{d}) \in B$ the vector $b(L, \tilde{d}) = (b_1, \dots, b_n)$ is the unique solution of the optimization problem*

$$\boxed{\begin{aligned} &\text{maximize} && g(b) = \prod_{l=1}^n (b_l - \tilde{d}_l) \\ &\text{subject to} && b \in L, \quad b \geq \tilde{d}. \end{aligned}} \quad (12.2.3)$$

The objective function of problem in Eq. (12.2.3) is usually called the *Nash product*. The condition $b \in L$ can be equivalente expressed as

$$\boxed{\min_{x \in X} f_l(x) \leq b_l \leq \max_{x \in X} f_l(x), \quad l = 1, \dots, n.} \quad (12.2.4)$$

12.2.2 Continuous-Time Bargaining

Detail description of the continuous-time bargaining game problem is given in Chap. 11. Rubinstein [48] defined seminal bargaining situation for two players ($n = 2$) who have to reach an agreement on the partition of a pie of size 1. Each player takes turns to make an offer to the other agents on how the pie should be divided between them. After player 1 has made such an offer, player 2 must decided whether to accept it, in this case the bargaining game ends and the players divide the

cake according to the accepted offer, or to reject it and continue with the bargaining process. If player 2 rejected, then this player has to make a counteroffer which player 1 would accept or reject it and continue with the negotiation process. The bargaining game continues until an offer is accepted. Offers are made at discrete points in time, and players experience an exponential discount factor which might be different across agents.

Rubinstein's [48] main results shows the existence of a subgame perfect equilibrium.

Such a game can be extended to a general set X of possible agreements. We denote by $\Phi(X) = \{(\psi^1(x), \psi^2(x)) | x \in X\}$ the set of possible utility pairs attainable at time 0, and Φ^e denote the Pareto frontier¹ of the set Φ .

When X is a compact and convex set, and the utility functions are continuous and concave, the Pareto frontier Φ^e can be represented by a graph function of a strictly decreasing and concave function, denoted by ϕ , whose domain is an interval $I^1 \subseteq \mathbb{R}$ and range an interval $I^2 \subseteq \mathbb{R}$. For simplicity, assume that $0 \in I^1$, $0 \in I^2$ and $\phi(0) > 0$. Then,

$$\Phi^e = \{(\psi^1, \psi^2) : \psi^1 \in I^1, \psi^2 \in \phi(\psi^1)\}.$$

Consider ϕ^{-1} the inverse of ϕ , a strictly decreasing and concave function from I^2 to I^1 , with $\phi^{-1}(0) > 0$. Then, for any $\psi^1 \in I^1$, $\phi(\psi^1)$ is the maximum utility that player 2 receives subject to player 1 receiving a utility ψ^1 ; in the same way, for any $\psi^2 \in I^2$, $\phi^{-1}(\psi^2)$ is the maximum utility that player 1 receives subject to player 2 receiving a utility ψ^2 .

Let Z^l , a non-empty subset of X , defined as follows

$$Z^l = \left\{ x^l := \arg \max_{x \in X} \psi^l(x) : \psi^m(x^l) = \beta^m \psi^m(x^m), (m \neq l) \right\}, \quad (12.2.5)$$

where $\beta^m = e^{(-r^m \Delta)}$ is the discount factor associated to player m .

Proposition 12.2 *For any $x^{*l} \in Z^l$, $l = 1, 2$, the following pair of strategies is a subgame perfect equilibrium of the general Rubinstein model:*

- Player 1 always offers x^{1*} and always accepts an offer x^2 if and only if $\psi^1(x^2) \geq \beta^1 \psi^{1*}$
- Player 2 always offers x^{2*} and always accepts an offer x^1 if and only if $\psi^2(x^1) \geq \beta^2 \psi^{2*}$

where $\psi^{1*} = \phi^{-1}(\beta^2 \psi^{2*})$ and $\psi^{2*} = \phi(\beta^1 \psi^{1*})$.

The generality of this Bargaining model and the proof of this result can be found in [42]. Note that if Z^l contains more than one element, then there exist more than one subgame perfect equilibrium in the general Rubinstein model. In any subgame

¹ A utility pair $(\psi^1, \psi^2) \in \Phi^e$ if and only if $(\psi^1, \psi^2) \in \Phi$ and there does not exist another utility pair $(\varphi^1, \varphi^2) \in \Phi$ such that $\varphi^1 \geq \psi^1, \varphi^2 \geq \psi^2$.

perfect equilibrium, if agreement is reached at time 0 and it is player 1 who makes the offer, then the equilibrium payoff for player 1 is ψ^{1*} and for player 2 is $\phi(\psi^{1*})$; similarly, if it is player 2 who makes the offer at time 0, then the equilibrium payoff for player 1 is $\phi^{-1}(\psi^{2*})$ and for player 2 is ψ^{2*} . This equilibrium pair is Pareto efficient.

12.3 Transfer Pricing

We consider a model of vertically integrated divisions [47]. The upstream division produces an intermediate product considering acquisition and production costs Θ . Divisions sell product either externally at positive market price p or internally to the group's downstream divisions at transfer price T . In order to focus the analysis on transfer pricing issues, the upstream division profit is given by $p - \Theta$. As well as, divisions' profit is given by $p - T$. The game consists of n divisions or players ($l = 1, \dots, n$ denoted by $l = \overline{1, n}$) which jointly make their decisions to maximize the global profits of the firm. We suppose that divisions are located in different markets to reduce the possibility of competition between the divisions. Our model considers a centralized structure allowing some divisions to act necessarily cooperatively, but divisions jointly make their decisions to maximize the profit.

The dynamics between divisions is as follows. For $l = 1, \dots, n - 1$, intermediate goods are shipped from level l to level $l + 1$, i.e., along the supply chain. Division l makes its market pricing decision p^l and sells $q^l(p^l)$ units of its final products to a given division l . Following [5, 20, 39], division l sales quantity q^l is determined by a linear demand function, i.e.

$$q^l(p^l) = \alpha^l - \beta^l p^l, \quad (12.3.1)$$

where $\alpha^l, \beta^l > 0$ and $p^l \leq \alpha^l / \beta^l$. Divisions are located in different marketing areas then they have independent demands $q^l(p^l)$ (see Fig. 12.3). Then, divisions profit is given by

$$\Phi^l(p^1, \tau^1, q^1) = p^1 q^1(p^1) = (p^1 - \Theta^1) (\alpha^1 - \beta^1 p^1), \quad (12.3.2)$$

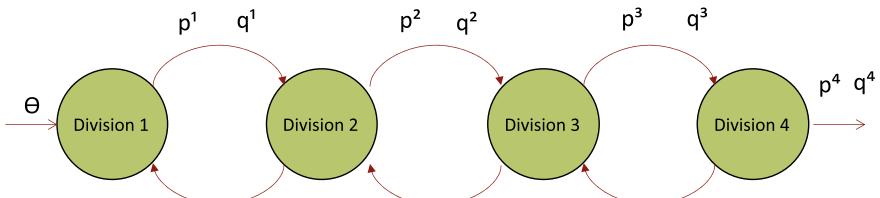


Fig. 12.3 Supply chain divisions

where Θ^l denotes acquisition and production costs, and

$$\Phi^l(p^l, \tau^l, q^l) = p^l q^l(p^l) = (p^l - T^l)(\alpha^l - \beta^l p^l), \text{ for } l \geq 2. \quad (12.3.3)$$

T^l corresponds to the transfer price that division l pays to division $l-1$ for $l \geq 2$. So Eq. (12.3.3) can be written as

$$\Phi^l(p^l, p^{l-1}, q^l) = (p^l - p^{l-1}) q^l(p^l) = (p^l - p^{l-1})(\alpha^l - \beta^l p^l), \text{ for } l \geq 2. \quad (12.3.4)$$

Because of the existence of economies of scale, we consider that the division's unit production cost is dependent on the production quantity. Then, the unit production cost which is incurred by division l when the division sells $q^l(p^l)$ units of intermediate products is represented by $c^l(q^l)$. The division's total sales quantity is

$$Q(\mathbf{q}) \equiv \sum_l q^l, \quad (12.3.5)$$

where $\mathbf{q} = (q^1(p^1), \dots, q^n(p^n))$. Then, the production cost can be written as $c(Q(\mathbf{q}))$. The corresponding effect on divisions' costs is given by $\kappa_c^l c^l(q^l) q^l$ where $\kappa_c^l \in [0, 1]$.

As well, we represent the taxes that a division l has to pay as function depending on the product and the quantity represented by $\tau^l(p^l q^l)$. We do not consider any specific function for the costs and the taxes, and we use the general form $c^l(q^l)$ and $\tau^l(p^l q^l)$ for our analysis. The corresponding effect on divisions' costs is given by $\omega_\tau^l p^l q^l$ where $\omega_\tau^l \in [0, 1]$.

Finally, the model is defined in following manner:

- Market prices p^l for one intermediate good (sold from l to $l+1$).
- Quantities q^l of intermediate good l shipped from l to $l+1$ for $l = 1, \dots, n-1$;
- Production costs

$$c^l(Q(\mathbf{q})) := \kappa_c^l c^l(q^l) q^l, \quad \kappa_c^l \in [0, 1] \quad (12.3.6)$$

(e.g.: transaction costs, raw materials, components, and their per period inventory costs in dollars) at each level $l = \overline{1, n}$;

- Taxes

$$\tau^l(p^l q^l) := \omega_\tau^l p^l q^l, \quad \omega_\tau^l \in [0, 1] \quad (12.3.7)$$

at each level $l = \overline{1, n}$.

Remark 12.1 The taxes $\tau^l(p^l q^l)$ can be eliminated from the computation process if it will be the case that a multidivisional firm does not use transfer pricing as a tool for reducing the firm's total tax payment.

We suppose that all the divisions are located in different marketing areas, therefore they face independent demands q^l , costs $c^l(q^l)$ and taxes $\tau^l(p^l q^l)$. Then, the division's utility Φ^l for each level $l = \overline{1, n}$ is given by

$$\left. \begin{aligned} \Phi^1(p^1, q^1) &= (p^1 - \Theta^1) q^1 - c^l(Q(\mathbf{q})) - \tau^1(p^1 q^1), \\ \Phi^l(p^l, p^{l-1}, q^l) &= (p^l - p^{l-1}) q^l - c^l(Q(\mathbf{q})) - \tau^l(p^l q^l), \text{ for } l = 2, \dots, n, \end{aligned} \right\} \quad (12.3.8)$$

where $l = 1$ represents the first division and $l = 2, \dots, n$ the rest of the divisions on the vertically integrated supply chain.

In the proposed model the cost and taxing information asymmetry between divisions validates divisional autonomy: delegate the production and marketing decisions. The proposed linear model for costing, taxing and pricing captures all the relevant production and marketing decisions: (1) individual deliberation on the production or selling cost, (2) separate consideration on the taxation and, (3) global computation of the transfer pricing policy for global profit maximization.

Usually, transfer pricing is concerned with an intra-firm transaction problem, it always involved strategic implications for competitive environment in which the divisions operates. The price at which transfers occur depends on the organizational structure adopted by the divisions: centralized or decentralized. The behavior of a MNE can vary widely within a market. A regulated subsidiary can intentionally have its unregulated subsidiaries overprice the parent subsidiary to increase the parent subsidiary's cost and the final price to consumers. In the meantime, the unregulated subsidiary can adopt a predatory price to prevent new divisions to entrant into the market. We consider the role of internal transfer prices within the context of an oligopoly model of competition and how transfer prices can be used as a strategic tool for competing firms to achieve tacit collusion.

Consider a transfer pricing game with n players with strategies $v^l \in V^l$ where V is a concave and compact set ($l = \overline{1, n}$). Denote by $v = (v^1, \dots, v^n)^\top \in V$ the joint strategy of the players and $v^{\hat{l}}$ is a strategy of the rest of the players adjoint to v^l , namely,

$$v^{\hat{l}} := (v^1, \dots, v^{l-1}, v^{l+1}, \dots, v^n)^\top \in V^{\hat{l}} := \bigotimes_{h=1, h \neq l}^n V^h,$$

such that $v = (v^l, v^{\hat{l}})$. A *Nash equilibrium* is a strategy $v^* = (v^{1*}, \dots, v^{n*})$ such that

$$\Phi(v^{1*}, \dots, v^l, \dots, v^{n*}) \leq \Phi(v^{1*}, \dots, v^{n*})$$

for any $v^l \in V$. A *strong Nash equilibrium* is a strategy $v^{**} = (v^{1**}, \dots, v^{n**})$ such that there does not exist any $v^l \in V$, $v^l \neq v^{l**}$

$$\Phi(v^{1**}, \dots, v^{n**}) \leq \Phi(v^{1**}, \dots, v^l, \dots, v^{n**})$$

for $v^l \in V$. A policy v^* is said to be a *Pareto policy* (or Pareto-optimal) if there is no policy v such that $\Phi(v^*) < \Phi(v)$, and similarly for weak or proper Pareto policies.

