DISSERTATION

in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
from University of Nice Sophia Antipolis

Specialization: Communication and Electronics

Lorenzo Maggi

Markovian Competitive and Cooperative Games with Applications to Communications

Thesis defended on the 9th of October, 2012
before a committee composed of:

Reporters	Prof. Jean-Jacques Herings (Maastricht University, The Netherlands)
	Prof. Roberto Lucchetti, (Politecnico of Milano, Italy)
Examiners	Prof. Pierre Bernhard, (INRIA Sophia Antipolis, France)
	Prof. Petros Elia, (EURECOM, France)
	Prof. Matteo Sereno (University of Torino, Italy)
	Prof. Bruno Tuffin (INRIA Rennes Bretagne-Atlantique, France)
Thesis Director	Prof. Laura Cottatellucci (EURECOM, France)
Thesis Co-Director	Prof. Konstantin Avrachenkov (INRIA Sophia Antipolis, France)

Université Nice SOPHIA ANTIPOLIS

EURECOM Sophia Antipolis

THESE

présentée pour obtenir le grade de

Docteur en Sciences
de l'Université de Nice-Sophia Antipolis

Spécialité: Automatique, Traitement du Signal et des Images

Lorenzo Maggi

Jeux Markoviens, Compétitifs et Coopératifs, avec Applications aux Communications

Thèse soutenue le 9 Octobre 2012 devant le jury composé de :

Rapporteurs	Prof. Jean-Jacques Herings (Maastricht University, Pays Bas)
	Prof. Roberto Lucchetti, (Politecnico of Milano, Italie)
Examinateurs	Prof. Pierre Bernhard, (INRIA Sophia Antipolis, France)
	Prof. Petros Elia, (EURECOM, France)
	Prof. Matteo Sereno (University of Torino, Italie)
	Prof. Bruno Tuffin (INRIA Rennes Bretagne-Atlantique, France)
Directrice de Thèse	Prof. Laura Cottatellucci (EURECOM, France)
Co-Directeur de Thèse	Prof. Konstantin Avrachenkov (INRIA Sophia Antipolis, France)

Abstract

In this dissertation we deal with the design of strategies for agents interacting in a dynamic environment. The mathematical tool of Game Theory (GT) on Markov Decision Processes (MDPs) is adopted. The agents' strategies control both the transition probabilities among the states and the rewards earned by each agent. Rewards are geometrically discounted over time. We first study the competitive case, in which two agents act selfishly. The game is zero-sum and the agents control disjoint sets of states. We devise two algorithms to compute the Nash equilibrium for all discount factors close enough to 1. Then we consider the long-run cooperative case, in which agents can coordinate their strategies. We utilize our two algorithms to compute the value of the coalitions in a routing game, in which several providers share the same network and control the routing in disjoint sets of nodes. Next we deal with dynamic cooperative GT on MDPs, in which coalitions can form throughout the game. We show how to enforce a common agreement for which the pay-off is distributed at each state, in a global optimum way and such that no coalition is ever enticed to break the agreement. We apply these concepts to a wireless multiple access channel, in which the channel is quasi-static. We assign the rate to users in each channel state in a fair and satisfactory manner. Next we provide three methods to compute a confidence interval for Shapley value on Markovian games. Such methods have polynomial complexity in the number of agents, while the complexity of the exact computation is exponential. Two methods are still valid when the values in each state are learned during the game. Finally we assess the performance of two strategies to dynamically select the frequency band to communicate on. We exploit an MDP formulation with uncountable state space.

Résumé

Dans cette thèse, nous étudions la théorie des jeux sur les Processus de Décision de Markov (PDM). Des agents interagissent dans un environnement dynamique, modélisée par une chaîne de Markov. Les stratégies des agents contrôlent les probabilités de transition entre les états et les récompenses gagnées par chaque agent. Les récompenses sont géométriquement pondérées avec le temps. Nous étudions d'abord le cas compétitif. Deux agents agissent égoïstement, le jeu est à somme nulle et les agents contrôlent des ensembles disjoints d'états. Nous élaborons deux algorithmes pour calculer l'équilibre de Nash pour tous les facteurs de pondération suffisamment proche de 1. Puis nous considérons le cas statique coopératif, dans lequel les joueurs peuvent coordonner leurs stratégies. Nous utilisons ces deux algorithmes pour calculer la valeur des coalitions dans un jeu de routage. Plusieurs fournisseurs partagent le même réseau. Ensuite, nous traitons des jeux dynamiques et coopératifs sur PDM. Des coalitions peuvent se former tout au long du jeu. Nous montrons comment distribuer la récompense dans chaque état, de sorte que la solution est optimale pour la communauté et tous les agents sont satisfaits pendant le jeu. Nous appliquons ces concepts à un canal à accès multiple par fil avec un canal quasi-statique. Nous assignons le débit du codage dans chaque état du canal d'une manière équitable. Enfin, nous proposons trois méthodes afin de calculer un intervalle de confiance pour la valeur de Shapley sur les jeux de Markov. Ces méthodes ont une complexité polynomiale en le nombre d'agents, tandis que la complexité du calcul exact est exponentielle. Enfin, nous évaluons la performance de deux stratégies pour sélectionner dynamiquement la bande de fréquence de communication. Nous exploitons une formulation PDM avec un espace d'états dénombrable.

Contents

List of Acronyms

AR(p)	:	Auto-Regressive model of order p
AWGN	:	Additive White Gaussian Noise
Co	:	Core set solution
CM	:	Cooperation Maintaining set solution
DCGT	:	Dynamic Cooperative Game Theory
HMC	:	Homogeneous Markov Chain
MAB	:	Multi Armed Bandit
MAC	:	Mutiple Access Channel
MDP	:	Markov Decision Process
MDP-CPDP	:	Cooperative Pay-off Distribution Procedure on MDP
NTU	:	Non-Transferable Utility
SCGT	:	Static Cooperative Game Theory
Sh	:	Shapley value
ShM	:	Shapley value in Markovian game
SINR	:	Signal to Interference plus Noise Ratio
SNR	:	Signal to Noise Ratio
SSM	:	Shapley-Shubik power index in Markovian game
TU	:	Transferable Utility

List of Notations

$\mathbb{R}_0$	:	$\mathbb{R} \cup \{0\}$
$\mathbb{N}_0$	:	$\mathbb{N} \cup \{0\}$
$:=$	:	defined as
$p(E)$	:	probability of event E
$A_{i,j}$	:	element (i, j) of matrix $\mathbf{A}$
$\mathbf{A}^T$	:	transpose of matrix $\mathbf{A}$
$\mathbf{I}_N$	:	N-by-N identity matrix
$\lvert\mathcal{A}\rvert$	:	cardinality of set $\mathcal{A}$
$\mathbb{I}$	:	indicator function
$\mathbb{E}$	:	expected value operator
$[a; b)$	:	set of $x: \ a \leq x < b$
$\mathcal{S}$	:	finite set of states of a Markov chain
N	:	$\lvert\mathcal{S}\rvert$
S_t	:	state reached at time t by the Markov process
$\mathbf{P}$	:	transition probability of a discrete-time Markov chain
$P_{i,j}$	:	$p(S_{t+1} = s_j \lvert S_t = s_i)$, $s_i, s_j \in \mathcal{S}$, for all $t \in \mathbb{N}_0$
Γ_s	:	long-run cooperative game starting from state s
Ω_s	:	static cooperative game in state s
$\Phi^{(\beta)}$	:	expected long-run β-discounted reward
$\mathcal{P}$	:	grand coalition of all players
$v(\Lambda)$	:	transferable value of players' coalition $\Lambda \subseteq \mathcal{P}$
$P(\mathbb{R})$	:	ring of the polynomials with real coefficients
$F(\mathbb{R})$	:	non-Archimedean ordered field of fractions of polynomials with coefficients in $\mathbb{R}$
$=_l$	:	equality in $F(\mathbb{R})$
$>_l$	:	inequality in $F(\mathbb{R})$

Acknowledgements

This research was supported by “Agence Nationale de la Recherche”, with reference ANR-09-VERS-001.

First and foremost, I would like to express my gratitude to my supervisors, Professors Konstantin Avrachenkov and Laura Cottatellucci. Their passion and enthusiasm for research inspired me throughout these three years. Thanks to them, I realized that doing research is what really fulfils me.

I thank my fellows at Eurecom, Turgut, Xiao Lei, Francesco, Carmelo, Emre, Aymen, Arun, Rajeev, Miltos, Samir, Anthony, Erhan, Umer, Erick, Axel. From each of them I learnt a piece of their culture, and they made this Ph.D. experience unforgettable.

I am deeply indebted to my parents for their unwavering moral support. Last but not the least, I thank Federica for having been such a wonderful life companion.

Introduction

Markov Decision Processes (MDP's) offer an elegant theoretical framework to a number of practical decision making issues, in which the evolution of the system is stochastic, but still depends on the decision taken at each time step by some controller agents, or players. More specifically, there exists a set of states, and in each of them the players have a set of actions at their disposal. The set of actions chosen jointly by the players in a state determines both an instantaneous reward for each of them and a probability distribution on the next state. Typically, the rewards are then either plainly summed over time, or geometrically discounted and then summed, or averaged. The fields in which MDP's found successful applications span from Computer Science, Engineering, Economics, Medicine, to Biology (see White, 1993 [99] for a general survey and Altman, 2002 [3] for applications to communication networks).

MDP's have been extensively studied over the last decades, mostly in the single-agent case. Its origins trace back to Bellman (1957, [18]) and Howard (1960, [42]). Typically, the main goal in single-agent MDP's is to find the optimal decision strategy that the controller has to implement to maximize the long-run sum of rewards (see Puterman, 1994 [77] for a survey). Interestingly, this model can also provide insightful results about the optimality of the controllers' decision with the variation of the horizon length, obtained by tuning the discount factor within $[0;1)$ (see e.g. [41]). Of particular importance is the so-called Blackwell optimality (Blackwell, 1962 [21]), which is the property of those strategies which are optimal for all discount factors sufficiently close to 1. Such strategies happen to be optimal under the average criterion, as well [77].

In this dissertation we will focus mainly on multi-agent MDP's, except for Chapter 4 in which an application of single-agent MDP's will be considered. The bulk of the literature on multi-agent MDP's focuses on the competitive case, in which the agents, or players, are contenders and act selfishly, with the aim of maximizing the long-run sum of individual rewards. Oddly enough, the origin of multi-agent MDP's predates the single-agent case, due to the visionary Shapley's paper in 1953 [81]. Under the competitive assumption, there is no hope that the players can coordinate their actions in order to achieve the social optimum, which coincides with the maximization of the sum of the long-run rewards for each player. Instead, in this scenario the celebrated Nash equilibrium (NE) (Nash, 1950 [64]) is typically utilized to predict the behaviour of conflicting players. It is defined as the set of strategies from which no player can unilaterally deviate by obtaining a benefit. Many of the results available on the literature on competitive MDP's deal with the two-player case (see Filar and Vrieze, 1996 [32] for a survey).

In the same spirit, in Chapter 2 we will deal with two-player Competitive MDP's, also called Stochastic Games, in which the game played in each state is zero-sum, i.e. for each pair of actions for the two players, the sum of rewards earned by the players sums up to zero. Hence, the game is purely antagonistic: no common agreement on any strategy solution can be reached by the players, since any benefit for one player is a loss for the other. In Section 2.1 we will focus on the computation of a pair of strategies at the Nash equilibrium in a two-player zero-sum Competitive MDP. The reward for each player is β-discounted over time, with $\beta \in [0; 1)$. We will assume that, in each state, at most one player has an effective control over it, i.e. the Competitive MDP is with perfect information. In essence, we will consider an MDP in which the control over rewards and transition probabilities switches from a player to the other, depending on the state in which the process finds itself, at each time step. In this scenario, we will devise two algorithms which compute the strategies for both players at the NE, for all discount factors sufficiently close to 1. One of them is proved to converge in a finite time. To do so, we will basically combine two techniques. The former is by Raghavan and Syed (2003, [78]), who provided an algorithm to compute the strategies at NE in the same scenario described above, for a *fixed* discount factor. The latter is by Hordijk, Dekker, and Kallenberg (1985, [41]), who first utilized linear programming on the field of rational functions with real coefficients to compute Blackwell optimal strategies in single-agent MDP's. Moreover, our algorithms produce the interval $[\beta^*; 1)$ in which the strategies are optimal, i.e. at NE. Thanks to the special structure of the Competitive MDP, such strategies are also optimal under the average criterion.

In Section 2.4 we show a possible application of the algorithms described in Section 2.1 to the study of a routing scenario by using Static Cooperative Game Theory (SCGT) with transferable utility (TU) assumption. In SCGT, the players can make binding agreements both on the adopted strategy and on the sharing procedure of the resulting pay-off, which, under the TU assumption, can be shared in any manner among the players. Potentially, each subset of players can form a coalition. The objective of static Cooperative Game Theory is two-fold: a social optimum solution has to be found, and the maximum pay-off has to be shared in a satisfactory way for all the play-

ers. Hence, all the players need to agree on a common contract, which has to be profitable, or at least fair, for all of them. Several pay-off allocation criteria that have been studied over the years, like the Core, Shapley value, Nucleolus, τ-value etc. (see [69] for a survey).
Typically, in SCGT the pay-off allocation is a function of the *value* of each coalition, which is the pay-off that a coalition can earn on its own, without the ability of coordinating its actions with other players. Such value can be computed *à la* Morgenstern-von Neumann [96], i.e. as the minimum cost (or maximum pay-off) that a (sub-)coalition can guarantee if the anti-coalition punishes it by adopting an adverse behaviour. Therefore, the game between a coalition and the anti-coalition is still a zero-sum game.

Now, let us describe our routing game of Section 2.4. We consider a system where several providers share the same network and control the routing in disjoint sets of nodes; each link has a different cost for each provider. They provide connection toward a unique server (destination) to their customers. In order to carry out a successful transmission, they need to cooperate and coordinate their own routing strategies so that the global transmission cost is minimized. Hence, in this case we assume that the players, i.e. the service providers, do not have a conflicting behaviour. Indeed, a selfish behaviour would backfire on each service providers, who needs the help of the other to deliver its packets. In this scenario, the transmission costs need to be shared among the service providers. If we translate the scenario into the MDP jargon, the states of the associated MDP are the nodes, and the actions for each player are the routing decision in each node, and the rewards associated to an action are the cost of the selected link. We study this situation by utilizing SCGT, and we would like to be able to allocate a cost to each provider, which depends on their ability to cooperate with the others and carry out a successful transmission. For this purpose, we need to compute the value for each coalition of providers, and we opt for a max-min approach. Hence, the same model introduced in Section 2.1 emerges, and we show how to adapt the algorithms described in Section 2.1 to compute the value of each coalition of providers.

We remark that, in the cooperative model described in Section 2.4, the interaction among the players is one-shot: the value of each coalition is computed off-line, once for all, the transmission costs are shared among the players at the beginning or at the end of the game indifferently, and coalitions cannot form once the transmission has started. This is the reason why we call this approach Long-Run Cooperative Game.
A different scenario is depicted in Chapter 3, in which the interaction among the players continues over time, on each different step of a Markov Decision Process, and coalitions may form throughout the game. We dub this scenario Cooperative MDP.

Our work on Cooperative MDP's fits in the more general context of

Dynamic Cooperative Game Theory (DCGT), which has been one of the most innovative and interesting topics in the field of Game Theory in the last few years. The bulk of the research on Cooperative Games, spurred by Morgenstern and von Neumann (1953, [96]), has focused ever since on the study of Static Games, modelling one-shot interactions among players.

Nevertheless, most of the interaction situations among players, e.g. countries, firms, or users in a communication systems, are not one-shot but continue over time, and the environment in which they take place is dynamic. This motivation stimulated the research on DCGT. The agreement is stipulated by the players at the beginning of the game, once for all, mainly for two reasons. Firstly, renegotiating a contract at each step in costly, in time and money. Secondly, revising an agreement at each time step may result in a myopic policy, that does not take into account the social optimum in the long run. As an example, curbing the emission of pollutants requires investment in cleaner technologies, that can be made only if the economic agents commit themselves to a far-sighted investment strategy.
In this dynamic context, reaching a common agreement which is enduring over time constitutes the real challenge, since coalitions are allowed to form throughout the game. One of the most important notions in dynamic games is the Time Consistency of a pay-off sharing (Petrosyan, 1977 [72]). According to it, at each intermediate step of the interaction process, the pay-off distribution in the subgame from that instant onwards must respect the same fairness criterion under which the agreement has been stipulated in the first place. A more specific property for a pay-off dynamic allocation is the sequential Core, for which no party should not be enticed to breach the agreement at any time step, preferring to adopt a more profitable non-cooperative mode of play from that instant onward. Kranich, Perea, and Peters (2005, [48]) introduced the weak sequential Core, restricting the focus to credible deviations, i.e. deviations cannot be counter-blocked by any subcoalition. Another interesting concept for DCGT is Cooperation Maintenance, for which, at each time step, each coalition should be persuaded to postpone the decision of breaching the agreement. By induction, hence, the social agreement is stable over time. Mazalov and Rettieva (2010, [59]) first introduced this property in a fish-war setting.

Over the last decade, the research on DCGT has ramified into different branches. The first one is on Repeated Cooperative Games, in which the *same* game is played repeatedly over time. The papers by Oviedo (2000, [66]) and by Kranich, Perea, and Peters (2000, [47]) are the two independent pioneering works in this field. A second research branch deals with different states that succeed each other, and in each of them a *different* static game is played. The state transition accounts for the dynamics of the environment in which the interaction takes place. Kranich, Perea, and Peters (2005, [48]) studied the Core concept solution for these kinds of games, in which the states succession is pre-determined. Predtetchinski (2007, [75]) considered that an endogenous Markov chain determines the probability of transition among the states. A third major research thread in DCGT is on Differential Games (see Zaccour 2008, [102] for a survey), in which the state of the game evolves continuously over time according to a deterministic differential equation, controlled by the players' strategies.

DCGT finds many applications in Economics and, not surprisingly, the connections between the two fields is interwoven. DCGT has capitalized on some concepts already existing in the vast Economics literature. For example, a concept similar to the Time Consistency property has been elaborated by the Nobel prize 2004 winners Kydland and Prescott (1977, [49]), about the strategy that a policymaker has to implement in order to trigger the desired response from the economic agents. Gale (1978, [35]) was the first to introduce the idea of strong sequential Core, in a monetary economy model. Moreover, the notion of Core in an exchange economy with incomplete information has been thoroughly studied over the last few decades (see Forges et. al 2002, [34] for a survey), much before the theoretical foundations of DCGT have been laid down.

On the other hand, the research on DCGT has been spurred by the real economic issues, and in the last few years the recent advances on DCGT have represented a valid tool for economists. Predtetchinski, Herings, and Peters (2002, [76]) studied the strong sequential Core in two-stage economies in which the trade in assets takes place at period zero and the trade in commodities occurs at period one. Herings, Predtetchinski, and Perea (2006, [38]) studied the weak sequential Core in the same context.

We apply DCGT for the first time to a communication network model, in Section 3.2.

Now, let us describe our TU Cooperative MDP model. It is essentially a multi-agent MDP model in which the agents, or players, are allowed to form coalitions throughout the game. Let us analyse first the long-run game on MDP in a Static Cooperative Game perspective. If we look at the game as a whole, we can still compute off-line the global optimum strategies for the grand coalition of players, via classic optimization techniques for single-agent MDP's, in which the grand coalition is considered to be the agent. In the long-run game, the value of each coalition is computed, as the long-run sum of rewards that each coalition can attain on its own, and a cooperative solution is assigned in the long-run game. So far, this formulation still relates to SCGT. Nevertheless, two issues arise. Firstly, the long-run Cooperative solution is actually the expected value of a random variable, hence, pragmatically, it is not clear how to allocate it to the players. Secondly, since the horizon of the game is infinite, or finite or with unknown duration, players may demand to be rewarded at *each* stage of the game, i.e. in each step of the controlled Markov chain. Therefore, the challenge becomes to devise a Cooperative Pay-off Distribution Procedure (CPDP) which distributes the long-run Cooperative solution throughout the game, in each of its states. From this seemingly harmless compromise, a number of issues arise. Since we assume that coalitions may form throughout the game, the CPDP has to content all the players, throughout the game. Indeed, if a long-run agreement has been stipulated by the players at the beginning of the game according to some (e.g. fairness) criteria, it is not clear whether the agreement can be sustained over time, i.e. whether such criterion keeps holding after some steps, or a renegotiation is needed. Real life situations abound with examples in which a long-term contract signed in the first place needs to be renegotiated over time, e.g. the current economic situation in the European Union. The property that a CPDP needs to possess in order to avoid a renegotiation is Time Consistency (Petrosjan 1977, [72]). A second major issue is on the stability of the grand coalition, throughout the game. A (sub-)coalition might be enticed to breach the agreement at some point of the dynamic game, since it can guarantee from that instant onwards a better allocation on its own. In order to avoid this, the CPDP needs to belong to the sequential Core (Kranich, Perea, and Peters 2005, [48]). Intuitively, this property claims that, whenever a coalition faces the dilemma "*do we break the agreement now or we cooperate forever?*", then it should always opts for the second option, since more profitable. Moreover, we demand that the CPDP satisfies the Cooperation Maintenance property (Mazalov and Rettieva 2010, [59]). Intuitively, this property suggests that, at each time step, if any coalition faces the dilemma "*do we break the agreement now or in one step?*", then it should choose the second option. Finally, we consider the presence of greedy players, having a myopic perspective of the game. In this case, the CPDP needs to belong to the Core of each static game played in each state of the MDP, in order to content the greedy players as well. Then, we devise a CPDP for MDP's, dubbed MDP-CPDP, and we find conditions for which all the properties enlisted above are fulfilled. Remarkably, we find that the Cooperation Maintenance property is a proper refinement of the concept of sequential Core on Cooperative MDP's.

In Section 3.2 we apply some of the concepts developed in Section 3.1 to a wireless communication scenario. We consider a Gaussian Multiple Access Channel (MAC) in which the channel is quasi-static, i.e. it varies slowly enough to be assumed constant for the whole duration of a codeword. Moreover, we assume that it follows an endogenous finite state Homogeneous Markov Chain (HMC) in which the state transitions occur at the end of each coherence period. We allocate a rate to each user in each state of the Markov chain. We stress that, in this scenario, the transition probabilities among the channel states of the system do not depend on the users' transmission strategies. In this sense, the scenario is simplified with respect to the one in Section 3.1. On the other hand, a new complication is brought on by the

introduction of the Non-Transferable Utility (NTU) assumption, for which the rate cannot be shared in any manner among the users, but only within the Shannon Capacity region.
In this scenario we tackle the issue of allocating the rate in each state, in a global optimum way, i.e. such that the sum-rate is maximum both in each state and in the long-run process. We call $\mathcal{M}$ this global optimal set. We investigate two procedures to select an allocation in $\mathcal{M}$: the former, called bottom-up, prescribes to allocate first the static allocations, while the latter, dubbed top-down, suggests to select first the long-run allocations, and then to derive the associated static allocations. The latter procedure would be more useful, since it permits to adhere to a selection criterion in the long-run process, being the one really concerning the users, which are endowed with a long-term perspective of the game. Unfortunately, this procedure does not always lead to feasible allocations, and we offer a remedy to it. Then we address the issue of allocating a fair rate to the users in the dynamic process. In the static case, the criteria of max-min, proportional, and α-fairness are always well defined [5]. In our dynamic case, the scenario is complicated by the the fact that we demand that the fairness criterion needs to hold throughout the process, i.e. it has to be time consistent. It is not always possible to find a time consistent fair allocation; nevertheless, we give a sufficient condition for its existence. Then, we utilize some tools developed in Section 3.1 to measure the users' satisfaction with the rate assigned. We find that $\mathcal{M}$ coincides with all the allocation rates belonging to the sequential Core. Moreover, $\mathcal{M}$ is exactly the set of Cooperation Maintaining solutions.

In Section 3.3 we deal with Cooperative MDP's under the TU assumption, in which the transition probabilities among the states do not depend on the players' actions. Hence, like in Section 3.2, an endogenous Markov chain governs the stochastic evolution of the state of the system, so we dub this scenario as Markovian TU cooperative game.
We deal with this model in a perspective different from Sections 3.1, 3.2. In fact, we do not study the pay-off allocation in each state, since, for a linearity argument, this is a trivial issue. Instead, we tackle a complexity issue relative to the computation of the Shapley value (Shapley 1953, [82]) in these kind of games. More specifically, we provide three methods to compute a confidence interval for the Shapley-Shubik index in Markovian games (SSM). We extend the approach of Bachrach et al. in 2010 [14] for static games to Markovian games. The Shapley-Shubik index (Shapley and Shubik, 1954 [85]) is the Shapley value applied to a simple game, i.e. a game in which the coalition values are binary, i.e. 1 or 0. The Shapley-Shubik index proves to be particularly suitable to assess *a priori* the power of the members of a legislation committee, and has many applications to politics (see Taylor and Pacelli 2008, [91] for an overview). In our case, the game in each state is simple. We prove that an exponential number of queries

is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Motivated by this, we propose three different randomized approaches to compute a confidence interval for SSM. Their complexity does not even depend on the number of players. Such approaches also hold for the classic Shapley value of any cooperative Markovian game. The first confidence interval, SCI, relies on the static assumption that the estimator agent has access to the coalition values in all the states at the same time, even before the Markov process initiates. Although SCI relies on an impractical assumption, it is still a valid benchmark for the performance of the approaches yielding the other two proposed confidence intervals, dubbed DCI1 and DCI2. DCI1 and DCI2 both hold also under the more realistic dynamic assumption that the estimator agent learns the value of coalitions along the course of the game. We propose a straightforward way to optimize the tightness of DCI1 and we compare the three proposed approaches in terms of tightness of the confidence interval. Finally, we provide a trade-off complexity/accuracy of our randomized algorithm, holding for any cooperative Markovian game.

Finally, in Chapter 4 we utilize an MDP formulation with an uncountable state space to solve a dynamic selection problem among different transmission channels. In the learning algorithms literature, this approach goes under the label of Multi Armed Bandit (MAB), in which there exists a pool of several random processes, or arms, that can be observed once at a time. The total reward is the (discounted or averaged) sum of the instantaneous values of the observed arms. Based on the past observations and on the statistical knowledge *a priori* of each arm, the goal is to maximize the expected total reward.

The so-called Rested MAB assume that the state of an arm which has not been pulled does not evolve in the next step. In this case, a brilliant optimization solution was found by Gittins (1989, [37]). He observed that the curse of state dimensionality in the number of arms caused by an MDP formulation can be overcome with the computation of an index, for each arm. The optimal solution consists in simply selecting the arm with the highest index, in each decision step. Thus, the complexity of the solution boils down from exponential to linear in the number of arms.

We deal instead with Restless MAB, in which the arms evolve even when not observed, which is of course the case of the attenuation coefficients of a transmission channel. In this case, an efficient solution has not been found yet. Whittle (1988, [100]) proposed an index whose optimality is not guaranteed though. It was proven by Papadimitriou and Tsitsiklis in [67] that restless MAB are PSPACE-hard in general. Hence, in order to deal with Restless MAB's, one needs to resort to a MDP formulation with uncountable state space, or equivalently to a Partially Observable MDP model. Typically, the state of the decision problem at each time step is the instantaneous value of each arm in the last time step. Since the state of the unobserved arms is unknown, the decision agent has at its disposal only its probability distribution, conditioned on its last observation.

In our case, the processes to be selected at each time are independent autoregressive processes of order 1, i.e. AR(1). They represent the slow fading channel attenuations on different frequency bands of a multi-access wireless network that a transmitter can utilize to communicate on. The goal is to maximize the long-run average Signal-to-Noise Ratio (SNR) for the user. Note that the user is assumed to know all the parameters of the AR(1) processes, hence also their (unconditioned) expected value. A naïve solution would prescribe to select invariably the channel with highest expected value. Roughly speaking, this procedure is highly suboptimal if the expected values are similar to each other. Indeed, it is possible to exploit the autocorrelation of the channels to select the one which is *instantaneously* the best one.

We propose two heuristic algorithms with linear complexity in the number of arms to solve the Restless MAB problem. The former, myopic, suggests to select the channel with the highest conditioned expected reward, at each time step. The latter, randomized, selects the seemingly sub-optimal arms with a certain probability. We find that the myopic strategy performs close to the optimal solution when all the arms are almost statistically equivalent. When one channel is characterized with a much higher autocorrelation, then the randomized approach outperforms the myopic approach and is quasi-optimal. We then propose a Competitive MDP formulation, in which different users access to the same pool of channels. We adapt the cited two algorithms to approximate the best response of a user against the strategy of a second user which is oblivious of the presence of the first.

Introduction (en français)

Les Processus de Décision Markoviens (MDP) offrent un cadre théorique très utile pour un certain nombre de problèmes décisionnel pratiques, dans lesquelles l'évolution du système stochastique dépend cependant de la décision prise par certains agents contrôleurs (ou joueurs). Plus précisément, il existe un ensemble d'états, et dans chacun d'eux les joueurs ont un ensemble d'actions à leur disposition. L'ensemble des actions choisies conjointement par les joueurs dans un état détermine à la fois une récompense instantanée pour chacun d'eux et une distribution de probabilité sur les états suivants. En règle générale, les récompenses sont soit additionnées dans le temps, ou actualisées géométriquement et ensuite additionnées, ou autrement on en prends la moyenne. Les domaines dans lesquels MDP a trouvé des applications réussies vont de l'informatique, de l'ingénierie, de l'économie, la médecine, la biologie (voir White, 1993 [99] pour une vue d'ensemble et Altman, 2002 [3] pour des applications aux réseaux de communication).

Les MDP ont été largement étudiés au cours des dernières décennies, surtout dans le cas mono-agent. Ses origines remontent à Bellman (1957, [18]) et Howard (1960, [42]). En règle générale, l'objectif principal des MDP avec mono-agent est de trouver la stratégie optimale de décision que le contrôleur doit mettre en oeuvre afin de maximiser la somme à long terme de récompenses (voir Puterman, 1994 [77] pour une vue d'ensemble). Il est intéressant de noter que ce modèle peut également fournir des résultats pertinentes sur l'optimalité de la décision des contrôleurs quand la longueur de l'horizon varie, obtenus en ajustant le facteur d'actualisation entre $[0; 1)$ (voir par exemple [41]). L'optimalité de Blackwell (Blackwell, 1962 [21]) est particulièrement important. Elle est la propriété de ces stratégies qui sont optimales pour tous les facteurs d'actualisation suffisamment proche de 1. Aussi, ces stratégies se révèlent être optimale selon le critère de la moyenne [77].

Dans cette thèse, nous nous concentrerons principalement sur les systèmes MDP avec multi-agents, sauf pour le Chapitre 4, dans lesquels nous ètudions une application des MDP avec mono-agent. La majeure partie de la littérature sur les systèmes MDP avec multi-agents se concentre sur la situation concurrentielle dans laquelle les agents ou les joueurs agissent égoïstement, dans le but de maximiser la somme à long terme des récompenses individuelles. Curieusement, l'origine des systèmes MDP avec multi-agents est antérieure le cas mono-agent, en raison de papier visionnaire de Shapley en 1953 [81]. Dans l'hypothèse competitive, il n'ya aucun espoir que les joueurs puissent coordonner leurs actions afin d'atteindre l'optimum social, qui coïncide avec la maximisation de la somme des récompenses à long terme pour chaque joueur. Au lieu de cela, dans ce scénario, le célèbre équilibre de Nash (NE) (Nash, 1950 [64]) est généralement utilisé pour prédire le comportement des acteurs en conflit. Aucun jouer ne peut modifier seul sa stratégie de Nash sans affaiblir sa position personnelle. La plupart des résultats disponibles dans la littérature sur les MDP compétitives traitent avec le cas de deux joueurs (voir Filar et Vrieze, 1996 [32] pour une vue d'ensemble).

Dans le même esprit, dans le Chapitre 2 nous allons étudier les MDPs compétitifs avec deux joueurs, aussi appelés "jeux stochastiques", dans lesquels le jeu pratiqué dans chaque état est à somme nulle. Ça signifie que pour chaque paire d'actions pour le deux joueurs, la somme des récompenses gagnées par les joueurs est nulle. Par conséquent, le jeu est purement antagoniste: pas de consensus sur les stratégies peut être atteint par les joueurs, puisque un profit pour un joueur est une perte pour l'autre. Dans la Section 2.1 nous allons nous concentrer sur le calcul d'une paire de stratégies à l'équilibre de Nash dans un MDP compétitif avec deux joueurs et à somme nulle. La récompense pour chaque joueur est β-actualisée au fil du temps, avec $\beta \in [0, 1)$. Nous supposons que, dans chaque état, au plus un joueur a un contrôle effectif, i.e. le MDP est á information parfaite. Nous allons considérer un MDP dans lequelle le contrôle sur les récompenses et les probabilités de transition passe d'un joueur à l'autre, en fonction de l'état dans lequel se trouve le processus. Dans ce scénario, nous allons mettre au point deux algorithmes qui calculent les stratégies pour les deux joueurs à la NE, pour tous les facteurs d'actualisation suffisamment proche de 1. Nous prouvons que l'un d'eux converge en un temps fini, en combinant essentiellement deux techniques. Le premier est par Raghavan et Syed (2003, [78]), qui fournit un algorithme pour calculer les stratégies au NE dans le même scénario décrit ci-dessus, pour un facteur d'actualisation *fixe*. Ce dernier est défini par Hordijk, Dekker et Kallenberg (1985, [41]), qui ont utilisé la programmation linéaire sur le corps des fonctions rationnelles à coefficients réels pour calculer le stratégies optimales de Blackwell pour MDP avec mono-agent. En outre, nos algorithmes calculent l'intervalle $[\beta^*; 1)$ dans lequelle les stratégies sont Nash optimales. Grace à la structure particulière des MDP compétitifs, ces stratégies sont également optimal selon le critère de la moyenne.

Dans la Section 2.4, nous montrons une éventuelle application des algorithmes décrits dans la Section 2.1 pour l'étude d'un scénario de routage statique en utilisant la théorie des jeux coopérative (SCGT) sous l'hypothèse

de utilité transférable (TU). En SCGT, les joueurs peuvent faire des accords contractuels à la fois sur la stratégie adoptée et sur la procédure de partage du résultant payoff, ce qui, dans l'hypothèse TU, peut être partagé en aucune manière parmi les joueurs. Potentiellement, chaque sous-ensemble de joueurs peut former une coalition. L'objectif de la théorie des jeux statiques coopératifs est double: une solution optimum social est à déterminer, et le maximum pay-off doit être partagée de manière satisfaisante pour tous les joueurs. Par conséquent, tous les joueurs doivent se mettre d'accord sur un contrat commun, qui doit d'être avantageux, ou du moins juste, pour chacun d'eux. Plusieurs critères d'attribution du payoff qui ont été étudiés au cours des années, comme le Core, la valeur de Shapley, le nucléolus, la τ-valeur, etc. (voir [69] pour une vue d'ensemble).
En règle générale, en SCGT le pay-off est une fonction de la *valeur* de chaque coalition, c'eat à dire la récompense qu'une coalition peut gagner toute seule, sans la possibilité de coordonner ses actions avec les autres joueurs. Cette valeur peut être calculée *à la* Morgenstern-von Neumann [96], i.e. le coût minimum (ou le payoff maximum) qu'une coalition peut garantir si l'anti-coalition la punit en adoptant un comportement antagoniste. Par conséquent, le jeu entre une coalition et la anti-coalition est toujours à somme nulle.

Maintenant, nous allons décrire notre jeu de routage de la Section 2.4. Nous considérons un système où plusieurs fournisseurs qui se partagent le même réseau et contrôlent le routage dans des ensembles disjoints de noeuds et chaque lien a un coût différent pour chaque fournisseur. Ils fournissent une connexion vers un serveur unique (destination) à leurs clients. Afin de réaliser une transmission réussie, ils ont besoin de coopérer et de coordonner leurs stratégies de routage de manière que le coût global de transmission est réduit au minimum. Par conséquent, dans ce cas, nous supposons que les joueurs, i.e. les fournisseurs de services, n'ont pas un comportement antagoniste. En effet, un comportement égoïste aurait un effet contre-productif pour chacun des fournisseurs de services, qui ont besoin de l'aide des autres pour livrer ses paquets. Dans ce scénario, les coûts de transmission doivent être partagés entre les prestataires de services. Si nous traduisons le scénario dans le jargon MDP, les états du MDP associé sont les noeuds et les actions de chaque joueur sont les décisions de routage dans chaque noeud, et la récompense associée à une action est le coût du lien sélectionné. Nous étudions cette situation en utilisant SCGT pour attribuer un coût à chaque fournisseur, qui dépend de leur capacité à coopérer avec les autres et de réaliser une transmission réussie. Pour cette raison, nous avons besoin de calculer la valeur de chaque coalition de fournisseurs, et nous optons pour une approche max-min. Par conséquent, le même modéle présenté dans la Section 2.1 apparait, et nous montrons comment adapter les algorithmes décrits dans la Section 2.1 pour calculer la valeur de chaque coalition de fournisseurs.

On remarque que, dans le modèle coopératif décrit dans la Section 2.4, l'interaction entre les joueurs est ponctuelle: la valeur de chaque coalition est calculée hors-ligne, une fois pour toutes, les coûts du transport sont partagés entre les joueurs au début ou à la fin du jeu, indifféremment, et des coalitions ne peuvent pas former quand la transmission a commencée. C'est la raison pour laquelle nous appelons cette approche “jeu coopératif à long terme”.
Un scénario différent est décrit dans le Chapitre 3, dans lequel l'interaction entre les joueurs se poursuit au fil du temps, à chaque étape d'un processus décisionnel de Markov, et des coalitions peuvent se former tout au long du jeu. Nous appélons ce scénario “MDP coopératif”.

Notre travail sur les MDP coopératifs s'inscrit dans le contexte plus général de la théorie des jeux coopératifs dynamiques (DCGT), qui a été l'un des sujets les plus novatrices et intéressantes dans le domaine de la théorie des jeux dans les dernières années. La majeure partie de la recherche sur les jeux coopératifs, stimulée par Morgenstern et von Neumann (1953, [96]), a mis l'accent sur l'étude des jeux statiques, qui modélisent les interactions ponctuelles entre les joueurs.

Néanmoins, la plupart des situations d'interaction entre les joueurs, e.g. pays, entreprises ou utilisateurs dans un système de communication, ne sont pas ponctuelles, mais continuent au fil du temps dans un environnement dynamique. Cette motivation a stimulé la recherche sur DCGT. L'accord est établi par les joueurs au début du match, une fois pour toutes, principalement pour deux raisons. Tout d'abord, la renégociation d'un contrat à chaque étape est coûteuse, en temps et en argent. Deuxièmement, la révision d'un accord à chaque étape temporelle peut amener à une politique myope, qui ne tient pas compte de l'optimum social à long terme. A titre d'exemple, dans la lutte contre les émissions de polluants, il faut investir dans les technologies propres, qui ne peut être faite que si les agents économiques eux-mêmes s'engagent dans une stratégie d'investissement clairvoyante.

Dans ce contexte dynamique, parvenir à un accord commun qui est durable dans le temps constitue le véritable défi, puisque les coalitions sont autorisées à former tout au long du jeu. L'une des notions les plus importantes dans les jeux dynamiques est la cohérence dans le temps d'un partage du payoff (Petrosyan, 1977 [72]). Selon lui, à chaque étape intermédiaire du processus d'interaction, la distribution du payoff dans le sous-jeu à partir de cet instant doit respecter le critère de l'équité même en vertu du quel le contrat a été stipulé en premier lieu. Une propriété plus spécifique pour une allocation dynamique du payoff est le Core séquentiel, pour laquelle aucun parti ne devrait pas être tenté de violer l'accord à n'importe quel instant temporel, à partir de ce moment préférant adopter un approche non coopératif plus rentable. Kranich, Perea, et Peters (2005, [48]) ont introduit le Core faible séquentielle, qui restreint la rechercheà des écarts crédibles, i.e. les écarts qui ne peuvent pas être contre-bloquée par une sous-coalition. Un autre concept intéressant pour DCGT est celui de la maintenance de la coopération, pour lesquel, à chaque étape temporelle, chaque coalition devrait être persuadée à différer la décision de violer l'accord. Par induction, l'accord social est stable dans le temps. Mazalov et Rettieva (2010, [59]) ont introduit pour la prémière fois cette propriété dans un cadre de "fish-war".