If we let $\Phi^{l*} = \sup_{v \in V} \Phi^l(v)$ and define the *utopia maximum* as $\Phi^* = (\Phi^{1*}, \dots, \Phi^{n*})$ (infeasible in general) the resulting problem is to find the values of

$$\lambda^* = \arg \max_{\lambda \in \mathcal{S}^n} \sum_l^n \lambda_l \Phi(v^*(\lambda)),$$

$$\mathcal{S}^n := \left\{ \lambda \in \mathbb{R}^n : \lambda_l \in [0, 1], \sum_{l=1}^n \lambda_l = 1 \right\},$$

in order to find the strong Nash equilibrium $v^*(\lambda)$ whose pay-off vector $\Phi(v^*(\lambda))$ is the “closest” to Φ^* in the usual Euclidean norm.

The bounds for p^l and q^l establish the maximum and minimum transfer price legally authorized (see Fig. 12.4). P_{adm} and Q_{adm} are established by the arm’s length price as well as the quantity of goods traded determining specific decision area where the optimal strategies can be selected (see Fig. 12.4).

$$p^l \in [p_-^l, p_+^l], q^l \in [q_-^l, q_+^l],$$

$$P_{adm} := \bigcap_{l=1}^n [p_-^l, p_+^l], Q_{adm} := \bigcap_{l=1}^n [q_-^l, q_+^l].$$

The admissible strategies are defined as $V_{adm} = P_{adm} \otimes Q_{adm}$.

The *Pareto set* [14, 19] can be defined as

$$\mathcal{P} := \left\{ v^*(\lambda) := \arg \max_{v \in V_{adm} = P_{adm} \otimes Q_{adm}} \Phi(v|\lambda), \lambda \in \mathcal{S}^n \right\},$$

where

$$\Phi(v|\lambda) := \lambda_1 \Phi^1(p^1, q^1) + \sum_{l=2}^n \lambda_l \Phi(p^l, p^{l-1}, q^l)$$

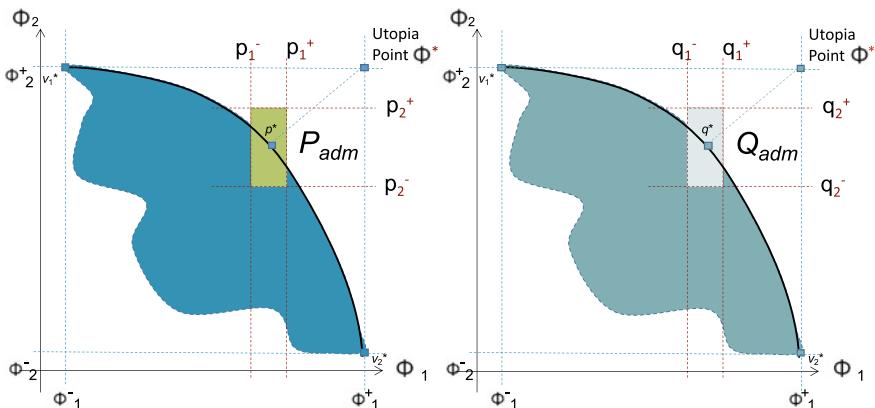


Fig. 12.4 Arm’s length price P_{adm} and quantity Q_{adm}

given

$$v = \{(p, q) : p = (p_1, \dots, p_n)^T \text{ and } q = (q_1, \dots, q_n)^T\}$$

with the Pareto front given by

$$\Phi(v^*(\lambda)) = (\Phi^1(v(\lambda)), \Phi^2(v^*(\lambda)), \dots, \Phi^n(v^*(\lambda))).$$

The vector v^* is called a *Pareto optimal solution* for \mathcal{P} .

12.4 The Transfer Pricing Nash Bargaining Solution

We start with several notations [17, 18, 53]. For a finite set of divisions (players) n with n elements, let \mathbb{R}^n denote the n -dimensional Euclidian space with coordinates indexed by $l = 1, \dots, n$. Any point in $X \subseteq \mathbb{R}^n$, called the joint strategy of the divisions, is denoted by $x = (x^l)_{l \in n}$, and also by $x = (x^1, \dots, x^n)$, $l = \overline{1, n}$. The set X is a convex and compact set. For $l \in n$ and $x = (x^l)_{l \in n} \in X$, \hat{x} denotes the $(n - 1)$ -dimensional vector constructed from x by deleting the l -th coordinate x^l . $x^{\hat{l}}$ is a strategy of the rest of the players adjoint to x^l , namely,

$$x^{\hat{l}} := (x^1, \dots, x^{l-1}, x^{l+1}, \dots, x^n)^T \in X^{\hat{l}} := \bigotimes_{m=1, m \neq l}^n X^m. \quad (12.4.1)$$

The point x is written as $(x^l, x^{\hat{l}})$.

A n -division game is defined by a triplet $\Gamma = (n, \{A_l\}_{l \in n}, \{\varphi_l\}_{l \in n})$ where n is the set of divisions and each A_l ($l \in n$) is finite set of division l 's actions. The Cartesian product $A = \bigotimes_{l=1}^n A^l$ is the set of action profiles $a = (a^1, \dots, a^n)$ for n players. Division l 's utility function φ_l is a real value function on A . Each player $l \in n$, for a given a strategy x^l , obtains the utility

$$\varphi_l(x) = \sum_{x^1 \in X^1} \dots \sum_{x^n \in X^n} \varphi_l(a^1, \dots, a^n) \prod_{l=1}^n x^l(a^l), \quad (12.4.2)$$

where $x^l(a^l)$ is the strategy of the action a^l .

Divisions try to reach the one of Nash equilibria, that is, each division l tries to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^l \in X^l$ and any $l = \overline{1, n}$

$$\varphi_l(x^l, x^{\hat{l}}) - \max_{x^l \in X^l} \varphi_l(x^l, x^{\hat{l}}) \leq 0, \quad (12.4.3)$$

for any $x^l \in X^l$ and all $l = \overline{1, n}$

where

$$\hat{x}(x) = (x^{\hat{l}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}, \quad (12.4.4)$$

and $p \geq 1$ [49, 50]. Here $\varphi_l(x^l, x^{\hat{l}})$ is the utility-function of the player l which plays the strategy $x^l \in X^l$ and the rest of the players the strategy $x^{\hat{l}} \in \hat{X}$.

Let us consider

$$G_{L_p}(x, \hat{x}(x)) := \left[\sum_{l=1}^n \left| \left(\max_{x^l \in X^l} \varphi_l(x^l, x^{\hat{l}}) \right) - \varphi_l(x^l, x^{\hat{l}}) \right|^p \right]^{1/p}. \quad (12.4.5)$$

If we define the utopia point

$$\bar{x}^l := \arg \max_{x^l \in X^l} \varphi_l(x^l, x^{\hat{l}}), \quad (12.4.6)$$

then, by replacing Eq. (12.4.6) in Eq. (12.7.2) the original problem given can be rewritten as

$$G_{L_p}(x, \hat{x}(x)) := \left[\sum_{l=1}^n \left| \varphi_l(\bar{x}^l, x^{\hat{l}}) - \varphi_l(x^l, x^{\hat{l}}) \right|^p \right]^{1/p}. \quad (12.4.7)$$

The functions $\varphi_l(x^l, x^{\hat{l}})$ ($l = \overline{1, n}$) are assumed to be concave in all their arguments.

Remark 12.2 The function $G_{L_p}(x, \hat{x}(x))$ satisfies *the Nash property*

$$\begin{aligned} \varphi_l(x^l, x^{\hat{l}}) - \varphi_l(\bar{x}^l, x^{\hat{l}}) &\leq 0 \\ \text{for any } x^l \in X^l \text{ and all } l = \overline{1, n}. \end{aligned} \quad (12.4.8)$$

If the function $G_{L_p}(x, \hat{x}(x))$ is strictly concave and the Hessian matrix is negative semi-definite, then $G_{L_p}(x, \hat{x}(x))$ attains a maximum at $(x, \hat{x}(x))$ and satisfies [53]

$$\begin{aligned} \nabla^2 G_{L_p}(x, \hat{x}(x)) &= \begin{bmatrix} \frac{\partial^2}{(\partial x_1)^2} G_{L_p}(x, \hat{x}(x)) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} G_{L_p}(x, \hat{x}(x)) \\ \frac{\partial^2}{\partial x_2 \partial x_1} G_{L_p}(x, \hat{x}(x)) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} G_{L_p, \delta}(x, \hat{x}(x)) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} G_{L_p}(x, \hat{x}(x)) & \dots & \frac{\partial^2}{(\partial x_n)^2} G_{L_p}(x, \hat{x}(x)) \end{bmatrix} = \\ &= \begin{bmatrix} \delta I_{n_1 \times n_1} & DG_{1,2}(\hat{u}_{1,2}) & \dots & DG_{1,n}(\hat{u}_{1,n}) \\ DG_{2,1}(\hat{u}_{2,1}) & \delta I_{n_2 \times n_2} & \dots & DG_{3,2}(\hat{u}_{3,2}) \\ \vdots & \ddots & \ddots & \vdots \\ DG_{3,1}(\hat{u}_{3,1}) & DG_{3,2}(\hat{u}_{3,2}) & \dots & \delta I_{n_n \times n_n} \end{bmatrix} < 0 \end{aligned}$$

or, equivalently, δ should provide the inequality

$$\delta > \max_{x \in X} [\Lambda_{\max} (\nabla^2 G_{L_p} (x, \hat{x}(x)))], \quad (12.4.9)$$

where Λ_{\max} is the maximum eigenvalue.

The bargaining game Γ is based on a model in which players are assumed to negotiate on a set of feasible pay-offs Φ . A fundamental element of the game is the disagreement point f_l^* (status quo) which plays a role of a deterrent. A bargaining solution is a single-valued function that selects an outcome from the feasible pay-offs for each bargaining problem. The agreement reached in the game is the most preferred alternative within the set of feasible outcomes Φ . Nash [43] proposed this approach by presenting four axioms and showing that they characterize the Nash bargaining solution.

Definition 12.1 (*Nash bargaining transfer pricing*). For a finite set of divisions n with n elements of a game Γ , a strategy x^* is called a *Nash bargaining solution for the transfer price* of Γ if x^* is an optimal solution of the maximization problem

$$\begin{aligned} \prod_{l=1}^n (\varphi_l (x^l, \hat{x}^l) - \tilde{f}_l (\tilde{x}^l, \hat{x}^l)) &\rightarrow \max_{x \in X} \\ \text{subject to} \\ x \in \Phi, \\ \varphi_l (x) \geq \tilde{f}_l (\tilde{x}) \text{ for all } l = 1, \dots, n, \end{aligned} \quad (12.4.10)$$

where φ is the payoff and \tilde{f}_l is the disagreement point. The pay-off $\Phi(x^*) = (\varphi_l(x^*))_{l=\overline{1,n}}$ of divisions generated by the Nash bargaining solution x^* is the *bargaining solution payoff*.

Let $\alpha^n = (\alpha_l)_{l=\overline{1,n}}$. Then, we rewrite the problem (12.4.10) as follows

$$\begin{aligned} g(x) = \sum_{l=1}^n \alpha_l \log (\varphi_l (x^l, \hat{x}^l) - \tilde{f}_l (\tilde{x}^l, \hat{x}^l)) &\rightarrow \max_{x \in X} \\ \text{subject to} \\ \varphi_l (x, \hat{x}(x)) > f_l^*, \end{aligned}$$

where α_l is called the individual weight of each division such that $\sum_{l=1}^n \alpha_l = 1$, \tilde{f}_l is the disagreement point ($l = \overline{1, n}$) or the status-quo and $\tilde{x} = (\tilde{x}^{l*}, \hat{x}^l)$ is the status-quo strategy.

Specifically, we will consider the disagreement point as division trying to reach the one of the L_p -Nash equilibria

$$\tilde{f}_l (\tilde{x}^{l*}, \hat{x}^l) = \left[\sum_{l=1}^n \left| (\varphi_l (\tilde{x}^{l*}, \hat{x}^l) - \varphi_l (x^l, \hat{x}^l)) \right|^p \right]^{1/p}.$$

12.5 Transfer Price Bargaining Solver with Additional Constraints

Let us introduce the “slack” vectors $c \in \mathbb{R}^n$ with nonnegative components, that is, $c_j \geq 0$ for all $j = 1, \dots, n$, the original problem (12.7.2) can be rewritten as

$$\left. \begin{aligned} \{g(x, \hat{x}(x))\} &\rightarrow \max_{x \in X_{adm}, c \geq 0} \\ X_{adm} := \{x \in X : x \geq 0, \quad A_{eq}x = b_{eq}, \quad A_{ineq}x - b_{ineq} + c = 0\}. \end{aligned} \right\} \quad (12.5.1)$$

Notice that this problem may have non-unique solution and $\det(A_{eq}^\top A_{eq}) = 0$. Define by $X^* \subseteq X_{adm}$ the set of all solutions of the problem (12.5.1).

Consider the *penalty function* given by

$$\tilde{\mathcal{V}}_k(x, c) := \mu\{g(x, \hat{x}(x))\} - k \left[\frac{1}{2} \|A_{eq}x - b_{eq}\|^2 + \frac{1}{2} \|A_{ineq}x - b_{ineq} + c\|^2 \right], \quad (12.5.2)$$

where the parameters k and c are positive. Notice also that

$$\arg \max_{x \in X_{adm}, c \geq 0} \tilde{\mathcal{V}}_{k,\delta}(x, c) = \arg \max_{x \in X_{adm}, c \geq 0} \mathcal{V}_{\mu,\delta}(x, c)$$

where $\mu := k^{-1} > 0$ and

$$\begin{aligned} \mathcal{V}_{\mu,\delta}(x, c) := \mu\{g(x, \hat{x}(x))\} - \frac{1}{2} \|A_{eq}x - b_{eq}\|^2 - \frac{1}{2} \|A_{ineq}x - b_{ineq} + c\|^2 - \\ \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 + \|c\|^2) \end{aligned}$$

Obviously, the optimization problem

$$\tilde{\mathcal{V}}_{\mu,\delta}(x, c) \rightarrow \max_{x \in X_{adm}, c \geq 0} \quad (12.5.3)$$

has a unique solution since the optimized function (12.5.2) is *strongly convex* [45] if $\delta > 0$. The Penalty Functions Method (PFM) consists on the following idea: If the penalty parameter μ and the regularizing parameter δ tend to zero by a particular manner, then we may expect that $x^*(\mu, \delta)$ and $c^*(\mu, \delta)$, which are the solutions of the optimization problem

$$\mathcal{V}_{\mu,\delta}(x, c) \rightarrow \max_{x \in X_{adm}, c \geq 0},$$

tend to the set V^* of all solutions of the original optimization problem (12.5.1), that is,

$$\rho \{x^*(\mu, \delta), c^*(\mu, \delta); X^*\} \xrightarrow[\mu, \delta \downarrow 0]{} 0, \quad (12.5.4)$$

where $\rho\{a; X^*\}$ is the Hausdorff distance defined as

$$\rho\{a; X^*\} = \min_{x \in X^*} \|a - x^*\|^2.$$

Below we define exactly how the parameters μ and δ should tend to zero to provide the property (12.5.4).

Then, if we assume that

- (a) the bounded set X^* of all solutions of the original optimization problem (12.5.1) is not empty and the Slater's condition holds, that is, there exists a point $\hat{x} \in X_{adm}$ such that

$$A_{ineq}\hat{x} < b_1, \quad (12.5.5)$$

- (b) the parameters μ and δ are time-varying, i.e.,

$$\mu = \mu_n, \quad \delta = \delta_n \quad (n = 0, 1, 2, \dots)$$

such that

$$0 < \mu_n \downarrow 0, \quad \frac{\mu_n}{\delta_n} \downarrow 0 \quad \text{when } n \rightarrow \infty. \quad (12.5.6)$$

Then

$$\left. \begin{aligned} x_n^* &:= x^*(\mu_n, \delta_n) \xrightarrow{n \rightarrow \infty} x^{**}, \\ c_n^* &:= c^*(\mu_n, \delta_n) \xrightarrow{n \rightarrow \infty} c^{**}, \end{aligned} \right\} \quad (12.5.7)$$

where $x^{**} \in X^*$ is the solution of the original problem (12.5.1) with the minimal weighted norm, i.e.,

$$\|x^{**}\| \leq \|x^*\| \quad \text{for all } x^* \in X^* \quad (12.5.8)$$

and

$$c^{**} = b_1 - A_1 x^{**}. \quad (12.5.9)$$

The format version ($n = 0, 1, \dots$) of the proximal method with some fixed admissible initial values ($x_0 \in X$, $\hat{x}_0(x) \in \hat{X}$, $c_0 \geq 0$) is as follows

$$\begin{aligned} x_{n+1} &= \arg \max_{x \in X} \left\{ -\frac{1}{2} \|x - x_n\|^2 + \gamma \mathcal{V}_{\mu_n, \delta_n}(x, \hat{x}_n(x), c_n) \right\}, \\ \hat{x}_{n+1}(x) &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\frac{1}{2} \|\hat{x}(x) - \hat{x}_n(x)\|^2 + \gamma \mathcal{V}_{\mu_n, \delta_n}(x_n, \hat{x}(x), c_n) \right\}, \end{aligned} \quad (12.5.10)$$

where $\mathcal{V}_{\mu_n, \delta_n}(x, \hat{x}(x), c)$ is given by

$$\begin{aligned}\mathcal{V}_{\mu_n, \delta_n}(x, \hat{x}_n(x), c) &:= \mu g(x, \hat{x}_n(x)) - \frac{1}{2} (A_{ineq}x - b_{ineq} + c) - \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 + \|c\|^2) \\ &= \mu g(x, \hat{x}_n(x)) - \frac{1}{2} (A_{ineq}x - b_{ineq} + c) - \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 + \|c\|^2).\end{aligned}$$

Then, developing we have

$$\begin{aligned}x_{n+1} &= -\frac{1}{2}\|x - x_n\|^2 + \gamma \mathcal{V}_{\mu_n, \delta_n}(x, \hat{x}_n(x), c_n) = \\ &= -\frac{1}{2}\|x - x_n\|^2 + \gamma [\mu g(x, \hat{x}_n(x)) - \frac{1}{2} (A_{ineq}x - b_{ineq} + c) - \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 + \|c\|^2)] \\ \tilde{x}_{n+1}(x) &= -\frac{1}{2}\|\hat{x}(x) - \hat{x}_n(x)\|^2 + \gamma \mathcal{V}_{\mu_n, \delta_n}(x_n, \hat{x}(x), c_n) = \\ &= -\frac{1}{2}\|\hat{x}(x) - \hat{x}_n(x)\|^2 + \gamma [\mu g(x_n, \hat{x}(x)) - \frac{1}{2} (A_{ineq}x - b_{ineq} + c) - \frac{\delta}{2} (\|x\|^2 + \|\hat{x}\|^2 + \|c\|^2)].\end{aligned}$$

12.5.1 Numerical Example for Nash's Bargaining Transfer Pricing

Our goal is to analyze a three-player non-cooperative bargaining situation in a class of discrete-time Markov chains. Let us consider a transfer pricing approach [15, 16, 20, 21], which considers three divisions. We are taking into account that the Markov chain game is ergodic.

Let the number of states $N = 3$ and the number of actions $M = 2$ for each division. The individual utility for each division are defined by

$$\begin{aligned}U_{i,j,1}^1 &= \begin{bmatrix} 11 & 8 & 12 \\ 7 & 11 & 8 \\ 11 & 13 & 12 \end{bmatrix}, \quad U_{i,j,1}^2 = \begin{bmatrix} 8 & 1 & 9 \\ 4 & 11 & 13 \\ 8 & 7 & 1 \end{bmatrix}, \quad U_{i,j,1}^3 = \begin{bmatrix} 6 & 8 & 19 \\ 14 & 11 & 22 \\ 6 & 2 & 9 \end{bmatrix}, \\ U_{i,j,2}^1 &= \begin{bmatrix} 3 & 11 & 6 \\ 2 & 17 & 13 \\ 17 & 8 & 1 \end{bmatrix}, \quad U_{i,j,2}^2 = \begin{bmatrix} 12 & 7 & 11 \\ 6 & 9 & 10 \\ 13 & 9 & 8 \end{bmatrix}, \quad U_{i,j,2}^3 = \begin{bmatrix} 1 & 12 & 5 \\ 4 & 3 & 5 \\ 21 & 2 & 4 \end{bmatrix}.\end{aligned}$$

The transition rate matrices, i.e. the matrices with the information about the behavior of each division, are defined as follows

$$\pi_{j|i,1}^1 = \begin{bmatrix} 0.8065 & 0.0482 & 0.1454 \\ 0.0936 & 0.7380 & 0.1684 \\ 0.0455 & 0.0445 & 0.9100 \end{bmatrix}, \quad \pi_{j|i,2}^1 = \begin{bmatrix} 0.4711 & 0.1180 & 0.4109 \\ 0.2094 & 0.3442 & 0.4464 \\ 0.1352 & 0.1124 & 0.7524 \end{bmatrix},$$

Fig. 12.5 SNE Strategies of player 1

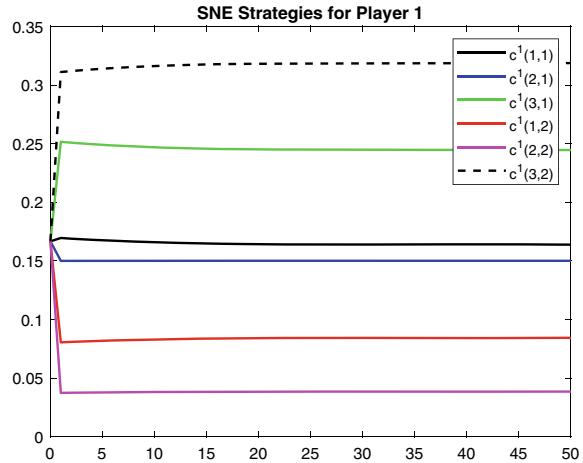
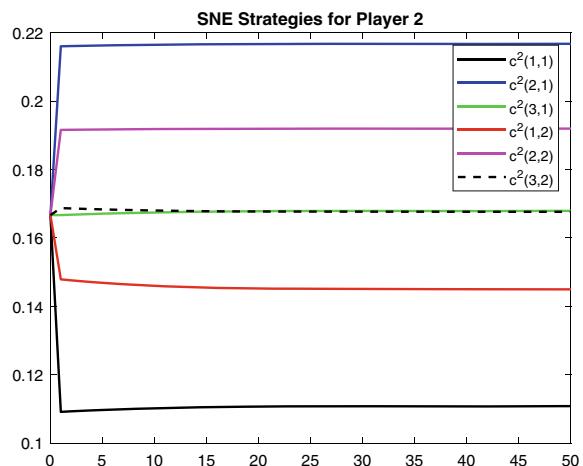


Fig. 12.6 SNE Strategies of player 2



$$\pi_{j|i,1}^2 = \begin{bmatrix} 0.4741 & 0.3486 & 0.1773 \\ 0.2143 & 0.5350 & 0.2507 \\ 0.1094 & 0.3291 & 0.5615 \end{bmatrix}, \quad \pi_{j|i,2}^2 = \begin{bmatrix} 0.8009 & 0.1478 & 0.0513 \\ 0.0941 & 0.8065 & 0.0995 \\ 0.0253 & 0.1350 & 0.8396 \end{bmatrix},$$

$$\pi_{j|i,1}^3 = \begin{bmatrix} 0.7298 & 0.2292 & 0.0410 \\ 0.0503 & 0.8587 & 0.0910 \\ 0.1956 & 0.2012 & 0.6032 \end{bmatrix}, \quad \pi_{j|i,2}^3 = \begin{bmatrix} 0.5079 & 0.4069 & 0.0852 \\ 0.1116 & 0.7446 & 0.1438 \\ 0.2817 & 0.3741 & 0.3442 \end{bmatrix}.$$

First let's calculate the starting point of the bargaining process applying the proximal method (12.5.10) to find the strong Nash equilibrium. We obtain the convergence of the strategies in terms of the variable $c_{(i,k)}^l$ for each player (division) $l = \overline{1, n}$ (see Figs. 12.5, 12.6 and 12.7) and the convergence of the parameter λ (see Fig. 12.8).

Fig. 12.7 SNE Strategies of player 3

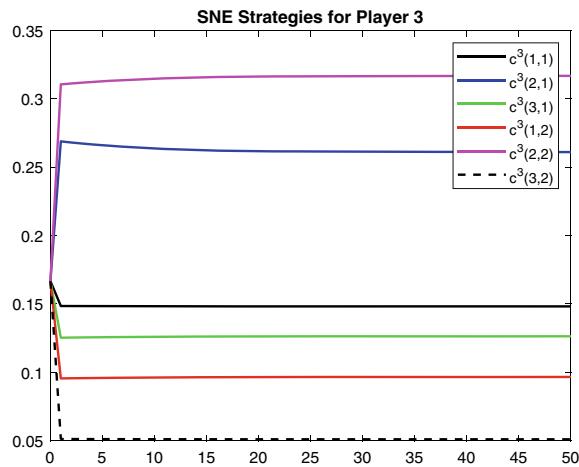
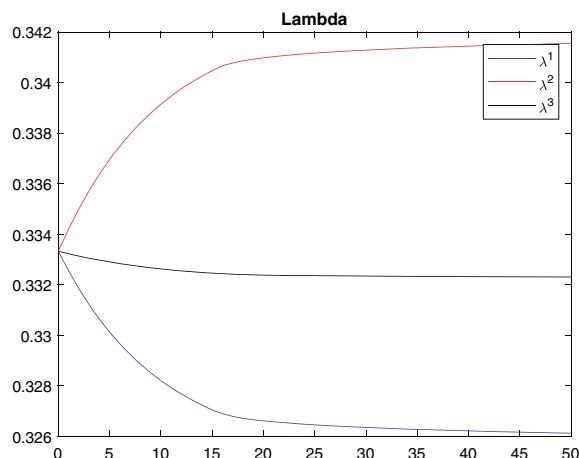


Fig. 12.8 Convergence of λ



The strong Nash equilibrium reached for all players is as follows:

$$c^1 = \begin{bmatrix} 0.1638 & 0.0843 \\ 0.1501 & 0.0385 \\ 0.2445 & 0.3187 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.1108 & 0.1449 \\ 0.2168 & 0.1920 \\ 0.1679 & 0.1676 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.1482 & 0.0966 \\ 0.2610 & 0.3168 \\ 0.1263 & 0.0510 \end{bmatrix}.$$

The utilities for each player in the strong Nash equilibrium are $\psi^1(c^1, c^2, c^3) = 8.4156$, $\psi^2(c^1, c^2, c^3) = 8.1339$ and $\psi^3(c^1, c^2, c^3) = 7.1423$. Once the starting point is set, the negotiation process between players begins, calculating the strategies until they converge.