Au cours de la dernière décennie, la recherche sur DCGT s'est ramifiée en différentes branches. Le premier est sur les jeux coopératifs répétés, dans lesquels le *même* jeu se joue à plusieurs reprises au fil du temps. Les articles par Oviedo (2000, [66]) et par Kranich, Perea, et Peters (2000, [47]) sont les deux travaux pionniers indépendants dans ce domaine.

Une autre branche de recherche s'occupe de différents états qui se succèdent, et dans chacun d'eux un jeu statique *différent* est joué. Les transition d'états réprésent la dynamique de l'environnement dans lequel l'interaction a lieu. Kranich, Perea, et Peters (2005, [48]) ont étudié le concept de solution du Core pour ces jeux, dans lesquels la succession des états est prédéterminée. Predtetchinski (2007, [75]) a considéré qu'une chaîne de Markov endogène détermine la probabilité de transition entre les états. Un troisième axe majeur de recherche en DCGT est répresenté par les jeux différentiels (voir Zaccour 2008, [102] pour une vue d'ensemble), dans lesquels l'état du jeu évolue continuellement au fil du temps selon une équation différentielle déterministe, contrôlée par les stratégies des joueurs.

DCGT s'applique à nombreuses applications en économie, et sans surprise, les connexions entre les deux champs sont intimement liées. DCGT a capitalisé sur certains concepts déjà existants dans la littérature économique. Par exemple, un concept similaire à la propriété de cohérence dans le temps a été élaboré par le prix Nobel de 2004 Kydland et Prescott (1977, [49]), à propos de la stratégie qu'un décideur doit mettre en oeuvre afin de déclencher la réponse souhaitée des agents économiques. Gale (1978, [35]) fut le premier à introduire l'idée de "Core séquentielle forte", dans un modèle d'économie monétaire. En outre, la notion de Core dans une économie d'échange d'informations incomplètes a été soigneusement étudiée au cours des dernières décennies (voir Forges et al. 2002, [34] pour une vue d'ensemble), bien avant que les fondements théoriques de DCGT ont été définies.

D'autre part, la recherche sur DCGT a été stimulée par des questions économiques réels, et dans les dernières années, les progrès sur DCGT ont représenté un outil valable pour les économistes. Predtetchinski, Herings, et Peters (2002, [76]) ont étudié le Core forte séquentielle en économies à deux niveaux dans lesquelles le commerce des biens a lieu dans la prémiére période et le commerce des produits se présente dans la périod successive. Herings, Predtetchinski et Perea (2006, [38]) ont étudié le Core faible séquentiel dans le même contexte.

Nous appliquons DCGT pour la première fois à un modèle de réseau de communication, dans la Section 3.2.

Maintenant, nous allons décrire notre modèle coopérative MDP sous l'hypothèse TU. Il s'agit essentiellement d'un modèle multi-agent MDP dans lequel les joueurs sont autorisés à former des coalitions tout au long du jeu. Nous analysons d'abord le jeu à long terme sur le MDP dans une perspective d'un jeu coopératif statique. Si on regarde le jeu dans son ensemble, nous pouvons toujours calculer hors-ligne la stratégie optimale pour la grande coalition de joueurs, par des techniques d'optimisation classiques pour MDPs mono-agent, dans lequel la grande coalition est considéré comme le seul agent. Dans le jeu à long terme, la valeur de chaque coalition est calculée comme la somme de long terme de récompenses que chaque coalition peut atteindre tout seul, et une solution coopérative est attribuée dans le jeu à long terme. Jusqu'à présent, cette formulation se rapporte encore à SCGT. Néanmoins, deux questions se posent. Tout d'abord, la solution coopérative à long terme est en fait la valeur prévue d'une variable aléatoire, et par conséquent, de faon pragmatique, il n'est pas clair comment la répartir entre les joueurs. Deuxièmement, quand l'horizon du jeu est infini, ou fini mais avec une durée inconnue, les joueurs peuvent exiger d'être récompensé au *chaque* étape du jeu, c'est à dire à chaque étape de la chaîne de Markov. Par conséquent, le défi est de mettre au point une procédure de distribution coopérative du payoff (CMDP) qui distribue la solution coopérative à long terme au long du jeu, dans chacun de ses états. De ce compromis apparemment inoffensif, un certain nombre de questions se posent. Puisque nous supposons que les coalitions peuvent se former tout au long du jeu, le CMDP doit satisfaire tous les joueurs, tout au long du match. En effet, si un accord à long terme a été stipulé par les joueurs au début du jeu en fonction de certains critères (par exemple, l'équité), il n'est pas clair si l'accord peut être maintenu au fil du temps, i.e. si un tel critère se garde après certaines étapes, ou une renégociation de l'accord est nécessaire. Des situations réelles abondent en exemples dans lesquels un contrat à long terme conclu en premier lieu doit être renégocié au cours du temps, e.g. la situation économique actuelle dans l'Union Européenne. La propriété que CMDP doit posséder afin d'éviter une renégociation est la cohérence temporelle (Petrosjan 1977, [72]). Un deuxième point important est la stabilité de la grande coalition, tout au long du match. Une coalition pourraient être tentée de violer l'accord à une certaine ètape du jeu dynamique, car elle ne peut garantir une meilleure allocation elle-même à partir de cet instant. Afin d'éviter cela, le CMDP a besoin d'appartenir au Core séquentiel (Kranich, Perea, et Peters 2005, [48]). Intuitivement, cette propriété declare que, chaque fois qu'une coalition est confrontée au dilemme "*On romp le contrat maintenant ou on coopére pour toujours?*", alors elle devrait opter pour la deuxième option puisqu'elle est plus rentable. De plus, nous exigeons que le CMDP satisfait la propriété de la maintenance de la coopération (Mazalov et Rettieva 2010, [59]). Intuitivement, cette propriété suggère que, à chaque étape temporelle, si une coalition est confrontée au dilemme "*On romp le contrat maintenant ou dans une étape temporelle?*", alors il faut choisir la deuxième option. Enfin, nous considérons la présence de joueurs avides, ayant une perspective myope du jeu. Dans ce cas, le CMDP a besoin d'appartenir au Core de chaque jeu statique joué dans chaque état du MDP, afin de contenter les joueurs avides aussi. Ensuite, nous élaborons un CMDP pour le MDP, baptisé MDP-CMDP, et nous trouvons des conditions pour lesquelles toutes les propriétés enrôlés ci-dessus sont satisfaites. Fait remarquable, nous constatons que la propriété de la maintenance de la coopération est un raffinement du concept du Core séquentiel pour les MDPs coopératifs.

Dans la Section 3.2 nous appliquons certains des concepts développés dans la Section 3.1 pour un scénario de communication sans fil. On considère un canal d'accès multiple gaussien (MAC) dans lequel le canal est quasi-statique, c'est à dire qu'il varie suffisamment lentement pour être supposée constant pendant toute la durée d'un mot de code. De plus, nous supposons qu'il suit une chaîne de Markov homogène (HMC) endogène avec un ensemble fini d'états dans lequel les transitions d'état se produisent à la fin de chaque période de cohérence. Nous attribuons un débit du codage à chaque utilisateur dans chaque état de la chaîne de Markov. Nous soulignons que, dans ce scénario, les probabilités de transition entre les états du canal du système ne dépendent pas de stratégies de transmission des utilisateurs. En ce sens, le scénario est simplifié par rapport à celui de la Section 3.1. D'autre part, une nouvelle complication est causée par l'introduction de l'hypothèse du payoff non-transférable (NTU), pour laquelle le débit du codage ne peut être partagé en aucune manière parmi les utilisateurs, mais seulement dans la région de la capacité de Shannon.

Dans ce scénario nous abordons la question de la répartition du débit du codage dans chaque état, d'une manière optimum global, c'est à dire que la somme des débits du codage est maximale à la fois dans chaque état et dans le processus à long terme. Nous appelons $\mathcal{M}$ cet ensemble de débit du codage optimal. Nous étudions deux procédures pour sélectionner une allocation de $\mathcal{M}$: la première, appelée "bottom-up", prévoit d'allouer d'abord les allocations statiques, tandis que le second, baptisé "top-down", suggère de choisir d'abord les allocations à long terme, et puis de calculer les allocations statiques associées. Cette dernière procédure serait plus utile, car elle permet de respecter le critère de sélection dans le processus à long terme, étant celle qui concerne les utilisateurs, qui sont dotés d'une perspective à long terme du jeu. Malheureusement, cette procédure ne génère pas toujours à des allocations réalisables, et nous offrons une solution pour ça. Ensuite, nous abordons la question de l'allocation d'un débit du codage équitable pour les utilisateurs dans le processus dynamique. Dans le cas statique, les critères de max-min, équité proportionnelle, et α-équité sont toujours bien définis [5]. Dans notre cas dynamique, le scénario est compliqué par le fait que nous exigeons que le critère de l'équité doit etre gardé tout au long du processus. Il n'est pas toujours possible de trouver une allocation constamment équitable; néanmoins, nous donnons une condition suffisante pour son existence. Ensuite, nous utilisons des outils développés dans la Section 3.1 pour mesurer la satisfaction des utilisateurs avec le débit du codage attribué. Nous constatons que $\mathcal{M}$ coïncide avec l'ensemble des débits du codage appartenant au Core séquentiel. De plus, $\mathcal{M}$ est exactement l'ensemble des débits du codage qui satisfaisaient la propriété de la maintenance de la coopération.

Dans la Section 3.3 nous nous occupons de MDP coopératifs sous l'hypothèse TU, dans lesquels les probabilités de transition entre les états ne dépendent pas des actions des joueurs. Comme dans la Section 3.2, une chaîne de Markov endogène gouverne l'évolution stochastique de l'état du système, donc nous appèlons ce scénario jeu coopératif Markovien.
Nous nous occupons de ce modèle dans une perspective différente de Sections 3.1, 3.2. En fait, nous n'avons pas étudié la répartition du payoff dans chaque état, puisque, par un argument de linéarité, il s'agit d'une question triviale. Au lieu de cela, nous abordons un problème de complexité relative au calcul de la valeur de Shapley (Shapley 1953, [82]) dans ce genre de jeux. Plus précisément, nous proposons trois méthodes pour calculer un intervalle de confiance pour l'indice de Shapley-Shubik dans les jeux de Markov (SSM). Nous étendons l'approche de Bachrach et al. en 2010 [14] pour les jeux statiques à des jeux de Markov. L'indice de Shapley-Shubik (Shapley et Shubik, 1954 [85]) est la valeur de Shapley appliquée à un jeu simple, c'est à dire un jeu dans lequel les valeurs des coalitions sont binaires, c'est à dire 1 ou 0. L'indice de Shapley-Shubik s'avère particulièrement approprié pour évaluer *a priori* le pouvoir des membres d'un comité de législation, et présente nombreuses applications à la politique (voir Taylor et Pacelli 2008, [91] pour une vue d'ensemble). Dans notre cas, le jeu dans chaque état est simple. Nous montrons que un nombre exponentiel de requêtes est nécessaire pour tout algorithme déterministe même pour approximer SSM avec une précision polynomial. Motivé par ça, nous proposons trois différentes approches aléatoires pour calculer un intervalle de confiance pour SSM. Leur complexité ne dépend pas du tout du nombre de joueurs. Ces approches sont valables aussi pour la valeur de Shapley classique de n'importe quel jeu coopératif Markovien. Le prémier intervalle de confiance, SCI, est valide sous l'hypothèse statique que l'agent estimateur a accès aux valeurs de la coalition dans tous les états en même temps, avant même que le processus de Markov commence. Bien que SCI repose sur une hypothèse irréaliste, il est encore un point de référence valable pour le calcul des deux autres intervalles de confiance proposés, surnommés DCI1 et DCI2. DCI1 et DCI2 sont valids également dans l'hypothèse plus réaliste dynamique: l'agent esti-

mateur apprend la valeur de coalitions au long du jeu. Nous proposons une méthode simple pour minimiser la longueur de DCI1 et nous comparons les trois approches proposées en termes de la longueur de l'intervalle de confiance. Enfin, nous proposons un compromis entre la complexité et la précision de notre algorithme randomisé, valide pour n'importe quel jeu coopératif Markovien.

Enfin, dans le Chapitre 4, nous utilisons une formulation MDP avec un espace d'états dénombrable pour résoudre un problème de sélection dynamique entre les différents canaux de transmission. Dans la littérature des algorithmes d'apprentissage, cette approche va sous l'étiquette de Multi Armed Bandit (MAB), dans lequel il existe un groupe de plusieurs processus aléatoires, ou des bras, qui peuvent être observés successivement. La récompense totale est la somme (actualisée ou en moyenne) des valeurs instantanées des bras observés. Sur la base des observations passées et de la connaissance statistique *a priori* de chaque bras, l'objectif est de maximiser la récompense attendue total.

Ce qu'on appelle Rested MAB suppose que l'état d'un bras qui n'a pas été tiré n'évolue pas à l'étape temporelle suivante. Dans ce cas, une solution d'optimisation brillante a été trouvé par Gittins (1989, [37]). Il a observé que la malédiction de la dimensionnalité des états dans le nombre de bras peut être surmontée avec le calcul d'un indice, pour chaque bras. La solution optimale consiste simplement en sélectionner les bras avec l'indice le plus élevé, dans chaque étape de décision. Ainsi, la complexité du calcul de la solution passe de exponentiel à linéaire dans le nombre de bras.

Nous nous occupons plutôt de Restless MAB, dans lesquels les bras évoluent même lorsqu'ils ne sont pas observés, ce qui est évidemment le cas des coefficients d'atténuation d'un canal de transmission. Dans ce cas, une solution efficace n'a pas encore été trouvée. Whittle (1988, [100]) a proposé un indice dont l'optimalité n'est pas garantie. Il a été prouvé par Papadimitriou et Tsitsiklis [67] que les restless MAB sont PSPACE-hard en général. Par conséquent, afin de faire face à Restless MAB, on a besoin de recourir à une formulation MDP avec un espace d'états dénombrable, ou de manière équivalente à un modèle MDP partiellement observable. Typiquement, l'état du MDP à chaque étape temporelle est les valeurs instantanées de chaque bras. Puisque l'état du bras non observés est inconnue, le seul agent de décision n'a à sa disposition que ses distributions de probabilité, conditionnées aux dernières observations.

Dans notre cas, les processus qui sont choisis à chaque fois sont des processus autorégressifs indépendants d'ordre 1, c'est á dir AR(1). Ils représentent les atténuations de canal à évanouissement lent sur les bandes de fréquences différentes d'un réseau sans fil à accès multiple. L'objectif est de maximiser le rapport signal sur bruit (SNR) à moyen au long terme pour l'utilisateur. L'utilisateur est censé connaître tous les paramètres des processus AR(1), donc aussi leur valeur attendue (inconditionnel). Une solution naïve suggère de choisir toujours le canal avec la plus grande valeur attendue. La procédure grosso modo est très sous-optimal si les valeurs attendues sont semblables les uns aux autres. En effet, il est possible d'exploiter l'auto-corrélation des canaux pour sélectionner celui qui est *instantanément* le meilleur.

Nous proposons deux algorithmes heuristiques avec une complexité linéaire avec le nombre d'armes à résoudre le problème Restless MAB. Le premier, myope, suggère de sélectionner le canal avec le plus haut SNR attendue conditionné, à chaque pas de temps. Celui-ci sélectionne les bras apparentement sous-optimales avec une certaine probabilité. Nous constatons que la stratégie myope se comporte presque optimalement lorsque tous les bras sont presque statistiquement équivalents. Lorsqu'un canal est caractérisé par une autocorrélation beaucoup plus élevé, alors l'approche randomisée surpasse l'approche myope et celle est quasi-optimal. Nous proposons ensuite une formulation en utilisant les MDP compétitifs, dans laquelle les différents utilisateurs accédent au même pool de canaux. Nous adaptons les deux algorithmes cités, rapprochant la "meilleur réponse" de l'utilisateur à l'encontre de la stratégie d'un second utilisateur qui est insensible à la présence du premièr.

Chapter 1

Markov Decision Processes (MDPs) and Game Theory: a brief overview

In this preliminary chapter we provide a short overview of Markov Decision Processes (MDPs), linear programming, Competitive and Cooperative Game Theory. It will provide to the reader the essential tools to understand the main results in this dissertation.

1.1 Discrete time Markov chains

Let $\mathbf{P}$ be a N-by-N stochastic matrix, i.e. non-negative and all its rows sum up to 1. Let $\{S_n\}_{n\in\mathbb{N}\cup\{0\}}$ be a discrete-time stochastic process on a finite state space $\mathcal{S}=\{s_1,\dots,s_N\}$. If for all integers $k\geq 0$:

$$\begin{aligned}p\big(S_{k+1}\!=\!s_{i_{k+1}}|S_i\!=\!s_{i_k},S_{k-1}\!=\!s_{i_{k-1}},\dots,S_0\!=\!s_{k_0}\big) &= p\big(S_{k+1}\!=\!s_{i_{k+1}}|S_k\!=\!s_{i_k}\big)\\ &= \mathbf{P}_{i_k,i_{k+1}},\end{aligned}$$

then the stochastic process is a discrete time Homogeneous Markov chain (HMC) with transition probability matrix $\mathbf{P}$.

The matrix $\mathbf{P}$ is irreducible when each state is reachable from any state with positive probability, i.e. for all s_i,s_j there exists n such that $(\mathbf{P}^n)_{i,j}>0$. The stationary distribution of $\mathbf{P}$ is the column vector $\mathbf{q}\in\mathbb{R}^N:\ \mathbf{q}^T\mathbf{P}=\mathbf{q}^T$, $\sum_i q_i=1$. If the HMC is irreducible, then $\mathbf{q}$ is the unique stationary distribution[1]. It can be interpreted as the probability distribution on the

[1] if the HMC is countably infinite, then the positive recurrence assumption is needed to guarantee the uniqueness of the stationary distribution. If the HMC is finite, instead,

states that is invariant over time. If $\mathbf{P}$ is irreducible, then it is also Cesàro summable and

$$\lim_{T\uparrow\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbf{P}^t = \mathbf{Q},$$

where each row of $\mathbf{Q}$ is the stationary distribution $\mathbf{q}^T$.

A good reference for the fundamentals of Markov chains is (Brémaud 1999, [24])

1.2 Linear programming

Linear programs are optimization problems that can be expressed in canonical primal form

$$\begin{aligned} \max\ & \mathbf{c}^T\mathbf{x} \\ \text{s.t.}\ & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}$$

The dual formulation is

$$\begin{aligned} \min\ & \mathbf{b}^T\mathbf{y} \\ \text{s.t.}\ & \mathbf{y}^T\mathbf{A} \geq \mathbf{c} \\ & \mathbf{y} \geq 0 \end{aligned}$$

Another primal form is

$$\begin{aligned} \max\ & \mathbf{c}^T\mathbf{x} \\ \text{s.t.}\ & \mathbf{A}\mathbf{x} \leq \mathbf{b} \end{aligned}$$

whose asymmetric dual form is

$$\begin{aligned} \min\ & \mathbf{b}^T\mathbf{y} \\ \text{s.t.}\ & \mathbf{y}^T\mathbf{A} = \mathbf{c} \\ & \mathbf{y} \geq 0 \end{aligned}$$

The strong duality theorem states that the optimal solutions of the primal and dual problems are equal. Classically, the linear programming problem is solved with the aid of the celebrated simplex algorithm, invented by Dantzig in 1947 (see also Dantzig 1998, [28]). It relies on the fact that an optimal solution is always found on an extreme point of the feasibility region, which is a convex polytope. The simplex algorithm is an efficient method to walk

irreducibility implies positive recurrence.

along the edges of the feasibility region through extreme points in which the objective function assume values being closer and closer to the optimum.