Fig. 12.9 Strategies of player 1

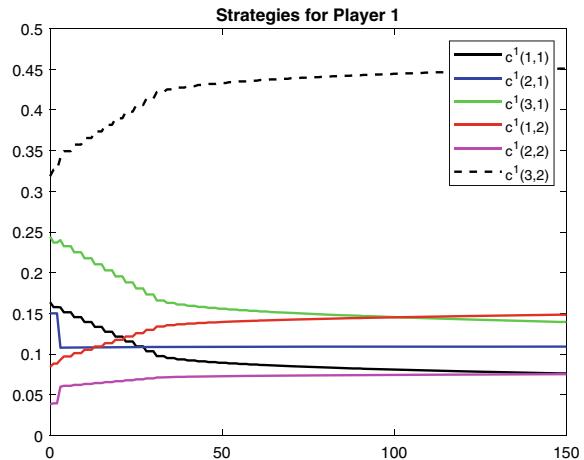
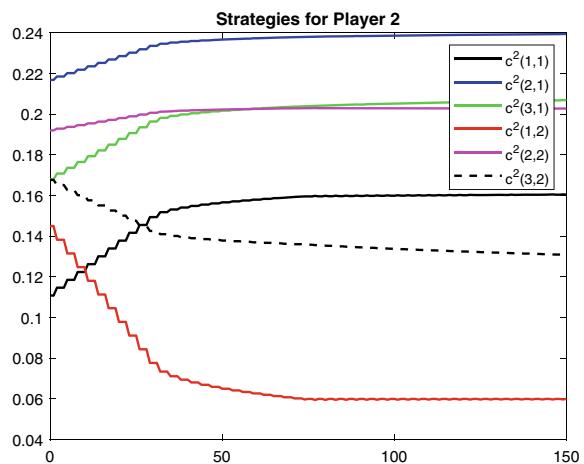


Fig. 12.10 Strategies of player 2



In this model each player calculates the strategies independently until they reach an agreement. Figures 12.9, 12.10 and 12.11 show the behavior of the offers (strategies) during the bargaining process.

Finally, the agreement reached is as follows:

$$c^1 = \begin{bmatrix} 0.0763 & 0.1485 \\ 0.1093 & 0.0754 \\ 0.1396 & 0.4508 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.1604 & 0.0600 \\ 0.2393 & 0.2027 \\ 0.2068 & 0.1308 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.1427 & 0.1160 \\ 0.1276 & 0.4287 \\ 0.1250 & 0.0601 \end{bmatrix}.$$

The mixed strategies obtained for players are as follows

Fig. 12.11 Strategies of player 3

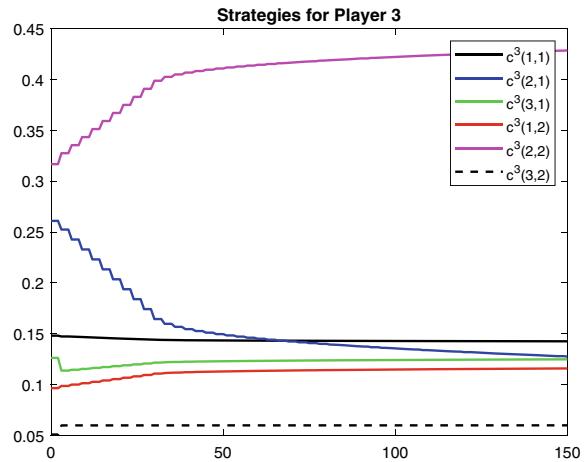
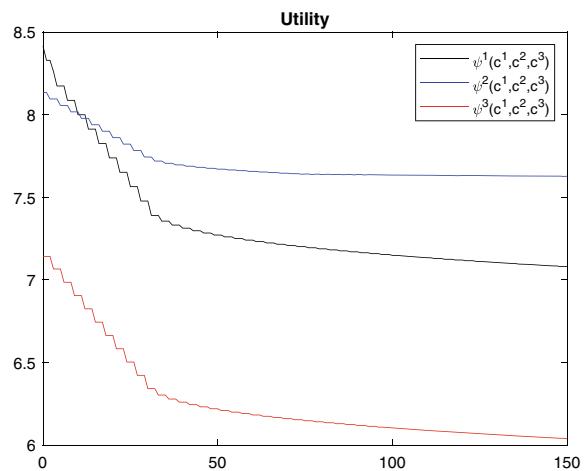


Fig. 12.12 Behavior of players' utilities



$$d^1 = \begin{bmatrix} 0.3395 & 0.6605 \\ 0.5917 & 0.4083 \\ 0.2365 & 0.7635 \end{bmatrix}, \quad d^2 = \begin{bmatrix} 0.7277 & 0.2723 \\ 0.5414 & 0.4586 \\ 0.6126 & 0.3874 \end{bmatrix}, \quad d^3 = \begin{bmatrix} 0.5515 & 0.4485 \\ 0.2294 & 0.7706 \\ 0.6754 & 0.3246 \end{bmatrix}.$$

With the strategies calculated at each step of the negotiation process, the utilities of each player showed a decreasing behavior as shown in the Fig. 12.12, i.e., at each step of the bargaining process, the utility of each player decreases until they reach an agreement. At the end of the bargaining process, the resulting utilities are as follows $\psi^1(c^1, c^2, c^3) = 7.0811$, $\psi^2(c^1, c^2, c^3) = 7.6264$ and $\psi^3(c^1, c^2, c^3) = 6.0419$ for each player.

12.6 Continuous-Time Transfer Pricing

12.6.1 Revenue of a Passenger Between Members of an Airline Alliance

Our goal is to analyze a three-player non-cooperative bargaining situation in a class of continuous-time Markov chains. Let us consider a transfer pricing approach [15, 16, 20, 21] which divide the revenue of a passenger between members of an airline alliance. The set of origin-destination time are made up of itineraries. The itineraries are either a direct flight or a series of connecting flights within the supply chain represented by the airlines network. The game penalizes the revenue taking into account the total time that a passenger takes for reaching the final destination. We are taking into account only round trips so the Markov chain game is ergodic.

Let the number of states $N = 3$ and the number of actions $M = 2$ for each airline. The individual utility for each airline are defined by

$$U_{i,j,1}^1 = \begin{bmatrix} 10 & 8 & 12 \\ 6 & 11 & 19 \\ 10 & 14 & 13 \end{bmatrix}, U_{i,j,1}^2 = \begin{bmatrix} 7 & 9 & 11 \\ 5 & 10 & 14 \\ 9 & 6 & 10 \end{bmatrix}, U_{i,j,1}^3 = \begin{bmatrix} 17 & 9 & 6 \\ 19 & 13 & 11 \\ 3 & 2 & 8 \end{bmatrix},$$

$$U_{i,j,2}^1 = \begin{bmatrix} 12 & 10 & 5 \\ 20 & 16 & 14 \\ 18 & 9 & 11 \end{bmatrix}, U_{i,j,2}^2 = \begin{bmatrix} 15 & 6 & 9 \\ 15 & 8 & 9 \\ 12 & 10 & 7 \end{bmatrix}, U_{i,j,2}^3 = \begin{bmatrix} 10 & 12 & 3 \\ 4 & 10 & 9 \\ 20 & 17 & 19 \end{bmatrix}.$$

The transition rate matrices, i.e. the matrices with the information about the behavior of each airline, are defined as follows

$$q_{j|i,1}^1 = \begin{bmatrix} -0.2230 & 0.0581 & 0.1649 \\ 0.1166 & -0.3131 & 0.1965 \\ 0.0504 & 0.0531 & -0.1034 \end{bmatrix} \quad q_{j|i,2}^1 = \begin{bmatrix} -0.8918 & 0.2323 & 0.6595 \\ 0.4664 & -1.2526 & 0.7862 \\ 0.2014 & 0.2122 & -0.4137 \end{bmatrix}$$

$$q_{j|i,1}^2 = \begin{bmatrix} -0.9336 & 0.7250 & 0.2086 \\ 0.4673 & -0.9428 & 0.4755 \\ 0.0862 & 0.6542 & -0.7405 \end{bmatrix} \quad q_{j|i,2}^2 = \begin{bmatrix} -0.2334 & 0.1813 & 0.0521 \\ 0.1168 & -0.2357 & 0.1189 \\ 0.0216 & 0.1636 & -0.1851 \end{bmatrix}$$

$$q_{j|i,1}^3 = \begin{bmatrix} -0.3297 & 0.2872 & 0.0426 \\ 0.0473 & -0.1738 & 0.1265 \\ 0.2912 & 0.2401 & -0.5313 \end{bmatrix} \quad q_{j|i,2}^3 = \begin{bmatrix} -0.7694 & 0.6700 & 0.0993 \\ 0.1103 & -0.4056 & 0.2953 \\ 0.6794 & 0.5602 & -1.2396 \end{bmatrix}$$

First let's calculate the starting point of the bargaining process applying the proximal method to find the strong Nash equilibrium. We obtain the convergence of the strategies in terms of the variable $c_{(i,k)}^l$ for each player (airline) $l = \overline{1, n}$ (see Figs. 12.13, 12.14 and 12.15) and the convergence of the parameter λ (see Fig. 12.16).

Fig. 12.13 SNE Strategies of player 1

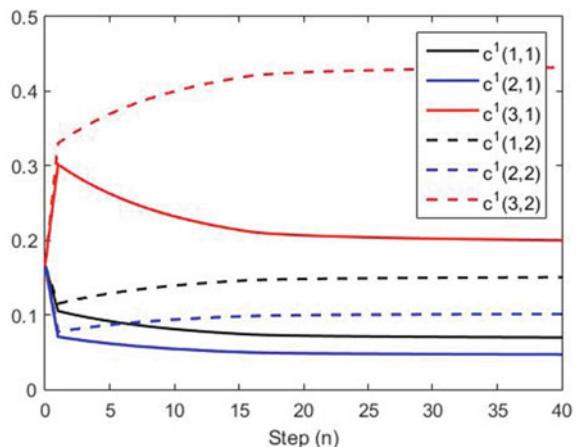


Fig. 12.14 SNE Strategies of player 2

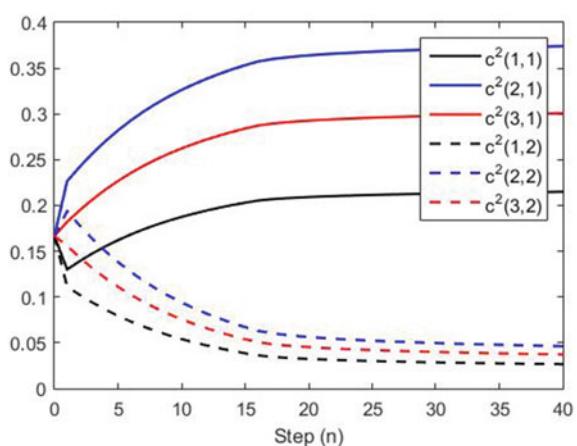


Fig. 12.15 SNE Strategies of player 3

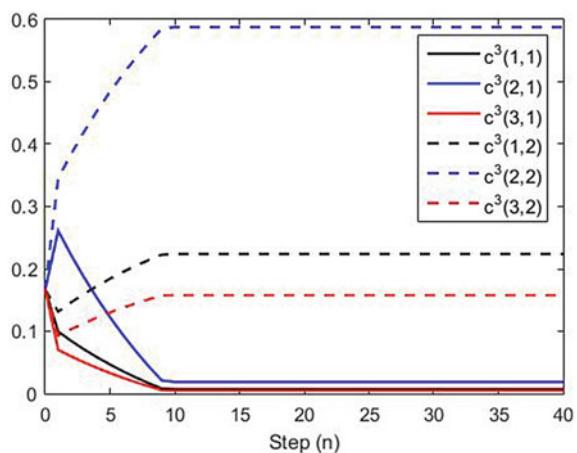
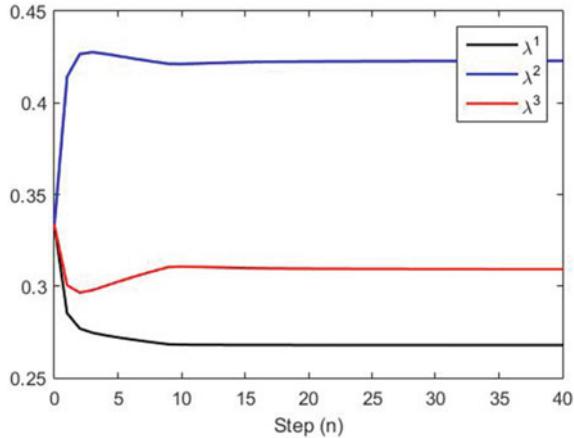


Fig. 12.16 Convergence of λ



The strong Nash equilibrium reached for all players is as follows:

$$c^1 = \begin{bmatrix} 0.0691 & 0.1510 \\ 0.0464 & 0.1015 \\ 0.1984 & 0.4336 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.2163 & 0.0253 \\ 0.3764 & 0.0440 \\ 0.3026 & 0.0354 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.0071 & 0.2237 \\ 0.0187 & 0.5876 \\ 0.0050 & 0.1579 \end{bmatrix}.$$

The utilities for each player in the strong Nash equilibrium are $\psi^1(c^1, c^2, c^3) = 3842.4$, $\psi^2(c^1, c^2, c^3) = 2961.7$ and $\psi^3(c^1, c^2, c^3) = 3560.3$. Once the starting point is set, the negotiation process between players begins, calculating the strategies until they converge.