A good reference for linear programming is Boyd and Vandenberghe (2004 [23])

1.3 Markov Decision Processes

Markov Decision Processes (MDPs) are essentially controlled Markov chains, in which the rewards gained by the agent and the transition probabilities depend on the agent's actions. In this dissertation we will mainly deal with MDPs with finite states and finite actions space. In this case, $\mathcal{S} = \{s_1, \dots, s_N\}$ is the set of states. In each state $s \in \mathcal{S}$, the agent has at its disposal the set of actions $A(s)$. Let us assume that in state s the agent selects an action $a \in A(s)$. Then, an instantaneous reward $r(s,a)$ is earned. The process evolves stochastically on a discrete time grid $t = 0, 1, \dots$, according to a stochastic kernel controlled by the agent's strategy defined by the transition probability $p(s'|s,a)$, where $s, s' \in \mathcal{S}$ and $a \in A(s)$. Such kernel is stationary, i.e. the probability that the state at time $t+1$ is s', given that $S_t = s$ and the action chosen in s is a, is independent of time and of previous actions.
We call $\mathbf{f}$ a strategy for the agent, which determines the probability to take an action at each time T, given the whole history $\mathbf{h}(T)$ of previous action and state succession up to time T. If we consider the β-discounted criterion, $\beta \in [0;1)$, the total reward for the agent, given that s is the initial state of the process, can be expressed as

$$\Phi^{(\beta)}(s,\mathbf{f}) = \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}_{\mathbf{f}}(R_t | S_o = s), \tag{1.1}$$

where R_t is the instantaneous reward earned at time t. In this dissertation we will also utilize the equivalent notation $\Phi_{(\beta)}$. We call β-discount optimal strategy $\mathbf{f}^{*(\beta)}$ the strategy maximizing the following problem:

$$\max_{\mathbf{f}} \Phi^{(\beta)}(s,\mathbf{f}), \quad \forall\, s \in \mathcal{S}. \tag{1.2}$$

It is possible to prove that (1.2) has a solution, and the optimal strategy $\mathbf{f}^{*(\beta)}$ exists. Moreover, an optimal solution can be found amongst stationary strategies $\mathbf{F}^{\mathrm{S}}$, for which the action taken at time t depends only on the state of the process at time t, S_t. Hence, we will restrict our focus on stationary strategies. Under a stationary strategy, the stochastic process on the set of states $\mathcal{S}$ is a HMC. The strategy $\mathbf{f} \in \mathbf{F}^{\mathrm{S}}$ determines a probability distribution $\mathbf{f}(s)$ on $A(s)$, such that $f(s,a)$ is the probability that the agent chooses action

$a \in A(s)$ in state s. Moreover, if $\mathbf{f} \in \mathbf{F}^{\mathrm{S}}$, then we can rewrite (1.1) as

$$\Phi^{(\beta)}(s, \mathbf{f}) = \sum_{t=0}^{\infty} \beta^t \, p_t(s'|s, \mathbf{f}(s)) \sum_{a \in A(s')} f(s', a) \, r(s', a), \qquad (1.3)$$

where $p_t(s'|s)$ is the probability that the state is in state s', t steps after being in state s. It is convenient to write (1.3) in matricial form, as

$$\begin{aligned} \Phi^{(\beta)}(\mathbf{f}) &= \sum_{t=0}^{\infty} \beta^t \, \mathbf{P}(\mathbf{f}) \, \mathbf{r}(\mathbf{f}) \\ &= (\mathbf{I} - \beta \mathbf{P})^{-1} \, \mathbf{r}(\mathbf{f}), \end{aligned}$$

where $\mathbf{P}(\mathbf{f})$ is the transition probability matrix associated to the strategy $\mathbf{f}$, i.e. $P_{i,j}(\mathbf{f}) = p(s_j|s_i, \mathbf{f})$, $\Phi^{(\beta)}(\mathbf{f}) := [\Phi^{(\beta)}(s, \mathbf{f})]_{s \in \mathcal{S}}$, and $\mathbf{r}(\mathbf{f})$ is a N-by-1 vector whose j-th component is $\mathbf{r}(s_j, \mathbf{f}) = \sum_{a \in A(s')} f(s_j, a) \, r(s_j, a)$.

Let us now consider the average criterion. In the case when the stationary strategy $\mathbf{f}$ is adopted, the long-run reward can be expressed as

$$\Phi^{av}(s, \mathbf{f}) = \limsup_{T \uparrow \infty} \frac{1}{T} \sum_{t=0}^{T} p_t(s'|s, \mathbf{f}) \sum_{a \in A(s')} f(s', a) \, r(s', a). \qquad (1.4)$$

Let us assume that $\mathbf{P}(\mathbf{f})$ is irreducible for each $\mathbf{f} \in \mathbf{F}$. Then, $\mathbf{P}(\mathbf{f})$ is Cesàro summable, and we can write

$$\Phi^{av}(\mathbf{f}) = \sum_{i=1}^{N} q_i(\mathbf{f}) \, \mathbf{r}(s_i, \mathbf{f})$$

where $\mathbf{q}(\mathbf{f})$ is the stationary distribution of $\mathbf{P}(\mathbf{f})$.

1.3.1 Optimality equation and linear programming

Let us now focus on the β-discounted criterion. The maximal total discounted reward $\Phi^{*(\beta)} := \Phi^{(\beta)}(\mathbf{f}^{*(\beta)})$ solves the following optimality equation:

$$\Phi^{*(\beta)}(s) = \max_{a \in A(s)} \left\{ r(s, a) + \beta \sum_{s' \in \mathcal{S}} p(s'|s, a) \, \Phi^{*(\beta)}(s') \right\}, \quad \forall \, s \in \mathcal{S}. \qquad (1.5)$$

Let us call $a^*(s)$ an optimal action achieving the maximum in (1.5). Then,

$$\begin{aligned} \mathbf{f}^{*(\beta)}(s, a) &= 1 \quad \text{for } a = a^*(s) \\ \mathbf{f}^{*(\beta)}(s, a) &= 0 \quad \text{for } a \neq a^*(s). \end{aligned}$$

Hence, the optimal strategy $\mathbf{f}^{*(\beta)}(s, a)$ can be found in the set of stationary pure (or deterministic) strategies. Also, the maximum reward $\Phi^{*(\beta)}$

can be computed as the optimal value of the following linear programming optimization problem:

$$\min_{\mathbf{u}\in\mathbb{R}^N} \sum_{s\in\mathcal{S}} u(s) \tag{1.6}$$
$$\text{s.t. } u(s) \geq\, r(s,a)+\beta \sum_{s'\in\mathcal{S}} p(s'|s,a)\, u(s'), \quad \forall\, a\in A(s);\ s\in\mathcal{S}.$$

The (asymmetric) dual problem of (1.6) is the following linear programming problem:

$$\max_{\mathbf{x}} \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} x(s,a)\, r(s,a) \tag{1.7}$$
$$\text{s.t. } \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} \big[\delta(s,s') - \beta p(s'|s,a)\big] x(s,a) =\, 1, \quad \forall\, s'\in\mathcal{S}$$
$$x(s,a) \geq\, 0, \quad \forall\, a\in A(s);\ s\in\mathcal{S}.$$

where the dual variable is $\mathbf{x}$. The optimal strategy $\mathbf{f}^{*(\beta)}$ can be computed as

$$\mathbf{f}^{*(\beta)}(s,a) =\, \frac{x^*(s,a)}{\sum_{a'\in A(s)} x^*(s,a')}, \tag{1.8}$$

where $\mathbf{x}^*$ is the optimal solution of (1.7).

In the case of average criterion, we can formulate the following linear programming:

$$\max_{\mathbf{x}} \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} x(s,a)\, r(s,a)$$
$$\text{s.t. } \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} \big[\delta(s,s') - p(s'|s,a)\big] x(s,a) =\, 0, \quad \forall\, s'\in\mathcal{S}$$
$$\sum_{a\in A(s)} x(s,a) =\, 1, \quad \forall\, s\in\mathcal{S}$$
$$x(s,a) \geq\, 0, \quad \forall\, a\in A(s);\ s\in\mathcal{S}.$$

The optimal average strategy is computed analogously to (1.8).

1.3.2 Value iteration

Let us consider the β-discounted criterion. We are going to introduce the approximation technique called value iteration, to solve an MDP optimization problem. When the state space is large, but still finite, the use of value iteration is to be preferred to a linear programming formulation.

Algorithm 1.3.1. *Value Iteration for finite MDPs.*

1. *Initialize* $u_0(s) \in \mathbb{R}$ *for all* $s \in \mathcal{S}$, $n = 0$, $\epsilon > 0$.
2. *For all* $s \in \mathcal{S}$, *compute* $u_{n+1}(s)$ *as*

$$\begin{aligned} u_{n+1}(s) = & \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s' \in \mathcal{S}} p(s'|s,a) u_n(s') \right\} \\ := & L\big(u_n(s)\big) \end{aligned}$$

3. *If*

$$||u_{n+1} - u_n|| < \epsilon(1-\beta)/2\beta, \tag{1.9}$$

then go to step 4, otherwise set $n := n+1$ *and return to step 2.*

4. *For each* $s \in \mathcal{S}$, *define*

$$a^{(\epsilon)}(s) \in \operatorname*{argmax}_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s' \in \mathcal{S}} p(s'|s,a) u_n(s') \right\}.$$

The value u_n converges to the optimal $\Phi^{*(\beta)}$ for $n \uparrow \infty$ and the stationary policy obtained from $a^{(\epsilon)}$ is ϵ-optimal, i.e. $||u_n - \Phi^{*(\beta)}|| < \epsilon/2$ whenever (1.9) holds.

There exists a variant of the value iteration for MDPs with continuous state space. The state space is discretized and step 2 of Algorithm 1.3.1 is applied to the state grid $\mathcal{S}_G$, while in other points the value is interpolated linearly. In other words, the operator L is redefined as

$$L'\big(u_n(s)\big) = \begin{cases} L\big(u_n(s)\big) & \text{if } s \in \mathcal{S}_G \\ \sum_k \lambda_k \, u_n(s_k), \; s_k \in \mathcal{S}_G, & \text{if } s \notin \mathcal{S}_G \end{cases}$$

for some convex coefficients $\{\lambda\}_k$ such that $s = \sum_k \lambda_k s_k$. It is possible to prove (Bäuerle and Rieder, 2011 [17]) that, if the mesh size is sufficiently small, then the algorithm converges and gives an approximation of the optimal value.

1.3.3 Blackwell optimality

The average criterion is under-selective, since it does only take into account the arbitrarily distant tail of rewards. The Blackwell optimal strategy (Blackwell, 1962 [21]) is a refinement of the average optimal strategy. It coincides with the β-discounted optimal strategy, for all β sufficiently close to 1. Hordjik, Dekker, and Kallenberg proposed in [41] to compute the Blackwell optimal strategy by utilizing the simplex algorithm in the non-Archimedean field of rational function with real coefficients.

A good reference for the fundamentals of single-agent MDPs is (Puterman, 1994 [77])

1.4 Game Theory

1.4.1 Competitive Game Theory with two players

Let us assume the presence of two players, who interact with each other. Player $i = 1, 2$ has at its disposal a set of actions A_i. When both players select an action, each of them obtains a reward depending also on the other player's action. Let $r_i(a_1, a_2)$ be the reward for player i when players 1 and 2 select the actions $a_1 \in A_1$ and $a_2 \in A_1$, respectively. We say that the game is zero-sum when, for each couple (a_1, a_2), $r_1(a_1, a_2) = -r_2(a_1, a_2) := r(a_1, a_2)$. In this case, both players adopt an antagonistic behaviour, since a benefit for one is a loss for the other. To examine this situation, the (mixed) Nash equilibrium (Nash, 1950 [64]) is typically utilized. It is defined as the couple of probability distributions $(\mathbf{f}_1^*, \mathbf{f}_2^*)$ on the action sets A_1, A_2 respectively, that no player has interest in deviating from, i.e.

$$r(\mathbf{f}_1, \mathbf{f}_2^*) \le r(\mathbf{f}_1^*, \mathbf{f}_2^*) \le r(\mathbf{f}_1^*, \mathbf{f}_2), \quad \forall \mathbf{f}_1 \in \mathbf{F}_1,\ \mathbf{f}_2 \in \mathbf{F}_2$$

where $\mathbf{F}_i$ is the set of mixed strategies for player i and $r(\mathbf{f}_1, \mathbf{f}_2)$ is the expected value of r under the strategies $\mathbf{f}_1, \mathbf{f}_2$. Thanks to the minimax Theorem (Von Neumann, 1928 [95]) the optimal couple $(\mathbf{f}_1^*, \mathbf{f}_2^*)$ exists and the reward at the equilibrium, can also be written as a max-min formulation

$$\begin{aligned} r(\mathbf{f}_1^*, \mathbf{f}_2^*) &= \max_{\mathbf{f}_1 \in \mathbf{F}_1} \min_{\mathbf{f}_2 \in \mathbf{F}_2} \sum_{a_1 \in A_1} \sum_{a_2 \in A_2} f_1(a_1) f_2(a_2) r(a_1, a_2) && (1.10) \\ &= \min_{\mathbf{f}_2 \in \mathbf{F}_2} \max_{\mathbf{f}_1 \in \mathbf{F}_1} \sum_{a_1 \in A_1} \sum_{a_2 \in A_2} f_1(a_1) f_2(a_2) r(a_1, a_2) && (1.11) \\ &:= \mathrm{val}(\mathbf{R}) \end{aligned}$$

where $\mathbf{R}$ is the $|A_1|$-by-$|A_2|$ matrix of the rewards. The strategy $\mathbf{f}_1^*$ ($\mathbf{f}_2^*$) maximizes (minimizes) expression (1.10) ((1.11)).

1.4.2 Static Cooperative games

In static cooperative games we consider the presence of P players which can cooperate and coordinate their actions. We call $\mathcal{P} = \{1, \dots, P\}$ the grand coalition. Again, let us assume that each player i has at its disposal the set of actions A_i and that the reward r_i depends on the players' actions $a_1, \dots, a_P$. To every coalition $\Lambda \subseteq \mathcal{P}$ the set of feasible pay-offs $\mathcal{V}(\Lambda)$ is assigned. Under the Transferable Utility (TU) assumption, there exists $v(\Lambda) \in \mathbb{R}$ such that $\mathcal{V}(\Lambda)$ is the hyperplane defined as:

$$\mathcal{V}(\Lambda) = \left\{ \mathbf{x} \in \mathbb{R}^{|\Lambda|} : \sum_{i=1}^{|\Lambda|} x_i \le v(\Lambda) \right\}.$$

We consider $v(\emptyset) = 0$. Under the Non-Transferable Utility (NTU), $\mathcal{V}(\Lambda)$ may assume any shape; for example, in Section 3.2 $\mathcal{V}(\Lambda)$ is a polymatroid. The feasible region $\mathcal{V}(\Lambda)$ can also be interpreted as the set of pay-offs that Λ can obtain when it is not capable to coordinate its strategies with the anti-coalition $\mathcal{P}\backslash\Lambda$. In the TU case, Morgenstern and von Neumann (1953, [96]) suggested to assume that the anti-coalition adopts an antagonist behaviour, hence to compute $v(\Lambda)$ as the value of the zero-sum game of Λ against $\mathcal{P}\backslash\Lambda$, i.e.

$$v(\Lambda) = \max_{\mathbf{f}_\Lambda \in \mathbf{F}_\Lambda} \min_{\mathbf{f}_{\mathcal{P}\backslash\Lambda} \in \mathbf{F}_{\mathcal{P}\backslash\Lambda}} \sum_{i \in \Lambda} r_i(\mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\backslash\Lambda}),$$

where $\mathbf{F}_\Lambda$ is the set of mixed strategies available to Λ. The coalition value $v(\Lambda)$ can be also interpreted as the maximum sum of pay-offs that Λ can guarantee, whatever the strategy of the anti-coalition is. Therefore, the antagonistic behaviour of $\mathcal{P}\backslash\Lambda$ is to be considered a worst-case scenario.

Let us discuss how to allocate the pay-off, or reward, to the players. Let us assume that the grand coalition $\mathcal{P}$ is formed.
Let us introduce the notion of Core. Let $\mathbf{x} \in \mathcal{V}(\mathcal{P})$. We say that $\mathbf{x}$ is blocked by coalition $\Lambda \subseteq \mathcal{P}$ if there exists $\mathbf{y} \in \mathcal{V}(\Lambda)$ such that $y_i > x_i$ for all $i \in \Lambda$, i.e. coalition Λ cannot accept the allocation $\mathbf{x}$ since it can achieve a better allocation on its own. Then, the Core $\mathbf{Co}$ is defined as the set of all unblocked allocations in $\mathcal{V}(\mathcal{P})$. In the TU case, $\mathbf{Co}$ has a nice formulation, as the set of all $\mathbf{x} \in \mathbb{R}^P$ such that

$$\begin{aligned} &\sum_{i \in \mathcal{P}} x_i = v(\mathcal{P}) \\ &\sum_{i \in \Lambda} x_i \geq v(\Lambda), \quad \forall\, \Lambda \subset \mathcal{P}. \end{aligned}$$

The Bondareva-Shapley Theorem (Bondareva, 1963 [22], Shapley, 1967 [83]) provides a necessary and sufficient condition for the Core to be non-empty, i.e. for all functions $\alpha : 2^P \to [0;1]$ such that

$$\sum_{\Lambda \ni i} \alpha(\Lambda) = 1, \quad \forall\, i \in \mathcal{P}$$

then

$$\sum_{\Lambda \subseteq \mathcal{P}} \alpha(\Lambda) v(\Lambda) \leq v(\mathcal{P}).$$

The proof stems directly from the feasibility conditions of the dual linear programming problem associated to the Core.

The Shapley value $\mathcal{S}h \in \mathbb{R}^P$ (Shapley, 1953 [82]) is another way to distribute the pay-off among the players, that we illustrate in the TU case:

$$\begin{aligned}\mathcal{S}h_i &= \frac{1}{P!}\sum_{\pi\in\Pi(P)} v(\Lambda(\pi)\cup\{i\}) - v(\Lambda(\pi)), \quad \forall\, i\in\mathcal{P}\\ &= \sum_{\Lambda\subseteq\mathcal{P}\backslash\{i\}} \frac{(P-|\Lambda|-1)!(|\Lambda|)!}{P!}\left[v(\Lambda\cup\{i\}) - v(\Lambda)\right]\end{aligned}$$

where $\Pi(P)$ is the set of $P!$ permutations of $\{1,\dots,P\}$ and $\Lambda(\pi)$ is the set of players preceding i in the permutation π. The Shapley value is the only allocation $\mathbf{x}\in\mathbb{R}^P$ for which these four properties are jointly satisfied:

- *Efficiency*: $\sum_i x_i = v(\mathcal{P})$;
- *Dummy*: if i is a dummy player, i.e. $v(\Lambda\cup\{i\}) - v(\Lambda) = 0$ for all $\Lambda\subseteq\mathcal{P}\backslash\{i\}$, then $x_i = 0$;
- *Symmetry*: if two players i and j are symmetric with respect with the game, i.e. $v(\Lambda\cup\{i\}) = v(\Lambda\cup\{j\})$ for all $\Lambda\subseteq\mathcal{P}\backslash\{i,j\}$, then $x_i = x_j$;
- *Linearity*: $\mathbf{x}(v_1+v_2) = \mathbf{x}(v_1)+\mathbf{x}(v_2)$, where v_1+v_2 is the game with coalition values $v(\Lambda) = v_1(\Lambda)+v_2(\Lambda)$, for all coalitions Λ.

Moreover, if the game is superadditive i.e.

$$v(\Lambda_1\cup\Lambda_2) \geq v(\Lambda_1)+v(\Lambda_2), \quad \forall\,\Lambda_1\cap\Lambda = \emptyset,$$

then the Shapley value is also individually rational, i.e. $\mathcal{S}h_i \geq v(\{i\})$, for each player i. We remark that the superadditive is a common assumption, since it is a necessary condition to enforce cooperation among players.