12.6.1.1 The Non-cooperative Bargaining Solution

In this model each player calculates the strategies independently and alternately following the relation (12.5.10) until they reach an agreement. Figures 12.17, 12.18 and 12.19 show the behavior of the offers (strategies) during the bargaining process. Finally, the agreement reached is as follows:

$$c^1 = \begin{bmatrix} 0.2028 & 0.0173 \\ 0.1363 & 0.0116 \\ 0.5824 & 0.0495 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.0691 & 0.1725 \\ 0.1202 & 0.3001 \\ 0.0967 & 0.2413 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.1320 & 0.0988 \\ 0.3469 & 0.2594 \\ 0.0932 & 0.0697 \end{bmatrix}.$$

The mixed strategies obtained for players are as follows

$$d^1 = \begin{bmatrix} 0.9216 & 0.0784 \\ 0.9216 & 0.0784 \\ 0.9216 & 0.0784 \end{bmatrix}, \quad d^2 = \begin{bmatrix} 0.2860 & 0.7140 \\ 0.2860 & 0.7140 \\ 0.2860 & 0.7140 \end{bmatrix}, \quad d^3 = \begin{bmatrix} 0.5721 & 0.4279 \\ 0.5721 & 0.4279 \\ 0.5721 & 0.4279 \end{bmatrix}.$$

Fig. 12.17 Strategies of player 1

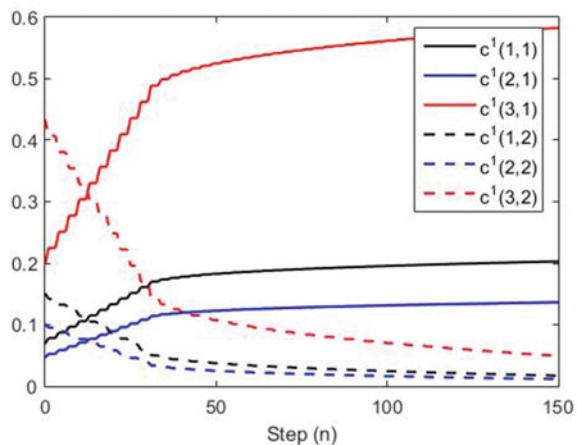


Fig. 12.18 Strategies of player 2

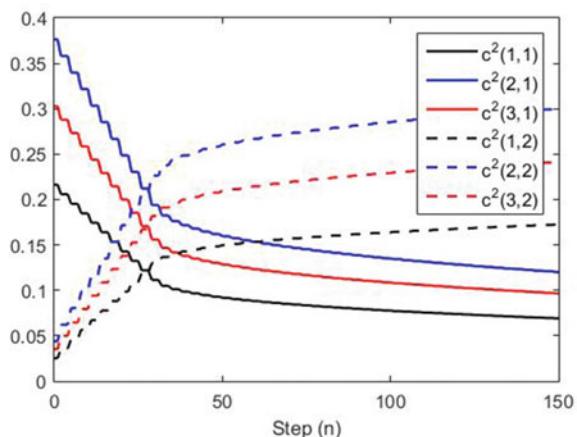


Fig. 12.19 Strategies of player 3

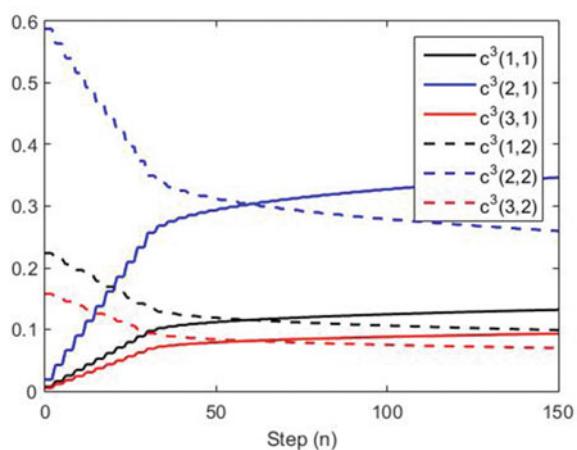
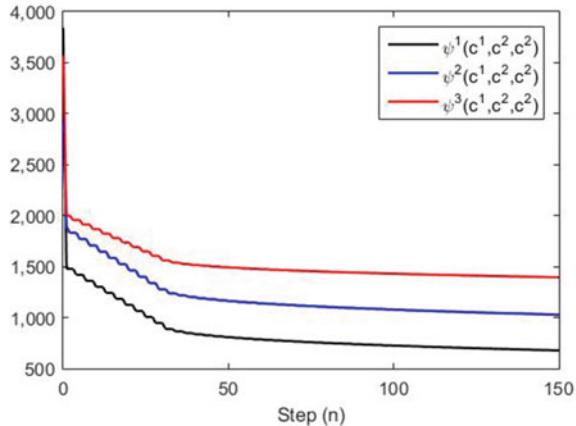


Fig. 12.20 Behavior of players' utilities



With the strategies calculated at each step of the negotiation process, the utilities of each player showed a decreasing behavior as shown in the Fig. 12.20, i.e., at each step of the bargaining process, the utility of each player decreases until they reach an agreement. At the end of the bargaining process, the resulting utilities are as follows $\psi^1(c^1, c^2, c^3) = 678.2$, $\psi^2(c^1, c^2, c^3) = 1028.0$ and $\psi^3(c^1, c^2, c^3) = 1394.3$ for each player.

12.7 Extensions

We now present two extensions of the bargaining game provided above that include the case when agents have different discount factors, and another where agents might coordinate on their demands. The convergence of results follows trivially from our general analysis presented in the appendix above.

12.7.1 Bargaining Under Different Discounting

In this approach we present a solution where at each step of the negotiation process players calculate the Nash equilibrium considering the utility functions of all players but with the particularity that internally each player reaches this equilibrium point in a different time. Following the description of the model presented previously, we redefine the advantage of propose a new offer that depends on the utility function

$$f(x_t, x_{t+1}) := \sum_{l=1}^n [\psi^l(x_{t+1}) - \psi^l(x_t)] \geq 0,$$

for all players to reject the offer x_t and making a new offer x_{t+1} given the time spent to benefit of this advantage $T(x_{t+1}) > 0$, and $\alpha^l(x_t)$ be the weight that players put on their advantages to reject the offer x_t . Thus, the advantages to reject the offer x_t and to propose a new offer x_{t+1} are given by $A(x_t, x_{t+1}) = \alpha(x_t)T(x_{t+1})f(x_t, x_{t+1})$.

Remark 12.3 The function $f(x_t, x_{t+1})$ satisfies the Nash condition

$$\psi^l(x_{t+1}) - \psi^l(x_t) \geq 0,$$

for any $x \in X$ and all players.

Definition 12.2 A strategy $x^* \in X$ is said to be a Nash equilibrium if

$$x^* \in \operatorname{Arg} \max_{x \in X} \{f(x_t, x_{t+1})\}.$$

Then, at each step of the bargaining game we have in proximal format that the players must select their strategies according to

$$x^* = \arg \max_{x \in X} \left\{ -\delta_t T(x) \| (x - x^*) \|^2 + \alpha_t T(x) f(x, x^*) \right\}, \quad (12.7.1)$$

where

$$f(x, x^*) := \sum_{l=1}^n [\psi^l(x) - \psi^l(x^*)].$$

At each step of the bargaining process, players calculate simultaneously the Nash equilibrium but considering that each player reach the equilibrium in a different time.

12.7.1.1 Markov Chains Description

Let us to define the Nash equilibrium as a strategy $x^* = (x^{1*}, \dots, x^n)$ such that

$$\psi(x^{1*}, \dots, x^{n*}) \geq \psi(x^{1*}, \dots, x^l, \dots, x^{n*})$$

for any $x^l \in X$.

Consider that players try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^l \in X^l$ and any $l = \overline{1, n}$

$$f(x, \hat{x}(x)) := \sum_{l=1}^n [\psi^l(x^l, x^{\hat{l}}) - \psi^l(\bar{x}^l, x^{\hat{l}})] \leq 0, \quad (12.7.2)$$

where $\hat{x} = (x^{\hat{1}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}$ [49, 50], \bar{x}^l is the utopia point defined as Eq. (12.4.6) and $\psi^l(x^l, x^{\hat{l}})$ is the concave cost-function of player l which plays

the strategy $x^l \in X^l$ and the rest of players the strategy $x^{\hat{l}} \in X^{\hat{l}}$ considering the time function.

Recall that a strategy $x^* \in X$ is a Nash equilibrium if

$$x^* \in \text{Arg} \max_{x \in X_{\text{adm}}} \{f(x, \hat{x}(x))\}.$$

Remark 12.4 If $f(x, \hat{x}(x))$ is strictly concave then

$$x^* = \arg \max_{x \in X_{\text{adm}}} \{f(x, \hat{x}(x))\}.$$

We redefine the utility function that depends of the average utility function of all players as follows

$$\begin{aligned} F(x, \hat{x}(x)) &:= f(x, \hat{x}(x)) - \frac{1}{2} \sum_{l=1}^n \sum_{j=1}^N \mu_{(j)}^l h_{(j)}^l(x^l) - \\ &\quad \frac{1}{2} \sum_{l=1}^n \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^M \xi_{(j)}^l q_{(j|i,k)}^l x_{(i,k)}^l - \frac{1}{2} \sum_{l=1}^n \sum_{i=1}^N \sum_{k=1}^M \eta^l (x_{(i,k)}^l - 1), \end{aligned}$$

then we may conclude that

$$x^* = \arg \max_{x \in X, \hat{x} \in \hat{X}} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} F(x, \hat{x}(x), \mu, \xi, \eta). \quad (12.7.3)$$

Finally we have that at each step of the bargaining process, players calculate the Nash equilibrium (but they reach the equilibrium at different time) according to the solution of the non-cooperative bargaining problem in proximal format defined as follows

$$\left. \begin{aligned} \mu^* &= \arg \min_{\mu \geq 0} \{-\delta \|\mu - \mu^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu, \xi^*, \eta^*)\}, \\ \xi^* &= \arg \min_{\xi \geq 0} \{-\delta \|\xi - \xi^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu^*, \xi, \eta^*)\}, \\ \eta^* &= \arg \min_{\eta \geq 0} \{-\delta \|\eta - \eta^*\|^2 + \alpha F(x^*, \hat{x}^*(x), \mu^*, \xi^*, \eta)\}, \\ x^* &= \arg \max_{x \in X} \{-\delta \|(x - x^*)\|_{\Lambda}^2 + \alpha F(x, \hat{x}^*(x), \mu^*, \xi^*, \eta^*)\}, \\ \hat{x}^* &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\delta \|(\hat{x} - \hat{x}^*)\|_{\Lambda}^2 + \alpha F(x^*, \hat{x}(x), \mu^*, \xi^*, \eta^*) \right\}. \end{aligned} \right\} \quad (12.7.4)$$

12.7.1.2 Transfer Pricing Simulation

Following the Section above, in this model each player calculates the strategies according the Nash equilibrium formulation where players calculate the Nash equilibrium simultaneously, but with the characteristic that they reach the equilibrium at different time, following the relation (12.7.4) until they reach an agreement (strate-

Fig. 12.21 Strategies of player 1 in the bargaining model 2

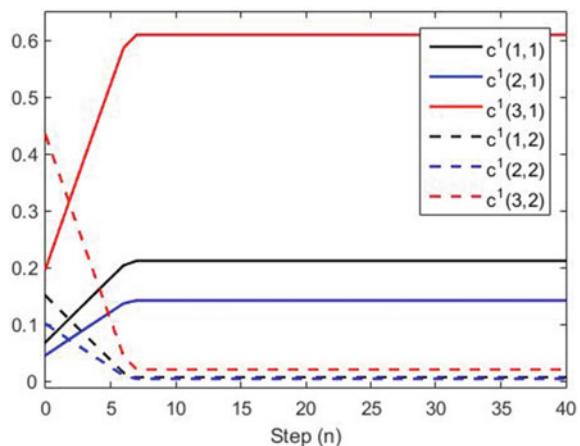
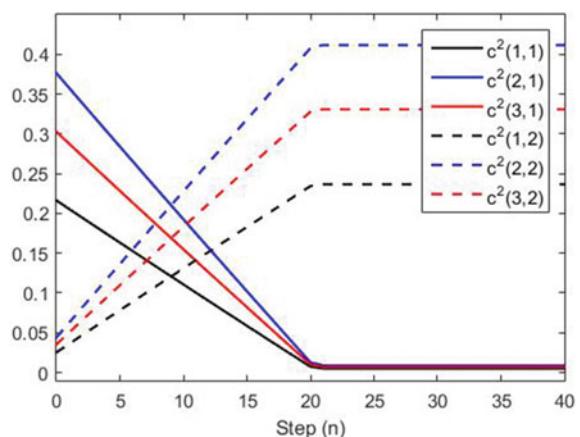


Fig. 12.22 Strategies of player 2 in the bargaining model 2



gies show convergence). Figures 12.21, 12.22 and 12.23 show the behavior of the offers (strategies) during the bargaining process.

Finally, the agreement reached is as follows:

$$c^1 = \begin{bmatrix} 0.2127 & 0.0074 \\ 0.1429 & 0.0050 \\ 0.6106 & 0.0214 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.0050 & 0.2366 \\ 0.0087 & 0.4117 \\ 0.0070 & 0.3310 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.2237 & 0.0071 \\ 0.5877 & 0.0186 \\ 0.1579 & 0.0050 \end{bmatrix}.$$

The mixed strategies obtained for players are as follows

$$d^1 = \begin{bmatrix} 0.9662 & 0.0338 \\ 0.9662 & 0.0338 \\ 0.9662 & 0.0338 \end{bmatrix}, \quad d^2 = \begin{bmatrix} 0.0207 & 0.9793 \\ 0.0207 & 0.9793 \\ 0.0207 & 0.9793 \end{bmatrix}, \quad d^3 = \begin{bmatrix} 0.9693 & 0.0307 \\ 0.9693 & 0.0307 \\ 0.9693 & 0.0307 \end{bmatrix}.$$

Fig. 12.23 Strategies of player 3 in the bargaining model 2

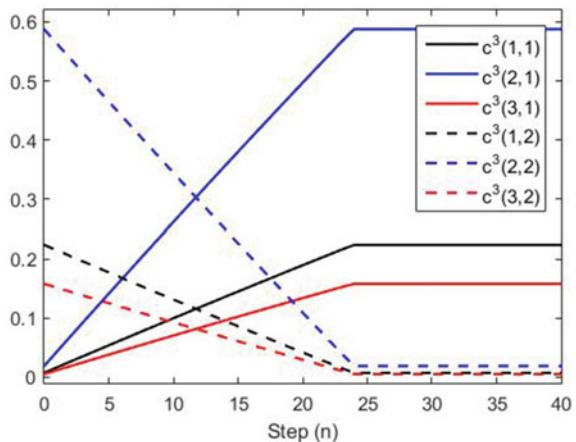
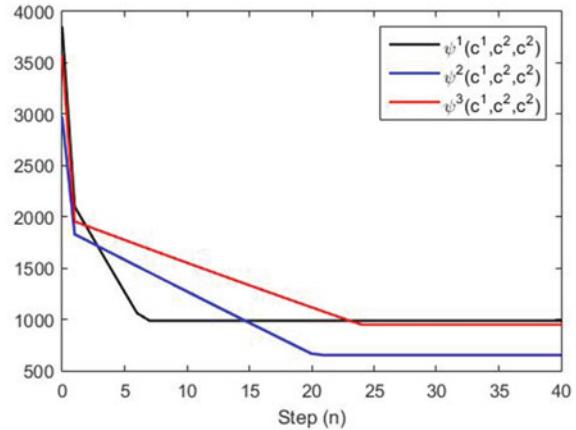


Fig. 12.24 Behavior of players' utilities in the bargaining model 2



With the strategies calculated at each step of the negotiation process, the utilities of each player showed a decreasing behavior as shown in the Fig. 12.24, i.e., at each step of the bargaining process, the utility of each player decreases until they reach an agreement. At the end of the bargaining process, the resulting utilities are as follows $\psi^1(c^1, c^2, c^3) = 986.8936$, $\psi^2(c^1, c^2, c^3) = 651.4633$ and $\psi^3(c^1, c^2, c^3) = 949.6980$ for each player.

12.7.2 Bargaining with Collusive Behavior

In this approach we analyze a bargaining situation where players make groups and alternately each group makes an offer to the others until they reach an equilibrium

point (agreement). We describe a bargaining model with two teams of players as follows. Let us consider a bargaining game with $n + m$ players. Let $n = \{1, \dots, n\}$ denote the set of players called team A and let's define the behavior of all players $l = \overline{1, n}$ as $x_t = (x_t^1, \dots, x_t^n) \in X$ where X is a convex and compact set. In the same way, the rest $M = \{1, \dots, m\}$ players are the team B and let the set of the strategy profiles of all player $m = 1, m$ be defined by $y_t = (y_t^1, \dots, y_t^m) \in Y$ where Y is a convex and compact set. Then, $X \times Y$ in the set of full strategy profiles. In this model the function $\psi(x, y)$ represents the utility function of team A which determines the decision of accept or reject the offer; similarly, team B makes the decision according to its utility function $\varphi(x, y)$.

Following the description of the model presented above, we redefine the advantage of propose a new offer considering the utility function for team A as follows

$$f(x_t, y_t, x_{t+1}, y_{t+1}) := \sum_{l=1}^n [\psi^l(x_{t+1}, y_t) - \psi^l(x_t, y_t)] \geq 0,$$

and, similarly the utility function for team B is as follows

$$g(x_t, y_t, x_{t+1}, y_{t+1}) := \sum_{m=1}^m [\varphi^m(x_t, y_{t+1}) - \varphi^m(x_t, y_t)] \geq 0.$$

Thus, the advantages for team A to reject the offer x_t and to propose a new offer x_{t+1} are given by $A(x_t, y_t, x_{t+1}, y_{t+1}) = \alpha(x_t)T(x_{t+1})f(x_t, y_t, x_{t+1}, y_{t+1})$; in the same way, the advantages for team B to reject the offer y_t and to propose a new offer y_{t+1} are given by

$$A(x_t, y_t, x_{t+1}, y_{t+1}) = \alpha(y_t)T(y_{t+1})g(x_t, y_t, x_{t+1}, y_{t+1}).$$

Remark 12.5 The function $f(x_t, y_t, x_{t+1}, y_{t+1})$ implies the Nash condition

$$\psi^l(x_{t+1}, y_t) - \psi^l(x_t, y_t) \geq 0,$$

for any $x \in X, y \in Y$ and $l = \overline{1, n}$ players.

Remark 12.6 The function $g(x_t, y_t, x_{t+1}, y_{t+1})$ implies the Nash condition

$$\varphi^m(x_t, y_{t+1}) - \varphi^m(x_t, y_t) \geq 0,$$

for any $x \in X, y \in Y$ and $m = \overline{1, m}$ players.

The dynamics of the bargaining game is as follows: at each step of the negotiation process the team A chooses a strategy $x \in X$ considering the utility function $f(x_t, y_t, x_{t+1}, y_{t+1})$, then team B must decide between to accept or reject the offer

calculating a new offer (strategies) $y \in Y$ considering the utility function of the group $g(x_t, y_t, x_{t+1}, y_{t+1})$. Following the description of the model 1, now we have that teams solve the problem in proximal format as follows:

$$\left. \begin{aligned} x^* &= \arg \max_{x \in X} \left\{ -\delta_t T(x) \| (x - x^*) \|^2 + \alpha_t T(x) f(x, y, x^*, y^*) \right\}, \\ y^* &= \arg \max_{y \in Y} \left\{ -\delta_t T(y) \| (y - y^*) \|^2 + \alpha_t T(y) g(x, y, x^*, y^*) \right\}, \end{aligned} \right\} \quad (12.7.5)$$

where

$$f(x, y, x^*, y^*) := \sum_{l=1}^n [\psi^l(x, y^*) - \psi^l(x^*, y^*)],$$

$$g(x, y, x^*, y^*) := \sum_{m=1}^m [\varphi^m(x^*, y) - \varphi^m(x^*, y^*)].$$

At each step, teams make a new offer according to Eq. (12.7.5), both teams solve the bargaining problem together but they reach the equilibrium at different time, the bargaining game continues until the offers (strategies) of all player show convergence.

12.7.2.1 Markov Chains Description

For this model, in the same way that we define the strategies $x \in X$, let us consider a set of strategies denoted by $y^m \in Y^m$ ($m = \overline{1, m}$) where $Y := \bigotimes_{m=1}^m Y^m$ is a convex and compact set,

$$y^m := \text{col } (c^m), \quad Y^m := C_{\text{adm}}^m,$$

where col is the column operator.

Denote by $y = (y^1, \dots, y^m)^\top \in Y$, the joint strategy of the players and $y^{\hat{m}}$ is a strategy of the rest of the players adjoint to y^m , namely,

$$y^{\hat{m}} := (y^1, \dots, y^{m-1}, y^{m+1}, \dots, y^m)^\top \in Y^{\hat{m}} := \bigotimes_{h=1, h \neq m}^m Y^h$$

such that $y = (y^m, y^{\hat{m}})$, $m = \overline{1, m}$.

Consider that players of team A try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $x^* = (x^{1*}, \dots, x^{n*}) \in X$ satisfying for any admissible $x^l \in X^l$ and any $l = \overline{1, n}$

$$f(x, \hat{x}(x)|y) := \sum_{l=1}^n [\psi^l(x^l, x^{\hat{l}}|y) - \psi^l(\bar{x}^l, x^{\hat{l}}|y)] \leq 0, \quad (12.7.6)$$

where $\hat{x} = (x^{\hat{1}\top}, \dots, x^{\hat{n}\top})^\top \in \hat{X} \subseteq \mathbb{R}^{n(n-1)}$ [49, 50], \bar{x}^l is the utopia point defined as Eq. (12.4.6) and $\psi^l(x^l, \hat{x}^l|y)$ is the concave cost-function of player l which plays the strategy $x^l \in X^l$ and the rest of players the strategy $x^i \in X^i$ fixing the strategies $y \in Y$ of team B, and it is defined as Eq. (11.4.15) considering the time function.

Similarly, consider that players of team B also try to reach the Nash equilibrium of the bargaining problem, that is, to find a joint strategy $y^* = (y^{1*}, \dots, y^{m*}) \in Y$ satisfying for any admissible $y^m \in Y^m$ and any $m = \overline{1, m}$

$$g(y, \hat{y}(y)|x) := \sum_{m=1}^m \left[\psi^m(y^m, y^{\hat{m}}|x) - \psi^m(\bar{y}^m, y^{\hat{m}}|x) \right] \leq 0, \quad (12.7.7)$$

where $\hat{y} = (y^{\hat{1}\top}, \dots, y^{\hat{m}\top})^\top \in \hat{Y} \subseteq \mathbb{R}^{m(m-1)}$, \bar{y}^m is the utopia point defined as Eq. (12.4.6) and $\psi^m(y^m, y^{\hat{m}}|x)$ is the concave cost-function of player m which plays the strategy $y^m \in Y^m$ and the rest of players the strategy $y^{\hat{m}} \in Y^{\hat{m}}$ fixing the strategies $x \in X$ of team A, and it is defined as Eq. (11.4.15) considering the time function.