A good reference for Competitive Game Theory is (Myerson 1997 [63]). For the cooperative case, we suggest (Peleg and Sudhölter, 2007 [69])

Chapter 2

Competitive and Long-run Cooperative MDPs

2.1 Algorithms for uniform optimal strategies in two-player zero-sum Competitive MDPs with perfect information

In Competitive Markov Decision Processes (MDPs) with perfect information, in each state at most one player has more than one action available. We deal with zero-sum two-player Competitive MDPs with perfect information. We propose two algorithms to find the uniform optimal strategies and one method to compute the optimality range of discount factors. We prove the convergence in finite time for one algorithm. The uniform optimal strategies are also optimal for the long run average criterion and, in transient games, for the undiscounted criterion as well.

2.1.1 Introduction

Competitive Markov Decision Processes (MDPs) are multi-stage interactions among several participants in an environment whose conditions change stochastically, influenced by the decisions of the players. Such games were introduced by Shapley (1953, [81]), who proved the existence of the discounted value and of the stationary discounted optimal strategies in two-player zero-sum games with finite state and action spaces. The problem of long term average reward games was addressed first by Gillette (1957, [36]). Bewley and Kohlberg (1976, [19]) proved that the field of real Puiseux series is an appropriate class to study the asymptotic behavior of discounted Competitive MDP when the discount factor tends to one. Mertens and Neyman (1981, [61]) showed the existence of the long term average value of Competitive MDPs. Then, Parthasarathy and Raghavan (1981, [68]) first introduced the notion of order field property. This property implies that the solution of a game lies in the same ordered field of the game data. Solan and Vieille (2009, [88]) presented an algorithm to find the ϵ-optimal uniform discounted strategies in two-player zero-sum Competitive MDPs, where $\epsilon > 0$.

Perfect information games were addressed by several researchers (e.g. see Thuijsman and Raghavan, 1997 [93], Altman, Feinberg, and Shwartz, 2000 [7]), since they are the most elementary form of Competitive MDPs: the reward and the transition probabilities in each state are controlled at most by one player. Recently, Raghavan and Syed (2002, [78]) provided an algorithm which finds the optimal strategies for two-player zero-sum perfect information games under the discounted criterion for a fixed discount factor.

Markov Decision Processes (MDPs) can be seen as Competitive MDPs in which only one player can possess more than one action in each state. It is well known (see e.g. Filar and Vrieze, 1997 [32]) that the optimal strategy in an MDP can be computed with the help of a linear programming formulation. Hordijk, Dekker and Kallenberg (1985, [41]) proposed to find the Blackwell optimal strategies (uniform optimal discount strategies) for MDPs by using the simplex method in the ordered field of rational functions with real coefficients. Altman, Avrachenkov and Filar (1999, [6]) analysed singularly perturbed MDP using the simplex method in the ordered field of rational functions. More generally, Eaves and Rothblum (1994, [29]) studied how to solve a vast class of linear problems, including linear programming, in any ordered field.

In this section we propose two algorithms to determine the uniform optimal discount strategies in two-player zero-sum games with perfect information. Such strategies are optimal in the long run average criterion as well. The proposed approaches generalize the works by Hordijk, Dekker, Kallenberg (1985, [41]) and Raghavan, Syed (2003, [78]) to the game model in the field $F(\mathbb{R})$ of the non-archimedean ordered field of rational functions with coefficients in $\mathbb{R}$.

Let Γ be a two-player zero-sum Competitive MDP with perfect information and $\Gamma_i(\mathbf{h}), i = 1, 2$ be the MDP that player i faces when the other player fixes his own strategy $\mathbf{h}$. Our first algorithm can be summed up in the following 3 steps:

1. Choose a stationary pure strategy $\mathbf{g}$ for player 2.

2. Find the uniform optimal strategy $\mathbf{f}$ for player 1 in the MDP $\Gamma_1(\mathbf{g})$.

3. Find the *first* state controlled by player 2 in which a change of strategy $\mathbf{g}'$ is a benefit for player 2 for all the discount factors close enough to 1. If it does not exists, then $(\mathbf{f}, \mathbf{g})$ are uniform optimal, otherwise set $\mathbf{g} := \mathbf{g}'$ and go to step 2.

It is evident that player 1 is left totally free to optimize the MDP that he faces at each iteration of the algorithm in the most efficient way.
Our second algorithm is a best response approach, in which the two players alternatively find their own uniform optimal strategies:

1. Choose a stationary pure strategy $\mathbf{g}$ for player 2.

2. Find the uniform optimal strategy $\mathbf{f}$ for player 1 in the MDP $\Gamma_1(\mathbf{g})$.

3. If $\mathbf{g}$ is uniform optimal for player 2 in the MDP $\Gamma_2(\mathbf{f})$, then $(\mathbf{f}, \mathbf{g})$ are uniform optimal. Otherwise, find the uniform optimal strategy $\mathbf{g}'$ in $\Gamma_2(\mathbf{f})$, set $\mathbf{g} := \mathbf{g}'$ and go to step 2.

The convergence in a finite time of the first algorithm is proven, while for the second we provide numerical analysis. We also show that the second algorithm has a lower complexity.

This section is organized as follows. In Section 2.1.2 we introduce formally the properties of Competitive MDPs, section 2.1.3 is dedicated to the description of the field of rational functions with real coefficients, while in section 2.1.4 we recall the linear programming procedures in the field $F(\mathbb{R})$ in order to find a Blackwell optimal policy for MDPs. We present some new useful results on perfect information games in Section 2.1.5 and section 2.1.6 is dedicated to the description and to the validation of our first algorithm. In section 2.1.7 we provide a numerical example. In Section 2.1.8 we introduce an algorithm whose convergence is only conjectured; we report some considerations and numerical results about the complexity of our algorithms in Section 2.1.8. In Section 2.2 we finally prove that, for transient stochastic games, $(\mathbf{f}^*, \mathbf{g}^*)$ are optimal under the undiscounted criterion as well.

Some notation remarks: the ordering relation between vectors of the same length $\mathbf{a} \geq (\leq)\mathbf{b}$ means that for every component i, $\mathbf{a}(i) \geq (\leq)\mathbf{b}(i)$. The discount factor and the interest rate are barred $(\overline{\beta}, \overline{\rho})$ if they are a fixed value; the symbols β, ρ represent the related variables.

2.1.2 The model

In a two-player Competitive MDP Γ we have a set of states $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, and for each state s the set of actions available to the i-th player is called $A^{(i)}(s) = \{a_1^{(i)}(s), \ldots, a_{m_i(s)}^{(i)}\}$, $i = 1, 2$. Each triple (s, a_1, a_2) with $a_1 \in A^{(1)}$, $a_2 \in A^{(2)}$ is assigned an immediate reward $r(s, a_1, a_2)$ for player 1,

$-r(s, a_1, a_2)$ for player 2 and a transition probability distribution $p(.|s, a_1, a_2)$ on $\mathcal{S}$.

A stationary strategy $\mathbf{u} \in \mathbf{U}_S$ for the i-th player determines the probability $u(a|s)$ that in state s player i chooses the actions $a \in [a_1^{(i)}, \dots, a_{m_i(s)}^{(i)}]$. We assume that both the number of states and the overall number of available actions are finite.

It is evident that a couple of strategies $\mathbf{f} \in \mathbf{F}_S,\ \mathbf{g} \in \mathbf{G}_S$ for player 1 and 2, respectively, sets up a Markov chain in which the transition probability equals

$$p(s'|s, \mathbf{f}, \mathbf{g}) = \sum_{p=1}^{m_1(s)} \sum_{q=1}^{m_2(s)} p(s'|s, a_p^{(1)}, a_q^{(2)})\, \mathbf{f}(a_p^{(1)}|s)\, \mathbf{g}(a_q^{(2)}|s)$$

$\forall\, s, s' \in \mathcal{S}$, while the average immediate reward $r(s, \mathbf{f}, \mathbf{g})$ equals

$$r(s, \mathbf{f}, \mathbf{g}) = \sum_{p=1}^{m_1(s)} \sum_{q=1}^{m_2(s)} r(s, a_p^{(1)}, a_q^{(2)})\, f(a_p^{(1)}|s)\, g(a_q^{(2)}|s)$$

Let $\overline{\beta} \in [0; 1)$ be the discount factor and $\overline{\rho}$ be the interest rate such that $\overline{\beta}(1 + \overline{\rho}) = 1$. Note that when $\overline{\beta} \uparrow 1$, then $\overline{\rho} \downarrow 0$. We define $\mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}, \mathbf{g})$ as a column vector of length N such that its i-th component equals the expected $\overline{\beta}$-discounted reward when the initial state of the Competitive MDP is s_i:

$$\mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}, \mathbf{g}) = \sum_{t=0}^{\infty} \overline{\beta}^t \mathbf{P}^t(\mathbf{f}, \mathbf{g}) \mathbf{r}(\mathbf{f}, \mathbf{g})$$

where $\mathbf{P}(\mathbf{f}, \mathbf{g})$ and $\mathbf{r}(\mathbf{f}, \mathbf{g})$ are the N-by-N transition probability matrix and the N-by-1 average reward vector associated to the couple of strategies $(\mathbf{f}, \mathbf{g})$ respectively.

Definition 1. *The $\overline{\beta}$-discounted value of the game Γ is such that*

$$\mathbf{\Phi}_{(\bar{\beta})}(\Gamma) = \sup_{\mathbf{f}} \inf_{\mathbf{g}} \mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}, \mathbf{g}) = \inf_{\mathbf{g}} \sup_{\mathbf{f}} \mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}, \mathbf{g}). \tag{2.1}$$

Definition 2. *An optimal strategy $\mathbf{f}^*_{(\bar{\beta})}$ for player 1 assures to him a reward which is at least $\mathbf{\Phi}_{(\bar{\beta})}(\Gamma)$*

$$\mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}^*_{(\bar{\beta})}, \mathbf{g}) \geq \mathbf{\Phi}_{(\bar{\beta})}(\Gamma) \qquad \forall\, \mathbf{g} \in \mathbf{G}$$

while $\mathbf{g}^*_{(\bar{\beta})}$ is optimal for player 2 iff

$$\mathbf{\Phi}_{(\bar{\beta})}(\mathbf{f}, \mathbf{g}^*_{(\bar{\beta})}) \leq \mathbf{\Phi}_{(\bar{\beta})}(\Gamma) \qquad \forall\, \mathbf{f} \in \mathbf{F}.$$

Let $\mathbf{\Phi}(\mathbf{f},\mathbf{g})$ be the long term average value of the game Γ associated to the couple of strategies $(\mathbf{f},\mathbf{g})$:

$$\mathbf{\Phi}(\mathbf{f},\mathbf{g}) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbf{P}^t(\mathbf{f},\mathbf{g})\mathbf{r}(\mathbf{f},\mathbf{g})$$

and $\mathbf{\Phi}(\Gamma)$ be the value vector for the long term average criterion of the game Γ, defined in an analogous way to expression (2.1).

The existence of optimal strategies in discounted Competitive MDPs is guaranteed by the following theorem (Filar and Vrieze, 1997 [32]):

Theorem 2.1.1. *Under the hypothesis of discounted pay-off, Competitive MDPs possess a value, the optimal strategies* $(\mathbf{f}^*_{(\bar\beta)}, \mathbf{g}^*_{(\bar\beta)})$ *exist among stationary strategies and moreover* $\mathbf{\Phi}_{(\bar\beta)}(\Gamma) = \mathbf{\Phi}_{(\bar\beta)}(\mathbf{f}^*_{(\bar\beta)}, \mathbf{g}^*_{(\bar\beta)})$.

Definition 3. *A stationary strategy* $\mathbf{h}$ *is said to be uniformly discount optimal for a player if* $\mathbf{h}$ *is optimal for every* $\overline{\beta}$ *close enough to 1 (or, equivalently, for all* $\overline{\rho}$ *close enough to 0).*

In the present Chapter we deal with perfect information Competitive MDPs.

Definition 4. *Under the hypothesis of perfect information, in each state at most one player has more than one action available.*

Let $\mathcal{S}_1 = \{s_1, \ldots, s_{t_1}\}$ be the set of states controlled by player 1 and $\mathcal{S}_2 = \{s_{t_1+1}, \ldots, s_{t_1+t_2}\}$ be the set controlled by player 2, with $t_1+t_2 \le N$.

2.1.3 The ordered field of rational functions with real coefficients

Let $P(\mathbb{R})$ be the ring of all the polynomials with real coefficients.

Definition 5. *The dominating coefficient of a polynomial* $f = a_0 + a_1x + \cdots + a_nx^n$ *is the coefficient* a_k*, where* $k = \min\{i : a_i \neq 0\}$ *and we denote it with* $\mathcal{D}(f)$.

Let $F(\mathbb{R})$ be the non-archimedean ordered field of fractions of polynomials with coefficients in $\mathbb{R}$:

$$f(x) = \frac{c_0 + c_1x + \cdots + c_nx^n}{d_0 + d_1x + \cdots + d_mx^m} \qquad f \in F(\mathbb{R})$$

where the operations of sum and product are defined in the usual way (see Hordijk, Dekker and Kallenberg, 1985 [41]). Two rational functions h/g, p/q are identical (and we say $h/g =_l p/q$) if and only if $h(x)q(x) = p(x)g(x)\ \forall x \in$

$\mathbb{R}$.

The following lemma (Hordijk et al., 1985 [41]) introduces the ordering in the field $F(\mathbb{R})$:

Lemma 2.1.2. *A complete ordering in $F(\mathbb{R})$ is obtained by the rule*

$$\frac{p}{q} >_l 0 \iff \mathcal{D}(p)\mathcal{D}(q) > 0 \qquad p, q \in P(\mathbb{R})$$

In the same way, we can also define the operations of maximum ($\max_l$) and minimum ($\min_l$) in $F(\mathbb{R})$.

The ordering law defined above is useful when one wants to compare the behavior of rational functions whose indipendent variable is positive and approaches to 0 (see Hordijk et al., 1985 [41]).

Lemma 2.1.3. *The rational function p/q is positive ($p/q >_l 0$) if and only if there exists $x_0 > 0$ such that $p(x)/q(x) > 0$ for every $x \in (0; x_0]$.*

Application to Competitive MDPs

From the next theorems the reader will start perceiving the importance of dealing with the field $F(\mathbb{R})$ in Competitive MDPs.

Theorem 2.1.4. *Let $\mathbf{f}, \mathbf{g}$ be two stationary strategies respectively for players 1 and 2 and $\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}) : \mathbb{R} \to \mathbb{R}^N$ be the discounted reward associated to the couple of strategies ($\mathbf{f}$,$\mathbf{g}$) expressed as a variable of ρ. Then, $\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}) \in F(\mathbb{R})$.*

Proof. For any couple of stationary strategies $(\mathbf{f}, \mathbf{g})$, we can write

$$\sum_{s'=1}^{N} [(1+\rho)\delta_{s,s'} - p(s'|s, \mathbf{f}, \mathbf{g})]\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}, s') = (1+\rho)r(s, \mathbf{f}, \mathbf{g}) \quad s \in [1; N] \tag{2.2}$$

where ρ is a variable. By solving the above system of equations in the unknown $\mathbf{\Phi}_{(\rho)}$ by Cramer rule, it is evident that $\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}) \in F(\mathbb{R})$. □

Generally, the discounted value of a Competitive MDP for all the interest rates close enough to 0 belongs to the field of real Puiseux series (see Filar and Vrieze, 1997 [32]). From Theorems 2.1.1 and 2.1.4 it is straightforward to obtain the following important Lemma.

Lemma 2.1.5. *Let Γ be a zero-sum Competitive MDP which possesses uniform discount optimal strategies for both players. Then, there exist $\overline{\rho}^* > 0$ and $\mathbf{\Phi}^*_{(\rho)}(\Gamma) \in F(\mathbb{R})$ such that $\mathbf{\Phi}^*_{(\bar{\rho})}(\Gamma)$ is the discounted optimal value for all the interest rates $\overline{\rho} \in (0; \overline{\rho}^*]$.*

Proof. Let $(\mathbf{f}^*, \mathbf{g}^*)$ be a couple of uniformly discount optimal strategies for players 1 and 2 respectively. Then, by definition, there exists $\overline{\rho}^* > 0$ such that $(\mathbf{f}^*, \mathbf{g}^*)$ are discounted optimal for all the interest rates $\overline{\rho} \in (0; \overline{\rho}^*]$. From Theorem 2.1.4 we know that $\mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}^*) \in F(\mathbb{R})$ and, from Theorem 2.1.1, the optimum uniform discounted value $\mathbf{\Phi}_{(\bar{\rho})}(\Gamma) = \mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}^*, \mathbf{g}^*)\ \forall\, \overline{\rho} \in (0; \overline{\rho}^*]$. So, $\mathbf{\Phi}^*_{(\rho)}(\Gamma) \in F(\mathbb{R})$ represents the discounted value of Γ for all the interest rates sufficiently close to 0. □

Lemma 2.1.6. *Let Γ be a zero-sum Competitive MDP which possesses uniform discount optimal strategies $\mathbf{f}^*, \mathbf{g}^*$ for players 1 and 2 respectively. Then,*

$$\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}^*) \leq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}^*) =_l \mathbf{\Phi}^*_{(\rho)}(\Gamma) \leq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}) \quad \forall\, \mathbf{f}, \mathbf{g} \tag{2.3}$$

where

$$\mathbf{\Phi}^*_{(\rho)}(\Gamma) =_l \max_{\mathbf{f}}{}_l \min_{\mathbf{g}}{}_l\, \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}) =_l \min_{\mathbf{g}}{}_l \max_{\mathbf{f}}{}_l\, \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}). \tag{2.4}$$

Proof. From Theorem 2.1.1 and by the definition of uniform discount optimal strategy, we assert that

$$\exists\, \overline{\rho}^* > 0 :\ \forall\, \overline{\rho} \in (0; \overline{\rho}^*] \ \Rightarrow\ \mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}, \mathbf{g}^*) \leq \mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}^*, \mathbf{g}^*) \leq \mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}^*, \mathbf{g}) \quad \forall\, \mathbf{f}, \mathbf{g}$$

which coincides with (2.3) for Lemma 2.1.3. The equation (2.4) is a direct consequence of (2.3). □

Definition 6. $\mathbf{\Phi}^*_{(\rho)}(\Gamma)$*, defined as in (2.4), is the uniform discount value of the Competitive MDP Γ.*

2.1.4 Computation of Blackwell optimum policy in MDPs

In this section we will discuss about some concepts of linear programming, which can be easily found on any book on linear optimization (e.g. see Luenberger and Ye, 2008 [51]).

Let Ψ be a Markov Decision Process, which can be seen as a two-player Competitive MDP in which one of the two players either fixes his own strategy or has only one available action in each state. We call $\mathbf{\Phi}_{(\rho)}(\mathbf{f})$ the value of the discounted MDP associated to the strategy $\mathbf{f}$ with interest rate variable ρ.

It is known (Puterman, 1994 [77]) that the interval of interest rate $(0; \infty)$ can be broken into a finite number n of subintervals, say $(0 \equiv \alpha_0; \alpha_1]$,$(\alpha_1; \alpha_2]$, …,$(\alpha_{n-1}; \infty)$ in such a way that for each one there exists an optimal pure strategy.

A Blackwell optimal policy is an optimal strategy associated to the first sub-interval.

Definition 7. *We say that the strategy* $\mathbf{f}^*$ *is Blackwell optimal iff there exists* $\bar{\rho}^* > 0$ *such that* $\mathbf{f}^*$ *is optimal in the* $(1/\bar{\rho} - 1)$*-discounted MDP for all the interest rates* $\bar{\rho} \in (0; \bar{\rho}^*]$*.*

Since for Theorem 2.1.4 $\mathbf{\Phi}_{(\rho)}(\mathbf{f}) \in F(\mathbb{R})$ for any $\mathbf{f} \in \mathbf{F}_S$, we can say

$$\mathbf{\Phi}_{(\rho)}(\mathbf{f}^*) \geq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}) \qquad \forall \, \mathbf{f} \in \mathbf{F}$$

where $\mathbf{F}$ is the set of all possible strategies.
Hordijk, Dekker, and Kallenberg (1985, [41]) provided a useful algorithm to compute the Blackwell optimum policy in MDPs. It consists in solving the following parametric linear programming problem:

$$\begin{cases} \max_{\mathbf{x}}{}_{l} \; \sum_{s=1}^{N} \sum_{a=1}^{m(s)} x_{sa}(\rho) r(s,a) \\ \sum_{s=1}^{N} \sum_{a=1}^{m(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,a)] \, x_{s,a}(\rho) =_l 1, \quad s' \in \mathcal{S} \\ x_{s,a}(\rho) \geq_l 0, \quad s \in \mathcal{S}, \; a \in A(s) \end{cases} \tag{2.5}$$

in the ordered field of rational functions with real coefficients $F(\mathbb{R})$. This means that

i) ρ is the variable of polynoms;

ii) all the elements of the related simplex tableau belong to $F(\mathbb{R})$;

iii) all the algebraic and ordering operations required by the simplex method are carried out in the field $F(\mathbb{R})$.