Then, we have that a strategy $x^* \in X$ of team A together with the collection $y^* \in Y$ of team B are defined as the equilibrium of a strictly concave bargaining problem if

$$(x^*, y^*) = \arg \max_{x \in X_{\text{adm}}, y \in Y_{\text{adm}}} \{f(x, \hat{x}(x)|y) \leq 0, g(y, \hat{y}(y)|x) \leq 0\}.$$

We redefine the utility function that depends of the average utility function of all players as follows

$$\begin{aligned} F(x, \hat{x}(x), y, \hat{y}(y)) &:= f(x, \hat{x}(x)|y) + g(y, \hat{y}(y)|x) - \frac{1}{2} \sum_{l=1}^n \sum_{j=1}^N \mu_{(j)}^l h_{(j)}^l(x^l) - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{j=1}^N \mu_{(j)}^m h_{(j)}^m(y^m) - \frac{1}{2} \sum_{l=1}^n \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^l q_{(j|i,k)}^l x_{(i,k)}^l - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \xi_{(j)}^m q_{(j|i,k)}^m y_{(i,k)}^m - \frac{1}{2} \sum_{l=1}^n \sum_{i=1}^N \sum_{k=1}^M \eta^l (x_{(i,k)}^l - 1) - \\ &\quad \frac{1}{2} \sum_{m=1}^m \sum_{i=1}^N \sum_{k=1}^M \eta^m (y_{(i,k)}^m - 1), \end{aligned}$$

then, we may conclude that

$$(x^*, y^*) = \arg \max_{x \in X, \hat{x} \in \hat{X}, y \in Y, \hat{y} \in \hat{Y}} \min_{\mu \geq 0, \xi \geq 0, \eta \geq 0} F(x, \hat{x}(x), y, \hat{y}(y), \mu, \xi, \eta). \quad (12.7.8)$$

Finally, we have that at each step of the bargaining process, players calculate their equilibrium according to the solution of the non-cooperative bargaining problem in proximal format defined as follows

$$\left. \begin{aligned} \mu^* &= \arg \min_{\mu \geq 0} \left\{ -\delta \|\mu - \mu^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu, \xi^*, \eta^*) \right\}, \\ \xi^* &= \arg \min_{\xi \geq 0} \left\{ -\delta \|\xi - \xi^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi, \eta^*) \right\}, \\ \eta^* &= \arg \min_{\eta \geq 0} \left\{ -\delta \|\eta - \eta^*\|^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta) \right\}, \\ x^* &= \arg \max_{x \in X} \left\{ -\delta \|x - x^*\|_\Lambda^2 + \alpha F(x, \hat{x}^*(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\}, \\ \hat{x}^* &= \arg \max_{\hat{x} \in \hat{X}} \left\{ -\delta \|\hat{x} - \hat{x}^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}(x), y^*, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\}, \\ y^* &= \arg \max_{y \in Y} \left\{ -\delta \|y - y^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}^*(x), y, \hat{y}^*(y), \mu^*, \xi^*, \eta^*) \right\}, \\ \hat{y}^* &= \arg \max_{\hat{y} \in \hat{Y}} \left\{ -\delta \|\hat{y} - \hat{y}^*\|_\Lambda^2 + \alpha F(x^*, \hat{x}^*(x), y^*, \hat{y}(y), \mu^*, \xi^*, \eta^*) \right\} \end{aligned} \right\}. \quad (12.7.9)$$

12.7.2.2 Transfer Pricing Simulation

For this example, the team 1 is only formed by player 1 while team 2 is composed of players 2 and 3. Although the players calculate the strategies together following the relation (12.7.9), we consider that players reach the equilibrium at different times. Figures 12.25, 12.26 and 12.27 show the behavior of the offers (strategies) during the bargaining process.

Finally, the agreement reached is as follows:

Fig. 12.25 Strategies of player 1 in the bargaining model 3

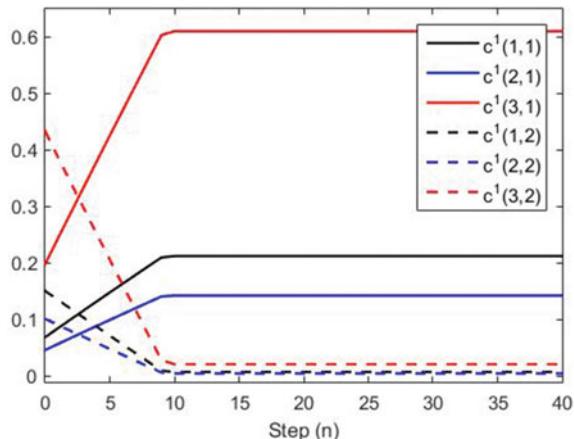


Fig. 12.26 Strategies of player 2 in the bargaining model 3

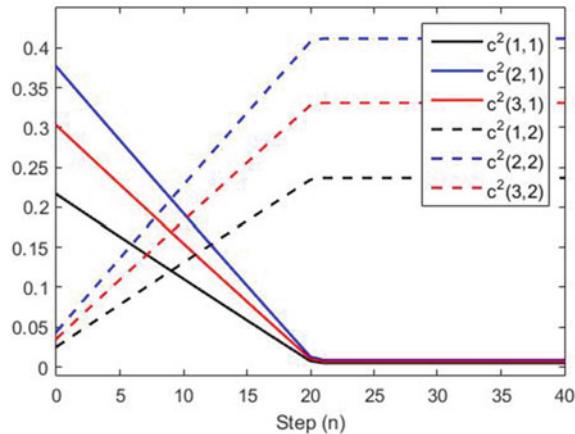
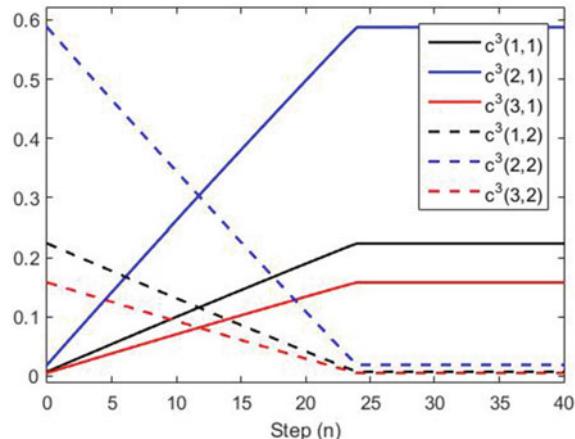


Fig. 12.27 Strategies of player 3 in the bargaining model 3



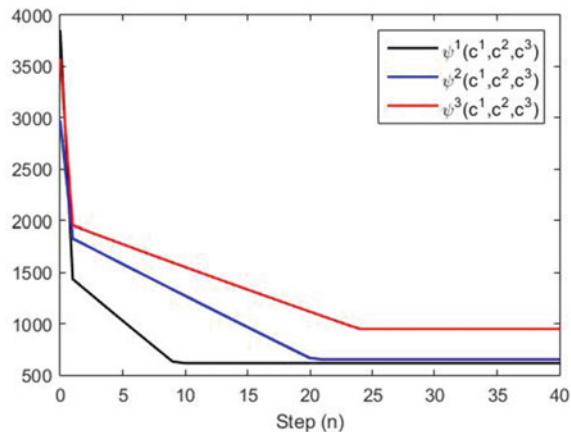
$$c^1 = \begin{bmatrix} 0.2127 & 0.0074 \\ 0.1429 & 0.0050 \\ 0.6106 & 0.0214 \end{bmatrix}, \quad c^2 = \begin{bmatrix} 0.0050 & 0.2366 \\ 0.0087 & 0.4117 \\ 0.0070 & 0.3310 \end{bmatrix}, \quad c^3 = \begin{bmatrix} 0.2237 & 0.0071 \\ 0.5877 & 0.0186 \\ 0.1579 & 0.0050 \end{bmatrix}.$$

The mixed strategies obtained for players are as follows

$$d^1 = \begin{bmatrix} 0.9662 & 0.0338 \\ 0.9662 & 0.0338 \\ 0.9662 & 0.0338 \end{bmatrix}, \quad d^2 = \begin{bmatrix} 0.0207 & 0.9793 \\ 0.0207 & 0.9793 \\ 0.0207 & 0.9793 \end{bmatrix}, \quad d^3 = \begin{bmatrix} 0.9693 & 0.0307 \\ 0.9693 & 0.0307 \\ 0.9693 & 0.0307 \end{bmatrix}.$$

With the strategies calculated at each step of the negotiation process, the utilities of each player showed a decreasing behavior as shown in the Fig. 12.28, i.e., at each step of the bargaining process, the utility of each player decreases until they reach an agreement. At the end of the bargaining process, the resulting utilities are as

Fig. 12.28 Behavior of players' utilities in the bargaining model 3



follows $\psi^1(c^1, c^2, c^3) = 986.8936$, $\psi^2(c^1, c^2, c^3) = 651.4631$ and $\psi^3(c^1, c^2, c^3) = 949.6978$ for each player.

The following figure shows the behavior of the utilities at each of the applied models (model 1 is the general bargaining model, model 2 corresponds to bargaining under different discounting and model 3 to bargaining with collusive behavior), we can see that the utilities begin at the same point, the strong Nash equilibrium, and then decrease until the strategies converge (see Fig. 12.29). From the results obtained we observed that model 1 favors the utilities of players 2 and 3, while model 2 and 3 are better for player 1. We also observed that even if models 2 and 3 reach the same agreement (equilibrium point) the strategies and, as a consequence, the utilities have a different behavior during the bargaining process.

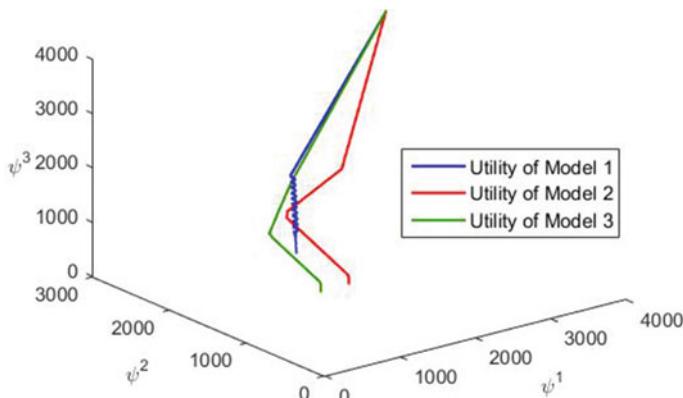


Fig. 12.29 Behavior of the utilities at each model

References

1. Abdel-Khalik, A., Lusk, E.: Transfer pricing - a synthesis. *Account Rev.* **49**(1), 8–23 (1974)
2. Alm, J.: Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. *Int. Tax Public Finance* **19**(1), 54–77 (2012)
3. Amershi, A., Cheng, P.: Intrafirm resource allocation: the economics of transfer pricing and cost allocations in accounting. *Contemp. Account Res.* **7**(1), 61–99 (1990)
4. Arrow, K.J.: Contributions to scientific research in management, chapter. In: Optimization, Decentralization, and Internal Pricing in Business Firms, pp. 9–18. Western Data Processing Center, Graduate School of Business Administration, UCLA (1959)
5. Baldenius, T., Reichelstein, S.: External and internal pricing in multidivisional firms. *J. Account Res.* **44**(1), 1–28 (2006)
6. Baldenius, T., Reichelstein, S., Sahay, S.: Negotiated versus cost-based-transfer pricing. *Rev. Account Stud.* **4**(2), 67–91 (1999)
7. Baumol, W.J., Fabian, T.: Decomposition, pricing for decentralization and external economies. *Manag. Sci.* **11**(1), 1–32 (1964)
8. Beer, S., Loepnick, J.: Profit shifting: drivers of transfer (mis)pricing and the potential countermeasures. *Int. Tax Public Finance* **22**(3), 426–451 (2015)
9. Besanko, D., Sibley, D.S.: Compensation and transfer pricing in a principal-agent model. *Int. Econ. Rev.* **32**(1), 55–68 (1991)
10. Blois, K.J.: Pricing of supplies by large customers. *J. Bus. Finance Account* **3**, 367–379 (1978)
11. Burton, R.M., Damon, W.W., Loughrid, D.W.: The economics of decomposition: re-source allocation versus transfer pricing. *Decis. Sci.* **5**(3), 297–310 (1974)
12. Chalos, P., Haka, S.: Transfer pricing under bilateral bargaining. *Account. Rev.* **65**(3), 624–641 (1990)
13. Chwolka, A., Martini, J.T., Simons, D.: The value of negotiating cost-based transfer prices. *BuR* **3**(2), 113–131 (2010)
14. Clempner, J.B.: Necessary and sufficient karush-kuhn-tucker conditions for multiobjective markov chains optimality. *Automatica* **71**, 135–142 (2016)
15. Clempner, J.B.: Strategic manipulation approach for solving negotiated transfer pricing problem. *J. Optim. Theory Appl.* **178**(1), 304–316 (2018)
16. Clempner, J.B.: Penalizing passenger's transfer time in computing airlines revenue. *Omega*, **97**, 102099 (2020). <https://doi.org/10.1016/j.omega.2019.08.006>
17. Clempner, J.B., Poznyak, A.S.: Convergence method, properties and computational complexity for lyapunov games. *Int. J. Appl. Math. Comput. Sci.* **21**(2), 349–361 (2011)
18. Clempner, J.B., Poznyak, A.S.: Convergence analysis for pure and stationary strategies in repeated potential games: nash, lyapunov and correlated equilibria. *Expert Syst. Appl.* **46**, 474–484 (2016)
19. Clempner, J.B., Poznyak, A.S.: Multiobjective markov chains optimization problem with strong pareto frontier: principles of decision making. *Expert Syst. Appl.* **68**, 123–135 (2017)
20. Clempner, J.B., Poznyak, A.S.: Negotiating the transfer pricing using the nash bargaining solution. *Int. J. Appl. Math. Comput. Sci.* **27**(4), 853–864 (2017)
21. Clempner, J.B., Poznyak, A.S.: Computing the transfer pricing for a multidivisional firm based on cooperative games. *Econ. Comput. Econ. Cybern. Stud. Res.* **52**(1), 107–126 (2018)
22. Dearden, J.: Cast Accounting and Financial Control Systems. Addison Wesley, Reno (1973)
23. Devereux, M., Maffini, G.: The Impact of Taxation on the Location of Capital, Firms and Profit: A Survey of Empirical Evidence. Oxford University Centre for Business Taxation (2007)
24. Edlin, A.S., Reichelstein, S.: Specific investment under negotiated transfer pricing: an efficiency result. *Account Rev.* **70**(2), 275–291 (1995)
25. Enzer, H.: The static theory of transfer pricing. *Nav. Res. Logist. Q.* **22**(2), 375–389 (1975)
26. Forgó, F., Szép, J., Szidarovszky, F.: Introduction to the Theory of Games: Concepts, Methods, Applications. Kluwer Academic Publishers (1999)
27. Fredrickson, J.W.: The strategic decision process and organizational structure. *Acad. Manag. Rev.* **11**(2), 280–297 (1986)