The practical technique to solve the linear optimization problem (2.5) proposed by Hordijk et al. (1985, [41]) is the so-called *two-phases method.*
In the *first phase* the artificial variables $z_1, \ldots, z_N$ are introduced as basic variables and the tableau of the following linear programming problem

$$\begin{cases} \max_{\mathbf{x}}{}_{l} \; \sum_{s=1}^{N} \sum_{a=1}^{m(s)} x_{sa}(\rho) r(s,a) \\ \sum_{s=1}^{N} \sum_{a=1}^{m(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,a)] \, x_{s,a}(\rho) + z_{s'}(\rho) =_l 1, \quad s' \in \mathcal{S} \\ x_{s,a}(\rho) \geq_l 0, \quad s \in \mathcal{S}, \; a \in A(s) \end{cases} \tag{2.6}$$

is built. Then, N successive pivot operations on all the artificial variables are carried out so that the feasibility of the solution is preserved. We call *entering variables* the basic variables of the tableau at the end of the first phase. In the *second phase* the columns of the tableau associated to the artificial variables $z_1, \ldots, z_N$ (which are now all non-basic) are removed and the simplex method is performed in the ordered field $F(\mathbb{R})$ on the obtained tableau.

We note that another approach for the solution of the parametric linear program (2.5) is given by simplex method in the field of Laurent series (see Filar, Altman and Avrachenkov, 2002 [33]).

The optimal Blackwell stationary pure strategy $\mathbf{f}^*$ is computed as:

$$\mathbf{f}^*(a|s) = \frac{x^*_{s,a}(\rho)}{\sum_{a=1}^{m(s)} x^*_{s,a}(\rho)} \quad \forall\, s \in \mathcal{S},\ a \in A(s) \tag{2.7}$$

where $\{x^*_{s,a}(\rho)\ \forall\, s, a\}$ is the solution of the optimization problem. The simplex method guarantees that the optimum strategy $\mathbf{f}^*$ is well-defined and pure (see Filar and Vrieze 1997 [32]).

2.1.5 Uniform optimality in perfect information games

As we said before, in a perfect information game in each state at most one player has more than one action available. A stationary strategy for the player $i = 1, 2$ is a function $\mathbf{f}_i : \mathcal{S} \to \bigcup_{k=1}^{N} A_i(s_k)$ with $f_i(.|s_t) \in A_i(s_t)$.

Theorem 2.1.7. *For a Competitive MDP with perfect information, both players possess uniform discount optimal pure stationary strategies, which are optimal for the average criterion as well.*

The Theorem 2.1.7 (see Filar and Vrieze, 1997 [32]) guarantees the existence of the optimal strategies for both players in the average criterion for games with perfect information. Moreover, it suggests that in order to find the optimal strategies for the average criterion one has to find the optimal strategies in the discounted criterion for a discount factor sufficiently close to 1.

Definition 8. *We call two pure stationary strategies adjacent if and only if they differ only in one state.*

Then the following property holds, which proof is analogous to the one in the field of real numbers.

Lemma 2.1.8. *Let* $\mathbf{g}$ *be a strategy for player 2 and* $\mathbf{f}, \mathbf{f}_1$ *be two adjacent strategies for player 1. Then either* $\mathbf{\Phi}_{(\rho)}(\mathbf{f}_1, \mathbf{g}) \geq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g})$ *or* $\mathbf{\Phi}_{(\rho)}(\mathbf{f}_1, \mathbf{g}) \leq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g})$*, which means that the two vectors are partially ordered.*

The property above allows us to give the following definition.

Definition 9. *Let* $(\mathbf{f}, \mathbf{g})$ *be a pair of pure stationary strategy respectively for player 1 and 2. We call* $\mathbf{f}_1$ $(\mathbf{g}_1)$ *a uniform adjacent improvement for player 1 (2) in state* s_t *if and only if* $\mathbf{f}_1$ $(\mathbf{g}_1)$ *is a pure stationary strategy which differs from* $\mathbf{f}$ $(\mathbf{g})$ *only in state* s_t *and* $\mathbf{\Phi}_{(\rho)}(\mathbf{f}_1, \mathbf{g}) \geq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g})$ $(\mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}_1) \leq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}, \mathbf{g}))$ *where the strict inequality holds in at least one component.*

As in the case in which the discount interest rate is fixed, we achieve the following results.

Lemma 2.1.9. *Let* Γ *be a perfect information Competitive MDP. A couple of pure stationary strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *is uniform discount optimal if and only if no uniform adjacent improvement is possible for both players.*

Proof. The *only if* implication is obvious. If the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are such that no uniform adjacent improvements are possible for both players, then no improvements are possible also for the first stage of the game too, that is

$$\mathbf{f}^*(s) = \operatorname*{argmax_l}_{a \in A_1(s)} \left\{ r(s,a) + (1+\rho)^{-1} \sum_{s'=1}^{N} p(s'|s,a) \mathbf{\Phi}_{(\rho)}(s', \mathbf{f}^*, \mathbf{g}^*) \right\} \qquad s \in \mathcal{S}_1$$

$$\mathbf{g}^*(s) = \operatorname*{argmin_l}_{a \in A_2(s)} \left\{ r(s,a) + (1+\rho)^{-1} \sum_{s'=1}^{N} p(s'|s,a) \mathbf{\Phi}_{(\rho)}(s', \mathbf{f}^*, \mathbf{g}^*) \right\} \qquad s \in \mathcal{S}_2$$

It is known (see Filar and Vrieze, 1997 [32]) that if the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ satisfy such equations then they are uniform discount optimal. □

In perfect information games, the following result (see Raghavan and Syed, 2002 [78]) holds

Lemma 2.1.10. *In a zero-sum, perfect information, two-player discounted Competitive MDP* Γ *with interest rate* $\overline{\rho} > 0$*, a pair of pure stationary strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *is optimal if and only if* $\mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}^*, \mathbf{g}^*) = \mathbf{\Phi}_{(\bar{\rho})}(\Gamma)$*, the value of the discounted Competitive MDP* Γ*.*

From the above result we can easily derive the analogous property in the ordered field $F(\mathbb{R})$.

Lemma 2.1.11. *In a zero-sum, two-player Competitive MDP* Γ *with perfect information, a pair of pure stationary strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are uniform discount optimal if and only if* $\mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}^*) =_l \mathbf{\Phi}^*_{(\rho)}(\Gamma) \in F(\mathbf{R})$*, where* $\mathbf{\Phi}^*_{(\rho)}(\Gamma)$ *is the uniform discount value of* Γ*.*

Proof. The *only if* statement coincides with the assertion of Theorem 2.1.1. The *if* condition is less obvious. If a pair of strategies $(\mathbf{f}^*, \mathbf{g}^*)$ has the property $\mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}^*) =_l \mathbf{\Phi}^*_{(\rho)}(\Gamma)$, then there exists $\rho^* > 0$ such that $\forall \overline{\rho} \in (0;\rho^*]$, $\mathbf{\Phi}_{(\bar{\rho})}(\mathbf{f}^*, \mathbf{g}^*)$ coincides with the value of the game Γ, $\forall \overline{\rho} \in (0;\rho^*]$. Then, thanks to Lemma 2.1.10, we can say that $\forall \overline{\rho} \in (0;\rho^*]$ the strategies $\mathbf{f}^*, \mathbf{g}^*$ are optimal in the discounted game Γ, which means that they are discount optimal. □

Let s_t be a state controlled by player i $(i = 1, 2)$ and $X \subset A_i(s_t)$. Let us call Γ^t_X the Competitive MDP which is equivalent to Γ except in state s_t, where player i has only the actions X available. Analogously to the result of Raghavan and Syed (2002, [78]), we propose the following Lemma.

Lemma 2.1.12. *Let $i=1,2$ and $s_t \in \mathcal{S}_i$, $X \subset A_i(s_t)$, $Y \subset A_i(s_t)$, $X \cap Y = \emptyset$. Then $\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_{X\cup Y}) \in F(\mathbb{R})$, which is the uniform value of the game $\Gamma^t_{X\cup Y}$, equals*

$$\begin{aligned}
\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_{X\cup Y}) &= \max\nolimits_l \{\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_X), \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_Y)\} \qquad \text{if} \quad i=1\\
\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_{X\cup Y}) &= \min\nolimits_l \{\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_X), \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_Y)\} \qquad \text{if} \quad i=2.
\end{aligned}$$

Proof. Let us suppose that the state s_t is controlled by player 2. We indicate with $\mathbf{G}^t_X$ the set of pure stationary strategies in which the choice in state s_t is restricted to the set X. We note that the restriction in state s_t does not affect player 1. Thus, $\mathbf{F}^t_X = \mathbf{F}$.
If it is possible to find optimal strategies for player 2 both in $\mathbf{G}^t_X$ and in $\mathbf{G}^t_Y$, then $\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_X) =_l \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_Y) =_l \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_{X\cup Y})$ for Lemma 2.1.11.
Otherwise, the uniform discount pure strategy of game $\Gamma^t_{X\cup Y}$ for player 2 belongs either to $\mathbf{G}^t_X$ or to $\mathbf{G}^t_Y$. For example, let us suppose that the optimal discount strategy in the Competitive MDP $\Gamma^t_{X\cup Y}$ for player 2 is found in Y. Then we have

$$\begin{aligned}
\mathbf{\Phi}^*_{(\rho)}(\Gamma^t_Y) &=_l \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_{X\cup Y})\\
&=_l \min_{\mathbf{g}\in\mathbf{G}}{}_l \max_{\mathbf{f}\in\mathbf{F}}{}_l \mathbf{\Phi}_{(\rho)}(\mathbf{f},\mathbf{g})\\
&\leq_l \min_{\mathbf{g}\in\mathbf{G^t_X}}{}_l \max_{\mathbf{f}\in\mathbf{F}}{}_l \mathbf{\Phi}_{(\rho)}(\mathbf{f},\mathbf{g})\\
&=_l \mathbf{\Phi}^*_{(\rho)}(\Gamma^t_X)
\end{aligned}$$

The proof for the situation in which $s_t \in \mathcal{S}_1$ is analogous. □

2.1.6 Algorithm description

Our task is to find an algorithm which allows to find the uniform discount optimal strategies for both players in a perfect information Competitive MDP Γ, which coincide with the optimal strategies for the long term average criterion for Theorem 2.1.7. Following the lines of the algorithm of Raghavan and Syed (2002, [78]) for optimal discount strategy, we propose an algorithm suitable to the ordered field $F(\mathbb{R})$.

Let Γ be a zero-sum two-player Competitive MDP with perfect information.

Algorithm 2.1.13.

Step 1 *Choose randomly a stationary deterministic pure strategy* $\mathbf{g}$ *for player 2.*

Step 2 *Find the Blackwell optimal strategy for player 1 in the MDP* $\Gamma_1(\mathbf{g})$ *by solving within the field* $F(\mathbb{R})$ *the following linear programming:*

$$\begin{cases} \max\limits_{\mathbf{x}} {}_l \sum_{s=1}^{N} \sum_{a=1}^{m_1(s)} x_{s,a}(\rho) r(s,a,\mathbf{g}) \\ \sum_{s=1}^{N} \sum_{a=1}^{m_1(s)} \left[(1+\rho)\delta_{s,s'} - p(s'|s,a,\mathbf{g})\right] x_{s,a}(\rho) =_l 1, \quad s' \in \mathcal{S} \\ x_{s,a}(\rho) \geq_l 0, \quad s \in \mathcal{S},\ a \in A_1(s) \end{cases} \tag{2.8}$$

and compute the pure strategy $\mathbf{f}$ *as*

$$f(a|s) = \frac{x^*_{s,a}(\rho)}{\sum_{a=1}^{m_1(s)} x^*_{s,a}(\rho)} \qquad \forall\, s \in \mathcal{S},\ a \in A_1(s) \tag{2.9}$$

where $\{x^*_{s,a}(\rho),\ \forall s, a\}$ *is the solution of (2.8).*

Step 3 *Find the minimum* k *such that in* $s_{t_1+k} \in \{s_{t_1+1}, \ldots, s_{t_1+t_2}\}$ *there exists an adjacent improvement* $\mathbf{g}'$ *for player 2, with the help of the simplex tableau associated to the following linear programming:*

$$\begin{cases} \max\limits_{\mathbf{x}} {}_l - \sum_{s=1}^{N} \sum_{a=1}^{m_2(s)} x_{s,a}(\rho) r(s,\mathbf{f},a) \\ \sum_{s=1}^{N} \sum_{a=1}^{m_2(s)} \left[(1+\rho)\delta_{s,s'} - p(s'|s,\mathbf{f},a)\right] x_{s,a}(\rho) =_l 1, \quad s' \in \mathcal{S} \\ x_{s,a}(\rho) \geq_l 0, \quad s \in \mathcal{S},\ a \in A_2(s) \end{cases} \tag{2.10}$$

where the entering variables are $\{x_{s,a} :\ g(a|s) = 1,\ \forall s\}$.
If no such improvement for player 2 is possible then go to step 4, otherwise set $\mathbf{g} := \mathbf{g}'$ *and go to step 2.*

Step 4 *Set* $(\mathbf{f}^*, \mathbf{g}^*) := (\mathbf{f}, \mathbf{g})$ *and stop. The strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are uniform discount and long term average optimal in the Competitive MDP* Γ *respectively for player 1 and player 2.*

□

Note that all the algebraic operations and the order signs $(<, >)$ are to be intended in the field $F(\mathbb{R})$.

Remark 1. *Unlike Raghavan and Syed's solution, the algorithm 2.1.13 does not require the strategy search for player 1 to be lexicographic. Player 1, in fact, faces in step 2 a classic Blackwell optimization.*

Remark 2. *Obviously, the roles of player 1 and 2 can be swapped in the algorithm 2.1.13. For simplicity, the player 1 will be assigned to step 2.*

Remark 3. *In step 3, once the state* s_{t_1+k} *is found, the adjacent improvement involves the pivoting of any of the non basic variable* $x_{s_{t_1+k},a}$ *to which corresponds a reduced cost* $c_{s_{t_1+k},a} \leq_l 0$.

Now, we prove the appropriateness of the algorithm 2.1.13. The proof is analogous to the one by Raghavan and Syed (2002, [78]).

Theorem 2.1.14. *The algorithm stops in a finite time and the couple of strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are uniform discount optimal in the Competitive MDP* Γ.

Proof. We assume that the overall number of actions

$$\mu = \sum_{k=1}^{t_1} m_1(s_k) + \sum_{k=1}^{t_2} m_2(s_{k+t_1})$$

is finite.
Without loss of generality, let us reorder the states so that in the first t_1 states player 1 has more than one action and the second t_2 states are controlled by player 2. Of course, $t_1 + t_2 \leq N$.

We can proceed by induction on μ. Trivially $\mu \geq 2N$, because $\mu = 2N$ is equivalent to the situation $t_1 = t_2 = 0$. In this case the algorithm finds the average optimal couple of strategies because it is the only existing.

Now we suppose by induction that the algorithm finds *without cycling* (that is, all pure stationary strategies are visited at most once) the couple of uniform optimal strategies when the number of actions is $\overline{\mu} \geq 2N$. We have to prove that the thesis is valid when the number of actions equals $\overline{\mu}+1$.

If $t_2 = 0$, then again there is nothing to prove, because, as we showed in section 2.1.4, the step 1 of our algorithm finds the Blackwell optimal policy $\mathbf{f}^*$ for player 1 in the MDP $\Gamma_1(g)$.

If $t_2 \geq 0$, then we focus on the state $s_{t_1+t_2} = s_\tau$, which is the last examined by our algorithm. The actions available in state s_τ are $A_2(s_\tau) \equiv X \cup a_i$, where $X = \{a_1 \dots a_{i-1}, a_{i+1} \dots a_n\}$ and $n \geq 2$ by hypothesis. By induction hypothesis, we suppose that the algorithm finds the uniform discount optimal strategies for both players in the game Γ_X^τ without cycling. Since no uniform improvements are possible in Γ_X^τ by definition of uniform optimal strategies, then the algorithm looks for an uniform adjacent improvement $\mathbf{g}'$, where $\mathbf{g}'(a_i|s_\tau) = 1$. There are now two possibilities.

If the uniform optimal strategy $\mathbf{g}$ for player 2 found in Γ_X^t is also optimal in Γ, then the algorithm terminates because still no adjacent improvements are possible for player 2 in s_t.

Otherwise, any uniform optimal strategy $\mathbf{g}^*$ for player 2 in Γ includes the action a_τ and the algorithm necessarily finds an adjacent improvement in state s_τ for Theorem 2.1.9 and it finds by induction hypothesis the uniform discount optimal strategies in the game $\Gamma_{a_n}^t$. So we have

$$\mathbf{\Phi}_{(\rho)}(\Gamma) =_l \min_l \{\mathbf{\Phi}_{(\rho)}(\Gamma_X^t), \mathbf{\Phi}_{(\rho)}(\Gamma_{a_n}^t)\} =_l \mathbf{\Phi}_{(\rho)}(\Gamma_{a_n}^t)$$

where the second equality holds because otherwise the optimal strategies of Γ_X^t would be uniform optimal in the game Γ for Lemma 2.1.11. Again thanks

to Lemma 2.1.11, we can assert that the uniform discount optimal strategies $(\mathbf{f}^*, \mathbf{g}^*)$ found in $\Gamma^t_{a_n}$ are optimal also for Γ, because $\mathbf{\Phi}_{(\rho)}(\mathbf{f}^*, \mathbf{g}^*) = \mathbf{\Phi}^*_{(\rho)}(\Gamma)$, which is the uniform discount value of the game.
Moreover, the algorithm terminates because for Theorem 2.1.9 no improvements are available to both players.

We gave a constructive proof of the fact that the algorithm passes through a path of pure strategies, it never cycles and it finds the uniform discount optimal strategies for both players. Since the overall number of actions is finite, then also the cardinality of pure strategies is finite; hence, the algorithm must terminate in a finite time and the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are uniform discount optimal, and for Theorem 2.1.7 they are long term average optimal as well. □

Computing the optimality range factor

The algorithm presented in section 2.1.6 suggests a way to determine the range of discount factor in which the long term average optimal strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are also optimal in the discounted game. Before, we report the analogous result to Lemma 2.1.9 when the discount factor is fixed (see Raghavan and Syed, 2002 [78]).

Lemma 2.1.15. *Let Γ be a perfect information Competitive MDP and $\overline{\beta} \in [0;1)$. The pure stationary strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are $\overline{\beta}$-discount optimal if and only if no uniform adjacent improvements are possible for both players in the $\overline{\beta}$-discounted Competitive MDP Γ.*

Let us define with $\zeta(f_{(\rho)})$, where $f_{(\rho)} \in F(\mathbb{R})$, the set of positive roots of $f_{(\rho)}$ such that $\frac{df_{(\rho)}}{d\rho}\,|_{\rho=u} < 0$, $\forall\, u \in \zeta(f_{(\rho)})$. Now we are ready to state the following Lemma.

Lemma 2.1.16. *Let C be the set of the reduced costs associated to the two optimal tableaux obtained at the step 2 and 3 of the last iteration of the algorithm 2.1.13 and*

$$\overline{\rho}^* = \min_c \zeta(c), \quad c \in C.$$

Then, $\overline{\beta}^* = (1+\overline{\rho}^*)^{-1}$ is the smallest value such that the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are $\overline{\beta}$-discount optimal in the game Γ, $\forall\, \overline{\beta} \in [\overline{\beta}^*; 1)$.

Proof. The existence of such $\overline{\rho}^*$ is guaranteed by Theorem 2.1.7. For all the value of the interest factor $\overline{\rho} \in (0; \overline{\rho}^*]$, the reduced costs are positive, hence no adjacent improvements are possible for both players. So, for Lemma 2.1.15 they are discounted optimal. If $\overline{\rho} > \overline{\rho}^*$ and $\overline{\rho}^* < \infty$, then at least one reduced cost is negative, hence at least an adjacent improvement is possible and $(\mathbf{f}^*, \mathbf{g}^*)$ are not $\overline{\beta}$-discount optimal, where $\overline{\beta} = (1+\overline{\rho})^{-1}$. □

Round-off errors sensitivity

The role of the first non-null coefficients of the polynomials (numerator and denominator) of the tableaux obtained throughout the algorithm unfolding is essential: they determine the positiveness of the elements of the tableaux themselves in the field $F(\mathbb{R})$. This knowledge is fundamental to choose the most suitable pivot elements.
The reader can easily understand that the algorithm is highly sensitive to the round-off errors that affect the null coefficients.

If the data of the problem (rewards and transition probabilities for each strategy) are rational, then it is possible to work in the exact arithmetic and such unconveniences are completely avoided. In fact, if all the input data are rational, they will stay rational after the algorithm execution.

Instead, if the data are irrational, a simple way to circumvent the round-off errors is to fix a tolerance value ϵ, and set to 0 all the polynomial coefficients of the tableaux obtained throughout the algorithm whose absolute value is smaller than ϵ.

2.1.7 An example

Here we present a run of our algorithm 2.1.13, where the input data are taken from Raghavan and Syed (2002, [78]). There are 5 states, the first two are controlled by player 1 and states 3 and 4 are for player 2; in the final state both players have no action choice. The immediate rewards and the probability transitions for every couple (state,action) for both players are shown in table 2.1.

We choose the initial strategy $(g(a_2|s_3) = 1, g(a_3|s_4) = 1)$ for player 2. We report the optimum tableau obtained by player 1 at the end of step 2 of the first iteration of our algorithm (tab.2.4) and the tableau of player 2 after the first improvement at step 3 (tab.2.5). Analogously, the tableaux 2.6 and 2.7 are associated to the second and last iteration of our algorithm. It is known (see Hordijk et al. 1985, [41]) that all the elements of simplex tableaux have a common denominator, stored in the top left-hand box. The last column of each tableau contains the numerator of the value of the basic variables, which are listed in the first column. The last row indicates the numerator of the reduced costs.

The optimum long term average strategy for player 1 is $f^*(a_1|s_1) = 1, f^*(a_2|s_2) = 1$, and for player 2 is $g^*(a_2|s_3) = 1, g^*(a_1|s_4) = 1$.

By computing the first positive root of the reduced costs of the two last optimal tableaux we find that the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are also $\overline{\beta}$-discount optimal for all the discount factor $\beta \in [\overline{\beta}^*; 1)$, where $\overline{\beta}^* \cong 0.74458$.

Table 2.1: Immediate rewards and transition probabilities for each player, state and strategy.

	(s,a)	r	$p(s_1\|s)$	$p(s_2\|s)$	$p(s_3\|s)$	$p(s_4\|s)$	$p(s_5\|s)$
pl. 1	(1,1)	5	0	0	0	0	1
	(1,2)	4	0	0	0.2	0	0.8
	(1,3)	3	0	0	0.6	0	0.4
	(2,1)	6	0	0	0	0	0.1
	(2,2)	1	1	0	0	0	0
	(2,3)	0	0	0	0.1	0	0
pl. 2	(3,1)	4	0	0	0	0.9	0.1
	(3,2)	2	0.1	0	0	0	0
	(3,3)	0	0.3	0	0.2	0.5	0
	(4,1)	2	0	0.1	0.6	0.3	0
	(4,2)	2	0.2	0	0.4	0.4	0
	(4,3)	3	0	0	0	0.9	0.1
	5	0	0	0.1	0.2	0.3	0.4

Note that the optimal strategies differ from the ones of Raghavan and Syed (2002, [78]), in which the discount factor is set to 0.999. We suspect that this is due to some clerical errors.

2.1.8 A lower complexity algorithm

Let Γ be a zero-sum two-player Competitive MDP with perfect information. Consider the following algorithm:

Algorithm 2.1.17.

Step 1 *Choose a stationary pure strategy* $\mathbf{g}_0$ *for player 2. Set* $k\!:=\!0$.

Step 2 *Find the Blackwell optimal strategy* $\mathbf{f}_k$ *for player 1 in the MDP* $\Gamma_1(\mathbf{g}_k)$.

Step 3 *If* $\mathbf{g}_k$ *is Blackwell optimal in* $\Gamma_2(\mathbf{f}_k)$, *then set* $(\mathbf{f}^*,\mathbf{g}^*) := (\mathbf{f}_k,\mathbf{g}_k)$ *and stop. Otherwise, find the Blackwell optimal strategy* $\mathbf{g}_{k+1}$ *for player 2 in the MDP* $\Gamma_2(\mathbf{f}_k)$, *set* $k\!:=\!k+1$ *and go to step 2.*

This is essentially a best reponse algorithm, in which at each step each player alternatively looks for his own Blackwell optimal strategy.
Obviously, if the above algorithm stops, $(\mathbf{f}^*,\mathbf{g}^*)$ forms a couple of uniform discount and long term average optimal strategies, since they are both Blackwell optimal in the respective MDPs, $\Gamma_1(\mathbf{g}^*)$ and $\Gamma_2(\mathbf{f}^*)$.
The proof that the algorithm 2.1.17 never cycles is still an open problem. It is quite natural to try to prove that $\mathbf{\Phi}_{(\rho)}(\mathbf{f}_{k+1},\mathbf{g}_{k+1}) \leq_l \mathbf{\Phi}_{(\rho)}(\mathbf{f}_k,\mathbf{g}_k)$, but it

is not difficult to find a counterexample.
Raghavan and Syed (2002, [78]) conjecture as follows:

Conjecture 2.1.1. *Let Γ be a two-player zero-sum Competitive MDP with perfect information and $\alpha = (\mathbf{f}, \mathbf{g})$ a couple of pure stationary strategies for the 2 players. For every discount factor $\overline{\beta} \in [0; 1)$, there are no sequences $\alpha_0, \alpha_1, \ldots, \alpha_k$ such that $\mathbf{\Phi}_{(\bar{\beta})}(\alpha_k) = \mathbf{\Phi}_{(\bar{\beta})}(\alpha_0)$, where α_i is an adjacent improvement with respect to α_{i-1} in the $\overline{\beta}$-discounted Competitive MDP Γ for only one player for any $i > 0$.*

If Conjecture 2.1.1 were valid, then we could conclude that the algorithm 2.1.17 terminates in finite time.

Complexity

In our first algorithm 2.1.13, player 1 faces at each step an MDP optimization problem in the field of rational functions with real coefficients, which is solvable in polynomial time. Player 2, instead, is involved in a lexicographic search throughout the algorithm unfolding, whose complexity is at worst exponential in time.

Player 2 lexicographically expands his search of his optimum strategy, and at the k-th iteration the two players find the solution of a subgame Γ_k which monotonically tends to the entire Competitive MDP Γ.

Analogously to what Raghavan and Syed (2002, [78]) remark, we can assert that the efficiency of our algorithm 2.1.13 is mostly due to the fact that most of the actions dominate totally other actions. In other words, it occurs very often that the optimum action $a^* \in A(s)$, $s \in \mathcal{S}$, found in an iteration k such that $A(s) \subset \Gamma_k$, is optimum also in Γ, and consequently remains the same in all the remaining iterations. This exponentially reduces the policy space in which the algorithm needs to search.

Remark 4. *As discussed in section 2.1.6, in the algorithm 2.1.13 players' roles are interchangeble. Since most of the actions dominate totally other actions, we suggest to assign the step 2 of the algorithm to the player whose total number of available actions is greater.*

Differently from Raghavan and Syed (2002, [78]), the search for player 1 does not need to be lexicographic, and player 1 is left totally free to optimize the MDP that he faces at each iteration of the algorithm in the most efficient way.

Let us compare in terms of number of pivoting the following three algorithms:

$\mathbf{M_1}$: Algorithm 2.1.13, in which in step 2 player 1 pivots with respect to the variable with the minimum reduced cost until he finds his own Blackwell optimal strategy.

$\mathbf{M_2}$: Algorithm 2.1.13, in which in step 2 player 1 pursues a lexicographic search, pivoting iteratively with respect to the *first* non-basic variable with a negative (in the field $F(\mathbb{R})$) reduced cost. This method is analogous to the one shown by Raghavan and Syed (2002, [78]), but in the field $F(\mathbb{R})$.

$\mathbf{M_3}$: Algorithm 2.1.17.

The results are shown in tables 2.2 and 2.3. The simulations were carried out on 10000 randomly generated Competitive MDPs with 4 states, 2 for player 1 and 2 for player 2. In each state 5 actions are available for the controlling player.

Table 2.2: Average number of pivotings for the 3 methods.

	n. pivoting
M_1	40.59
M_2	41.87
M_3	24.93

Table 2.3: $M_i > M_j$ when, fixing the game, the number of pivotings in M_i is strictly smaller than the number of pivoting in M_j.

$>$ (%)	M_1	M_2	M_3
M_1	-	52.85	18.57
M_2	42.18	-	15.26
M_3	80.05	82.75	-

It is evident that the algorithm M_3 is much faster than the other two, but unfortunately its convergence is not proven yet. However, in our numerical experiment with 10000 randomly generated Competitive MDPs, it never cycles. The difference between M_1 and M_2 is due to the more efficient simplex method used by player 1 in M_1.

2.2 Transient games

Let $p_t(s'|s)$ be the probability that the process is in state s' at time t given that s is the initial state. Let us give the definition of transient games.

Definition 1. *A stochastic game is transient if and only if* $\sum_{t=0}^{\infty}\sum_{s'\in S} p_t(s'|s,\mathbf{f},\mathbf{g})$ *is finite for all* $s \in S$ *and for any pair of stationary strategies* $(\mathbf{f},\mathbf{g})$.

Here we present the result of this section.

Theorem 2.2.1. *The uniform optimal strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *for a transient stochastic game with perfect information are optimal in the undiscounted criterion, i.e.* $\overline{\beta} = 1$*, as well.*

Proof. The uniform optimal strategies are still optimal when $\overline{\rho} \downarrow 0$, since the reduced costs of the tableaux built at the end of Algorithm 2.1.13 are non-negative when $\overline{\rho} \downarrow 0$. We know from [32] that, for transient stochastic games, the reward associated to each pair of stationary strategies $(\mathbf{f}, \mathbf{g})$ is finite. By invoking Abel's Theorem on power series [46], we claim that the reward associated to any stationary $(\mathbf{f}, \mathbf{g})$ tends to the undiscounted reward when $\overline{\rho} \downarrow 0$. Hence, the saddle-point relation (2.3) is still valid when $\overline{\rho} = 0$ and $(\mathbf{f}^*, \mathbf{g}^*)$ are optimal in the undiscounted criterion as well. □

2.3 Conclusions

In this section we dealt with zero-sum Competitive Markov Decision Processes with two players and perfect information, i.e. each state is either of nature or controlled by only one player. A finite set of states and actions is considered. There exist strategies for both players which are uniform optimal, i.e. at the Nash equilibrium for all β sufficiently close to 1. Hence, the optimal value belongs to the non-Archimedean ordered field $F(\mathbb{R})$ of rational functions with real coefficients. According to this ordering, a rational function $f \in F(\mathbb{R})$ is said to be not smaller than f' when $f(x) \geq f'(x)$ for all x sufficiently close to 0. We proposed Algorithm 2.1.13 to compute the uniform optimal strategies for both players by extending an approach by Raghavan and Syed (2002, [78]) for fixed discount factor to the field $F(\mathbb{R})$. This new method exploits linear programming techniques for MDPs developed by Hordijk, Dekker, and Kallenberg (1985, [41]). Algorithm 2.1.13 is proven to converge to the uniform optimal strategies in a finite time. We then developed Algorithm 2.1.17, which is a best response one. According to extensive simulations, Algorithm 2.1.17 requires a smaller number of pivoting operations to reach the optimal solution than Algorithm 2.1.13 does. Nevertheless, we could only conjecture that Algorithm 2.1.17 does not cycle. As a by-product, both proposed algorithms also produce the range of optimality of uniform optimal strategies, i.e. the interval of discount factors β in which such strategies are at Nash equilibrium. The uniform optimal strategies are also optimal in the average criterion. In transient games, they are optimal in the undiscounted criterion as well, i.e. when $\beta = 1$.

Table 2.4: Optimum tableau for player 1 at the first iteration.

$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,3}$	
$x_{1,1}$	$0.0198+0.6698\rho+3.06\rho^2+5.11\rho^3+3.7\rho^4+\rho^5$	$0.0234+0.6934\rho+3.04\rho^2+5.07\rho^3+3.7\rho^4+\rho^5$	$0.0288+0.7468\rho+2.418\rho^2+2.7\rho^3+\rho^4$	$0.0297+0.7527\rho+2.413\rho^2+2.69\rho^3+\rho^4$	$0.087+1.707\rho+4.42\rho^2+3.8\rho^3+\rho^4$
$x_{2,2}$	$0.0018+0.022\rho+0.042\rho^2+0.02\rho^3$	$0.0054+0.066\rho+0.126\rho^2+0.06\rho^3$	$0.027+0.756\rho+3.149\rho^2+5.12\rho^3+3.7\rho^4+1\rho^5$	$0.0279+0.767\rho+3.17\rho^2+5.13\rho^3+3.7\rho^4+1\rho^5$	$0.059+\rho+2.75\rho^2+2.8\rho^3+\rho^4$
$x_{3,1}$	$-0.084\rho-0.402\rho^2-0.5\rho^3-0.2\rho^4$	$-0.252\rho-1.206\rho^2-1.5\rho^3-0.6\rho^4$	$0.018+0.196\rho+0.158\rho^2-0.02\rho^3$	$0.018+0.154\rho-0.043\rho^2-0.27\rho^3-0.1\rho^4$	$0.1+1.36\rho+3.07\rho^2+2.9\rho^3+\rho^4$
$x_{4,1}$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.162+0.522\rho+0.54\rho^2+0.18\rho^3$	$0.27+0.51\rho+0.21\rho^2-0.03\rho^3$	$0.297+0.597\rho+0.3\rho^2$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+1\rho^4$
$x_{5,1}$	$0.018+0.238\rho+0.64\rho^2+0.62\rho^3+0.2\rho^4$	$0.054+0.714\rho+1.92\rho^2+1.86\rho^3+0.6\rho^4$	$0.09+1.07\rho+1.77\rho^2+0.69\rho^3-0.1\rho^4$	$0.099+1.189\rho+2.09\rho^2+\rho^3$	$0.41+4.01\rho+6.8\rho^2+4.2\rho^3+\rho^4$
	$0.1908+1.2838\rho+3.891\rho^2+7.028\rho^3+7.53\rho^4+4.3\rho^5+1\rho^6$	$0.5544+3.1754\rho+7.945\rho^2+12.884\rho^3+13.76\rho^4+8.2\rho^5+2\rho^6$	$0.909+3.373\rho+0.229\rho^2-14.525\rho^3-25.79\rho^4-18.5\rho^5-5\rho^6$	$1.1034+7.7329\rho+22.6785\rho^2+34.089\rho^3+26.54\rho^4+9.5\rho^5+1\rho^6$	$4.924+30.709\rho+74.775\rho^2+88.29\rho^3+50.3\rho^4+11\rho^5$

Table 2.5: Optimum tableau for player 2 at the first iteration.

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{3,1}$	$x_{3,3}$	$x_{4,3}$	$x_{4,2}$	
$x_{1,1}$	$-0.0576-0.0416\rho+0.246\rho^2+0.33\rho^3+0.1\rho^4$	$-0.1404-0.5904\rho-0.93\rho^2-0.68\rho^3-0.2\rho^4$	$-0.0036+0.0964\rho+0.26\rho^2+0.16\rho^3$	$-0.0432-0.2472\rho-0.544\rho^2-0.54\rho^3-0.2\rho^4$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+1\rho^4$
$x_{2,1}$	$-0.0576-0.208\rho-0.254\rho^2-0.1\rho^3$	$-0.054-0.152\rho-0.15\rho^2-0.05\rho^3$	$-0.0036+0.056\rho+0.216\rho^2+0.26\rho^3+0.1\rho^4$	$0.0144+0.152\rho+0.356\rho^2+0.32\rho^3+0.1\rho^4$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,2}$	$1.088\rho+4.584\rho^2+6.76\rho^3+4.3\rho^4+1\rho^5$	$1.136\rho+4.416\rho^2+6.36\rho^3+4.1\rho^4+1\rho^5$	$0.368\rho+1.404\rho^2+1.6\rho^3+0.6\rho^4$	$0.192\rho+0.608\rho^2+0.6\rho^3+0.2\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$-0.432-2.442\rho-4.38\rho^2-3.27\rho^3-0.9\rho^4$	$-0.306-1.466\rho-2.46\rho^2-1.8\rho^3-0.5\rho^4$	$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$0.216+1.956\rho+5.5\rho^2+6.96\rho^3+4.2\rho^4+\rho^5$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$-0.144-0.214\rho-0.24\rho^2-0.27\rho^3-0.1\rho^4$	$-0.234-0.594\rho-0.56\rho^2-0.2\rho^3$	$-0.054-0.134\rho-0.35\rho^2-0.37\rho^3-0.1\rho^4$	$-0.072-0.292\rho-0.42\rho^2-0.2\rho^3$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+1\rho^4$
	$2.3616+14.7176\rho+35.132\rho^2+43.526\rho^3+30.65\rho^4+11.9\rho^5+2\rho^6$	$1.368+5.132\rho+4.652\rho^2-4.782\rho^3-11.87\rho^4-8.2\rho^5-2\rho^6$	$0.8496+5.1836\rho+11.99\rho^2+15.096\rho^3+11.64\rho^4+5.2\rho^5+1\rho^6$	$0.3456+1.7496\rho+3.632\rho^2+4.128\rho^3+2.6\rho^4+0.7\rho^5+2.3008e-006\rho^6$	$-12.232-56.642\rho-108.24\rho^2-105.33\rho^3-51.5\rho^4-10\rho^5$

Table 2.6: Optimum tableau for player 1 at the second iteration.

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,3}$	
$x_{1,1}$	$0.306+2.324\rho+6.018\rho^2+7.3\rho^3+4.3\rho^4+\rho^5$	$0.342+2.356\rho+5.974\rho^2+7.26\rho^3+4.3\rho^4+\rho^5$	$0.4068+2.1148\rho+4.008\rho^2+3.3\rho^3+\rho^4$	$0.4158+2.1228\rho+3.997\rho^2+3.29\rho^3+\rho^4$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+\rho^4$
$x_{2,2}$	$0.018+0.058\rho+0.06\rho^2+0.02\rho^3$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.378+2.478\rho+6.11\rho^2+7.31\rho^3+4.3\rho^4+\rho^5$	$0.387+2.507\rho+6.14\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,1}$	$-0.24\rho-0.66\rho^2-0.62\rho^3-0.2\rho^4$	$-0.72\rho-1.98\rho^2-1.86\rho^3-0.6\rho^4$	$0.288+0.436\rho+0.128\rho^2-0.02\rho^3$	$0.288+0.316\rho-0.202\rho^2-0.33\rho^3-0.1\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.162+0.522\rho+0.54\rho^2+0.18\rho^3$	$0.27+0.51\rho+0.21\rho^2-0.03\rho^3$	$0.297+0.597\rho+0.3\rho^2$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$0.126+0.586\rho+\rho^2+0.74\rho^3+0.2\rho^4$	$0.378+1.758\rho+3\rho^2+2.22\rho^3+0.6\rho^4$	$0.63+2.09\rho+2.19\rho^2+0.63\rho^3-0.1\rho^4$	$0.693+2.383\rho+2.69\rho^2+\rho^3$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+\rho^4$
	$0.504+2.818\rho+7.344\rho^2+11.15\rho^3+10.02\rho^4+4.9\rho^5+\rho^6$	$1.224+5.858\rho+13.684\rho^2+20.09\rho^3+18.44\rho^4+9.4\rho^5+2\rho^6$	$1.8+2.896\rho-8.318\rho^2-29.624\rho^3-36.71\rho^4-21.5\rho^5-5\rho^6$	$3.636+18.583\rho+41.268\rho^2+49.431\rho^3+32.21\rho^4+10.1\rho^5+\rho^6$	$12.232+56.642\rho+108.24\rho^2+105.33\rho^3+51.5\rho^4+10\rho^5$

Table 2.7: Optimum tableau for player 2 at the second iteration.