28. Ghosh, P., Roy, N., Das, S.K., Basu, K.: A game theory based pricing strategy for job allocation in mobile grids. In: Proceedings of the 18th International Parallel and Distributed Processing Symposium, pp. 82–92. Santa Fe, New Mexico, USA (2004)
29. Grabski, S.V.: Readings in accounting for management control, chapter. In: Transfer Pricing in Complex Organizations: A Review and Integration of Recent Empirical and Analytical Research, pp. 453–495. Springer US (1982)
30. Haake, C.J., Martini, J.T.: Negotiating transfer prices. *Group Decis. Negot.* **22**(4), 657–680 (2013)
31. Henderson, B.D., Dearden, J.: New system for divisional control. *Harv. Bus. Rev.* **44**(5), 144–146 (1966)
32. Hirshleifer, J.: On the economics of transfer pricing. *J. Bus.* **29**, 172–184 (1956)
33. Hirshleifer, J.: Economics of the divisionalized firm. *J. Bus.* **30**(2), 96–108 (1957)
34. Jennergren, L.P.: The static theory of transfer pricing. *Nav. Res. Logist. Q.* **24**(2), 373–376 (1977)
35. Johnson, N.: Divisional performance measurement and transfer pricing for intangible assets. *Rev. Account Stud.* **11**(2/3), 339–365 (2006)
36. Kanodia, C.: Risk sharing and transfer price systems under uncertainty. *J. Account Res.* **17**(1), 74–75 (1979)
37. Karpowicz, M.P.: Nash equilibrium design and price-based coordination in hierarchical systems. *Int. J. Appl. Math. Comput. Sci.* **22**(4), 951–969 (2012)
38. Leng, M., Parlarb, M.: Transfer pricing in a multidivisional firm: a cooperative game analysis. *Oper. Res. Lett.* **40**(5), 364–369 (2012)
39. Leng, M., Parlarb, M.: Transfer pricing in a multidivisional firm: a cooperative game analysis. *Oper. Res. Lett.* **40**(5), 364–369 (2012)
40. Markides, C.C., Williamson, P.J.: Corporate diversification and organizational structure: a resource-based view. *Acad. Manag. J.* **39**(2), 340–367 (1996)
41. McAulay, L., Scrace, A., Tomkins, C.: Transferring priorities: a three-act play on transfer pricing. *Crit. Perspect. Account* **12**(1), 87–113 (2001)
42. Muthoo, A.: Bargaining Theory with Applications. Cambridge University Press (2002)
43. Nash, J.F.: The bargaining problem. *Econometrica* **18**(2), 155–162 (1950)
44. OECD: OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations 2010. OECD Publishing (2010)
45. Poznyak, A.S.: Advanced Mathematical tools for Automatic Control Engineers. Deterministic technique, vol. 1. Elsevier, Amsterdam (2008)
46. Ronen, J., Balachandran, K.R.: An approach to transfer pricing under uncertainty. *J. Account Res.* **26**(2), 300–3114 (1988)
47. Rosenthal, E.C.: A game theoretic approach to transfer pricing in a vertically integrated supply chain. *Int. J. Prod. Econ.* **115**(2), 542–552 (2008)
48. Rubinstein, A.: Perfect equilibrium in a bargaining model. *Econometrica* **50**(1), 97–109 (1982)
49. Tanaka, K.: The closest solution to the shadow minimum of a cooperative dynamic game. *Comput. Math. Appl.* **18**(1–3), 181–188 (1989)
50. Tanaka, K., Yokoyama, K.: On ϵ -equilibrium point in a noncooperative n-person game. *J. Math. Anal.* **160**(2), 413–423 (1991)
51. Thomas, A.: A behavioral analysis of joint cost allocation and transfer pricing. Technical report, Arthur Andersen & Co. Lecture Series 1977. Stipes Publishing Company (1980)
52. Trejo, K.K., Clempner, J.B.: New perspectives and applications of modern control theory: in honor of Alexander S. Poznyak, chapter. In: Continuous Time Bargaining Model in Controllable Markov Games: Nash versus Kalai-Smorodinsky. Springer International Publishing (2018)
53. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Computing the stackelberg/nash equilibria using the extraproximal method: convergence analysis and implementation details for markov chains games. *Int. J. Appl. Math. Comput. Sci.* **25**(2), 337–351 (2015)
54. Trejo, K.K., Clempner, J.B., Poznyak, A.S.: Nash bargaining equilibria for controllable markov chains games. In: The 20th World Congress of The International Federation of Automatic Control (IFAC), pp. 12772–12777. Toulouse, France (2017)

55. Vaysman, I.: A model of negotiated transfer pricing. *J. Account Econ.* **25**(3), 349–384 (1998)
56. Wahab, O.A., Bentahar, J., Otrok, H., Mourad, A.: A stackelberg game for distributed formation of business-driven services communities. *Expert Syst. Appl.* **45**(1), 359–372 (2016)
57. Watson, D.J.H., Baumler, J.V.: Transfer pricing: a behavioral context. *Account Rev.* **50**(3), 466–574 (1975)
58. Wielenberg, S.: Negotiated transfer pricing, specific investment, and optimal capacity choice. *Rev. Account Stud.* **5**(3), 197–216 (2000)

Index

A

Adaptive policies, 49
Agreements, 255
Allocation rule, 141
Arm-length rule, 290
Arm's length price, 299
Arm's-length principle, 289
Arm's length standard, 290
ASG continuous-time algorithm, 91
Average cost function, 87

B

Banks Marketing Planning, 127
Bargaining game
 collusive behavior, 270
 different discounting, 268
 disagreement penalties, 256
Bargaining paradigm, 185
Bargaining problem, 190, 293
Bargaining solution
 non-cooperative, 264
Bargaining solution payoff, 302
Bayesian incentive-compatible mechanisms, 139
Bayesian-Nash equilibrium, 139, 162
Bayesian partially observable system, 160
Best reply strategies, 115, 119
Birth-death process, 75

C

C variables, 9, 52, 88
Chapman-Kolmogorov equation, 3
Chemical Master Equation, 76

Chemical Reaction Network, 74
Coefficient of ergodicity, 6
Conflict situation, 89
Continuous-time bargaining, 294
Continuous-time bargaining game, 254, 294
Continuous-time Markov decision process, 65, 67
Contracting problem, 150
Controllable Markov chain, 8, 140
Control policy, 8
 randomized, 8
 stationary, 8
Cost function
 average, 9

D

Disagreement point, 208, 302
Disagreement point (status quo), 292, 293
Disagreement vector, 196, 293
Discounted reward, 65
Discount factor, 295
Divisional autonomy, 298
Division game, 300
Divisions, 290
Divisions profit, 296
Division's utility, 297
Division utility function, 300
Duel game, 129
Dynamic of the game, 87

E

Efficient frontier, 30
Equilibrium

Bayesian-Nash, 142
Equilibrium point
stable, 116
Ergodicity constraints, 73, 165
Euler approach, 94
Expected returns, 29, 31, 32
Extraproximal method, 93
Extraproximal procedure, 102

F

Fair agreement, 290
Feasibility, 196
Feasible joint observers, 52, 161
Feasible payoffs, 293
Function vector format, 121

G

Gradient solver, 238

H

Hausdorff distance, 304
Hybrid optimization, 222

I

Independence, 197
Independence of irrelevant alternatives, 189
Independent demands, 296
Indicator function, 147
Individual aim, 89
leader
follower, 101
Intermediate goods, 296
Intersection avoiding, 242
Intra-firm transaction problem, 298
Invariance, 197
Invariant to affine transformations, 189

J

Joint observer, 52, 161
Joint strategy variable, 71

K

Kalai-Smorodinsky, 186
Kalai-Smorodinsky solver, 202
Kiefer-Wolfowitz procedure, 21
Kolmogorov forward equations, 68

L

Lagrange function
regularized, 89
Lineal programming, 11
Linear demand function, 296
Linear programming, 66
solver, 72
Linear scalarization, 19
Local-optimal, 119
Local-optimal (best-reply) strategy, 117
Local-optimal policies, 119
Long-run expected average reward, 65, 70
Lyapunov equilibrium point, 116
Lyapunov function, 116
Lyapunov function design, 125
Lyapunov games, 116, 123
Lyapunov games concept, 116
Lyapunov-like function, 116, 124

M

Markov chain, 1
controllable, 7
ergodic, 4
Markov condition, 1, 140
Markov decision process, 10
optimal, 11
Markov property, 2
Markowitz bullet, 30
Markowitz function, 31
Markowitz portfolio
partially observable, 57
Mean-variance diagram, 32
Mean-variance portfolio, 33
Mechanism, 140, 161
Mechanism design, 137, 156
Mechanisms
admissible, 141
Message, 140
Monotonicity, 198
Multidivisional firm, 290
Multi-objective optimization problem, 18
Myopic policy, 119

N

Nash axioms, 189
Nash bargaining, 196
Nash bargaining issue, 186
Nash bargaining solver, 201
Nash equilibrium, 85, 116
global unique, 89
regularity property, 188

strong, 89
 ε -Nash equilibrium, 90
 ε -Nash equilibrium condition, 90
 Nash's bargaining game, 293
 Nash solution, 294
 Non-cooperative bargaining, 249

O

Observation kernel, 48
 Observation process, 49, 159
 Optimal transfer price, 290

P

Pareto front, 18
 Pareto optimality, 189, 197
 Pareto optimum, 18
 Pareto points
 parametrization, 19
 Pareto policy, 18
 Partially observable Markov chains, 47
 Partially observable Markov decision process, 159
 Partly Observable Markov Games, 156
 Patrolling, 172
 Penalty function, 23, 303
 Poisson distribution, 231
 Policies, 72
 admissible, 142
 POMDP, 48, 49
 Portfolio optimization, 29
 Portfolio return variance, 31
 Pricing, 296
 Principles of fairness, 189
 Prisoner's Dilemma, 127
 Probability space, 1
 Production costs, 297
 Profit allocation approach, 290
 Profit-maximizing allocation, 292
 Projection Gradient Method, 239
 Proximal format, 93
 Proximal method, 304

R

Random walk, 174
 Rational investor, 32
 Rationality, 196
 Recover
 behavior strategies, 164
 mechanism, 163
 observer, 164
 stationary distributions, 164

Recovering relationships, 145
 Reinforcement learning (RL), 146, 167
 Repeated games, 116
 Revelation principle, 143
 Risk, 32
 Risk-averse agents, 150
 Risk-aversion parameter, 58
 Risk-free asset, 33
 Risk-neutral, 187
 Rubinstein bargaining scenario, 250

S

Sales quantity, 296
 Signal controller, 230
 Simplex, 20
 Slack vectors, 303
 Social choice function, 140
 Stackelberg model, 100
 Stackelberg-Nash equilibrium, 102
 Standard deviation, 30
 State-value function, 120, 121
 Stationary distribution, 72
 Stationary distribution vector, 69
 Stochastic matrix, 2
 Stockholder, 58
 Strategy
 behavioral, 142
 Subgame perfect equilibrium, 254, 255, 295
 Supermarkets chain, 108
 Supply chain, 296
 Symmetry, 189, 197

T

Tanaka's function, 90, 101
 regularized, 91, 102
 Tatonnement equilibrium, 257
 Taxes, 297
 Time constraints, 73
 Time-homogeneous, 2
 Traffic optimization, 222
 Traffic-signal control, 222
 Traffic-signal-control problem, 230
 Transfer price, 297
 maximum and minimum, 299
 Transfer price negotiation, 290
 Transfer pricing, 289, 290, 296
 airline alliance, 310
 continuous-time, 310
 non-cooperative bargaining, 312
 under different discounting, 314
 with collusive behavior, 318

Transfer pricing bargaining game, 302

Transfer pricing game, 298

Transfer pricing model, 297

Transfer pricing Nash bargaining, 302

Transition equation, 7

Transition matrix, 2

controlled, 8

Transition probability, 2

Transition rates, 68

U

Uniqueness of the SNE, 96

Unit production cost, 297

Unit simplex, 122

Unsophisticated agent, 250

Urban modeling, 27

Utopia point, 21

V

Valuation function, 140

Vector average cost function, 119

Vehicle routing planning, 27

Vertically integrated divisions, 296

Volatility, 30

W

Weighted objective function, 20