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{3,1}$	$x_{3,3}$	$x_{4,2}$	$x_{4,3}$	
$x_{1,1}$	$-0.0576-0.0416\rho+0.246\rho^2+0.33\rho^3+0.1\rho^4$	$-0.1404-0.5904\rho-0.93\rho^2-0.68\rho^3-0.2\rho^4$	$-0.0432-0.2472\rho-0.544\rho^2-0.54\rho^3-0.2\rho^4$	$-0.0036+0.0964\rho+0.26\rho^2+0.16\rho^3$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+\rho^4$
$x_{2,1}$	$-0.0576-0.208\rho-0.254\rho^2-0.1\rho^3$	$-0.054-0.152\rho-0.15\rho^2-0.05\rho^3$	$0.0144+0.152\rho+0.356\rho^2+0.32\rho^3+0.1\rho^4$	$-0.0036+0.056\rho+0.216\rho^2+0.26\rho^3+0.1\rho^4$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,2}$	$1.088\rho+4.584\rho^2+6.76\rho^3+4.3\rho^4+\rho^5$	$1.136\rho+4.416\rho^2+6.36\rho^3+4.1\rho^4+\rho^5$	$0.192\rho+0.608\rho^2+0.6\rho^3+0.2\rho^4$	$0.368\rho+1.404\rho^2+1.6\rho^3+0.6\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$-0.432-2.442\rho-4.38\rho^2-3.27\rho^3-0.9\rho^4$	$-0.306-1.466\rho-2.46\rho^2-1.8\rho^3-0.5\rho^4$	$0.216+1.956\rho+5.5\rho^2+6.96\rho^3+4.2\rho^4+\rho^5$	$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$-0.144-0.214\rho-0.24\rho^2-0.27\rho^3-0.1\rho^4$	$-0.234-0.594\rho-0.56\rho^2-0.2\rho^3$	$-0.072-0.292\rho-0.42\rho^2-0.2\rho^3$	$-0.054-0.134\rho-0.35\rho^2-0.37\rho^3-0.1\rho^4$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+\rho^4$
	$2.3616+14.7176\rho+35.132\rho^2+43.526\rho^3+30.65\rho^4+11.9\rho^5+2\rho^6$	$1.368+5.132\rho+4.652\rho^2-4.782\rho^3-11.87\rho^4-8.2\rho^5-2\rho^6$	$0.3456+1.7496\rho+3.632\rho^2+4.128\rho^3+2.6\rho^4+0.7\rho^5$	$0.8496+5.1836\rho+11.99\rho^2+15.096\rho^3+11.64\rho^4+5.2\rho^5+\rho^6$	$-12.232-56.642\rho-108.24\rho^2-105.33\rho^3-51.5\rho^4-10\rho^5$

2.4 Stochastic Games for Cooperative Network Routing and Epidemic Spread

We consider a system where several providers share the same network and control the routing in disjoint sets of nodes. They provide connection toward a unique server (destination) to their customers. Our objective is to facilitate the design the available network links and their costs such that all the network providers are interested in cooperating and none of them withdraw from the coalition. More specifically, we establish the framework of a coalition game and we apply Algorithm 2.1.13 proposed in Section 2.1 to compute the transferable coalition values. As by-product, we apply the proposed algorithm to two-player games both in networks subject to hacker attacks and in epidemic networks.

2.4.1 Introduction

Sharing resources among competitive operators is a fundamental issue in 4G wireless systems. Cooperation enables a better exploitation of the resources and promises higher revenues to network providers. However, cooperation among competitive entities is complicated by the sensitive issue of conflicting interests. Thus, it becomes imperative to motivate and guarantee a fair cooperation among these entities. This can be achieved by a careful distribution of the costs or the incremental revenues obtained by cooperating. Coalition games offer a suitable theoretical framework to address this problem.

Several providers share a network to provide connection towards a unique common destination to their customers. We provide a framework of a coalition game to facilitate the design of the available network links and their costs such that there exists an optimum routing strategy and a cost sharing satisfying all the subsets of providers. More specifically, we provide algorithms to compute the coalition values, i.e. the minimum costs that each coalition can ensure for itself. The proposed algorithm is based on some results for two-player zero-sum Competitive MDPs with perfect information in Section 2.1.

It is worth noticing that the analysed problem differs substantially from the noncooperative routing games thoroughly studied in literature (for additional details see e.g. Nisan et al. 2007, [65] and references therein). At the best of the authors' knowledge, this work is the first one applying coalition games to determine an optimum routing solution and cost allocation in a shared network.

2.4.2 Routing model

We consider a network consisting of a set of nodes $V=\{1,\ldots,N\}$. P service providers (SPs) share the network to offer their customers connection toward a single destination node N. The customers' traffic is injected in the network at $n \leq N-1$ nodes, called sources, located in nodes $\mathcal{T} = \{v_1,\ldots,v_n\} \subseteq V/\{N\}$. There is only one destination, in node N. We assume that all the sources transmit at the same rate the packets of a provider k, for all possible k. Let $c_k(i,j) > 0$ represent the cost per unit time that provider k has to sustain to convey its own packets, sent by any of the sources in $\mathcal{T}$, through the link $i \to j$.

The k-th SP controls the routing, i.e. the activation of outgoing links, in the set of nodes V_k. We suppose that a node is controlled at most by one SP, i.e., $V_i \cap V_j = \emptyset$, $\forall\, i \neq j$ and $\bigcup_i V_i \subseteq V$. Each node i is assigned a subset $\alpha_i \subseteq V$, such that the *directed* link $i \to j$ can be activated if and only if $j \in \alpha_i$. In the generic node $i \in V_k$, SP k can assign a probability distribution $\mathbf{f}_k$ to the each node $j \in \alpha_i$ such that the probability that the network link (i,j) is utilized for routing is $\mathbf{f}_k(i,j)$ at *any* routing decision moment. The destination node is a "sink", and it does not route the incoming packets to any of the other nodes. We remark that all the nodes $\{1,\ldots,N-1\}$, included the sources, serve as routing nodes.

Let $\mathbf{\Phi}_k^{(\beta)}$, with $\beta \in [0;1]$, be a N-by-1 vector whose i-th component is the expected β-discounted sum of costs for a path originating in node i, i.e.

$$\mathbf{\Phi}_k^{(\beta)}(i) = \mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_P}\left[\sum_{t\geq 0} \beta^t c_k(V_t,V_{t+1})\Big| V_0 = i\right],$$

where V_t is the t-th node crossed by the packets. It is worth noticing that, for $\beta=1$, $\mathbf{\Phi}_k^{(1)}$ is a plain sum of costs and $\mathbf{\Phi}_k^{(1)}(v_i)$ is the cost per unit time that SP k incurs for the stream of packets going from the i-th source, v_i, to the destination.

2.4.3 Routing Long-run Cooperative game

In this section we tackle the problem of cost distribution among the SPs by adopting a cooperative game theory approach. First, let us provide some preliminary notions and notations. Let $\mathcal{P} = \{1,\ldots,P\}$ be the grand coalition of SPs. Let $\mathcal{C} \subseteq \mathcal{P}$ be a coalition of players. Let us define, for any $\beta \in [0;1]$, the expected β-discounted sum

$$\mathbf{\Phi}_{\mathcal{C}}^{(\beta)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}},\mathbf{f}_{\mathcal{C}}) = \sum_{\{k\}\in\mathcal{C}} \mathbf{\Phi}_k^{(\beta)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}},\mathbf{f}_{\mathcal{C}})$$

Here we are interested in the case $\beta=1$, since the quantity $\sum_{i=1}^n \mathbf{\Phi}_{\mathcal{C}}^{(\beta=1)}(v_i)$ is the total cost per unit time that $\mathcal{C}$ incurs to sustain its $n|\mathcal{C}|$ information

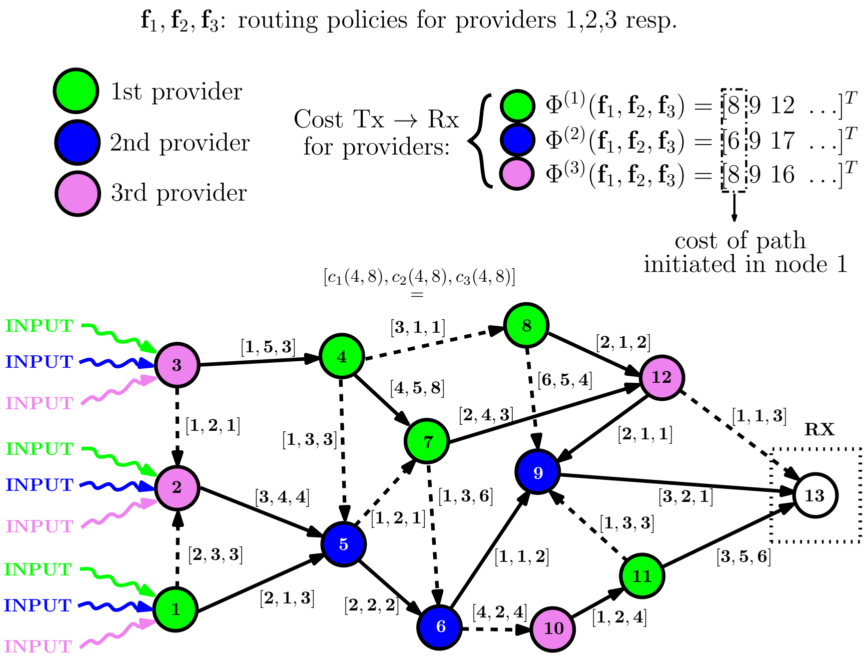
$\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$: routing policies for providers 1,2,3 resp.
1st provider
2nd provider
3rd provider
Cost Tx → Rx for providers:
$\Phi^{(1)}(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3) = [8\ 9\ 12\ \ldots]^T$
$\Phi^{(2)}(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3) = [6\ 9\ 17\ \ldots]^T$
$\Phi^{(3)}(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3) = [8\ 9\ 16\ \ldots]^T$
cost of path initiated in node 1
$[c_1(4,8), c_2(4,8), c_3(4,8)]$
=
INPUT
INPUT
INPUT
INPUT
INPUT
INPUT
INPUT
INPUT
INPUT
1
2
3
4
5
6
7
8
9
10
11
12
13
RX
[1, 5, 3]
[3, 1, 1]
[2, 1, 2]
[1, 2, 1]
[4, 5, 8]
[6, 5, 4]
[1, 3, 3]
[2, 4, 3]
[1, 1, 3]
[2, 1, 1]
[3, 4, 4]
[1, 3, 6]
[3, 2, 1]
[1, 2, 1]
[1, 3, 3]
[2, 3, 3]
[1, 1, 2]
[3, 5, 6]
[2, 1, 3]
[2, 2, 2]
[4, 2, 4]
[1, 2, 4]

Figure 2.1: Example of routing policy with 3 service providers, 3 transmitters, 13 nodes.

streams. We actually assume that all providers cooperate with each other, and coordinate their routing decisions in order to minimize the overall costs for the network. We define the value for coalition $\mathcal{P}$, $v(\mathcal{P})$, as the minimum global transmission cost:

$$v(\mathcal{P}) = \sum_{i=1}^{n} \mathbf{\Phi}_{\mathcal{P}}^{(1)}(v_i, \mathbf{F}^{\mathrm{o}}) = \min_{\mathbf{f}_{\mathcal{P}} \in \mathcal{F}_{\mathcal{P}}} \sum_{i=1}^{n} \mathbf{\Phi}_{\mathcal{P}}^{(1)}(v_i, \mathbf{f}_{\mathcal{P}})$$

where $\mathbf{F}_{\mathcal{P}}$ is the set of strategies available to the grand coalition $\mathcal{P}$ and $\mathbf{F}^{\mathrm{o}}$ is the optimum global routing strategy. Under the Transferable Utility (TU) assumption, the total cost $v(\mathcal{P})$ can be shared in any manner among the SPs, thanks to a binding agreement the members of $\mathcal{P}$. A Cooperative Game Theory approach (see Peleg and Sudhölter, 2007 [69]) suggests to assign first a value v to each coalition $\mathcal{C} \subset \mathcal{P}$ of SPs. All the cooperative solutions, like Shapley value, Core, Nucleoulus etc. (see Chapter 1), sharing $v(\mathcal{P})$ among the SPs according to different criteria, are indeed a function of $\{v(\mathcal{C})\}_{\mathcal{C} \subseteq \mathcal{P}}$.

In this section we will focus on the computation of the coalition values. In the literature, there are several ways to compute the value of a coalition. One of the most utilized is arguably the minimax one, by von Neumann and Morgenstern (1944, [96]), suggesting that the value of a coalition $\mathcal{C}$ of SPs should be computed as the minimum cost that $\mathcal{C}$ can guarantee against any routing strategy of the anti-coalition $\mathcal{P}\backslash\mathcal{C}$, i.e.

$$v(\mathcal{C}) = \min_{\mathbf{f}_{\mathcal{C}}\in\mathbf{F}_{\mathcal{C}}} \max_{\mathbf{f}_{\mathcal{P}/\mathcal{C}}\in\mathbf{F}_{\mathcal{P}/\mathcal{C}}} \sum_{i=1}^{n} \mathbf{\Phi}_{\mathcal{C}}^{(1)}(v_i, \mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}), \quad \forall\, \mathcal{C}\subset\mathcal{P}. \tag{2.11}$$

Hence, $v(\mathcal{C})$ can be interpreted as the value of the zero-sum game between $\mathcal{C}$ and the remaining SPs $\mathcal{P}\backslash\mathcal{C}$, which adopt an antagonistic behaviour towards $\mathcal{C}$. The coalition value $v(\mathcal{C})$ is a *measure of the power of a coalition of SPs*, rather than a depiction of a real antagonistic scenario. We point out the minimax approach is important because it ensures the superadditivity property of the characteristic function v:

$$v(\mathcal{C}_1) + v(\mathcal{C}_2) \geq v(\mathcal{C}_1 \cup \mathcal{C}_2),\ \forall\, \mathcal{C}_1, \mathcal{C}_2 \subset \mathcal{P},\ \mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset.$$

In the Appendix we show that Algorithm 2.1.13, developed in Section 2.1, can be utilized to compute the minimax coalition values $\{v(\mathcal{C})\}_{\mathcal{C}\subseteq\mathcal{P}}$. In Section 2.4.3 we show its adaptation to the transient case, for which no loops in the network can occur.

Remark 5. *We dub the Cooperative game described above* long-run *since the interaction among the SPs is to be meant as one-shot: the costs are shared among SPs at the beginning of the transmission, once for all, and any chance of a renegotiation is ruled out. In contrast, in Chapter 3 we will deal with a more dynamic situation, in which the payoff are distributed to the players along the game, and coalitions are allowed to form throughout the game.*

Algorithm for computing coalition values

The coalition values $v(\mathcal{C})$ may be *infinite*. If $v(\mathcal{C}) = +\infty$, then the optimal strategies for the players, i.e. the strategies at Nash equilibrium, impede at least one source-destination path by causing a loop in the network. In practice, $v(\mathcal{C}) = +\infty$ is not the cost that coalition $\mathcal{C}$ has to bear; anyway, it indicates that any service provider cannot accept to lose its own packets.
In order to *avoid infinities in the computation of coalition values*, the idea is to compute the optimal strategies $(\mathbf{f}^*_{\mathcal{P}/\mathcal{C}}, \mathbf{f}^*_{\mathcal{C}})$, for coalitions $\mathcal{P}/\mathcal{C}$ and $\mathcal{C}$ respectively, for *all* the discount factors sufficiently close to 1. Then, we adopt the strategy that is still optimal in the limit for $\beta \to 1$ to compute the coalition value.

In the Appendix, Lemma 2.4.7, we show that we are legitimated to utilize Algorithm 2.1.13 to compute $v(\mathcal{C})$, for all $\mathcal{C} \subseteq \mathcal{P}$, as the limit for $\beta \uparrow 1$ of the uniform value of the zero-sum game between $\mathcal{C}$ and $\mathcal{P}\backslash\mathcal{C}$.

Before showing the algorithm, let us refresh the reader's memory about some useful definitions. Let $\mathbf{f}_{\mathcal{P}/\mathcal{C}}$ be a pure strategy for coalition $\mathcal{P}/\mathcal{C}$. We say that the pure strategy $\mathbf{f}'_{\mathcal{C}}$ is an improvement for coalition $\mathcal{C}$ with respect to $\mathbf{f}_{\mathcal{C}}$ for the discount factor β if and only if

$$\mathbf{\Phi}_{\mathcal{C}}^{(\beta)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}'_{\mathcal{C}}) \leq \mathbf{\Phi}_{\mathcal{C}}^{(\beta)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$$

where the relation $\leq$ is component-wise and $<$ is valid for at least one node in V. Let $\Gamma_{\mathcal{P}/\mathcal{C}}(\overline{\mathbf{f}}_{\mathcal{C}})$ be the optimization problem that $\mathcal{P}/\mathcal{C}$ faces when $\mathcal{C}$ fixes its own strategy $\overline{\mathbf{f}}_{\mathcal{C}}$. Then, the optimum strategy for $\mathcal{P}/\mathcal{C}$ in $\Gamma_{\mathcal{P}/\mathcal{C}}(\overline{\mathbf{f}}_{\mathcal{C}})$ maximizes $\mathbf{\Phi}_{\mathcal{C}}^{(\beta)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \overline{\mathbf{f}}_{\mathcal{C}})$ component-wisely.

Algorithm 2.4.1. *Set $\mathcal{C} \subseteq \mathcal{P}$. Consider only pure routing strategy, i.e. only strategies $\mathbf{f}_k$ such that, for all $i \in V_k$, $\exists\, j : \mathbf{f}_k(i,j) = 1$.*

1. *Pick a pure routing strategy $\mathbf{f}_{\mathcal{C}}$ for coalition $\mathcal{C}$.*

2. *Find the best strategy $\mathbf{f}_{\mathcal{P}/\mathcal{C}}$ for coalition $\mathcal{P}/\mathcal{C}$ in the optimization problem $\Gamma_{\mathcal{P}/\mathcal{C}}(\mathbf{f}_{\mathcal{C}})$, for all the discount factors close enough to 1.*

3. *Find the first node controlled by coalition $\mathcal{C}$ in which a change of strategy $\mathbf{f}'_{\mathcal{C}}$ is an improvement for coalition $\mathcal{C}$ for all the discount factors close enough to 1. If it does not exists, then set $(\mathbf{f}^*_{\mathcal{P}/\mathcal{C}}, \mathbf{f}^*_{\mathcal{C}}) := (\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ and go to step 4. Otherwise, set $\mathbf{f}_{\mathcal{C}} := \mathbf{f}'_{\mathcal{C}}$ and go to step 2.*

4. *Compute the coalition value*

$$v(\mathcal{C}) = \lim_{\beta \to 1} \sum_{i=1}^{n} \mathbf{\Phi}_{\mathcal{C}}^{(\beta)}(v_i, \mathbf{f}^*_{\mathcal{P}/\mathcal{C}}, \mathbf{f}^*_{\mathcal{C}}).$$

We remark that the optimal strategy in step 2 and the strategy refinement in step 3 are found with the help of simplex tableaux in the non-Archimedean ordered field $F(\mathbb{R})$ of rational functions with real polynomial coefficients (for any detail, see Section 2.1).

Transient case

Under the assumption that no loops can be ever be present in the network, Algorithm 2.4.1 can be simplified by considering directly $\beta = 1$. In other words, in this section we suppose that the following assumption holds.

Assumption 1. *For any couple of pure strategies $(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ for $\mathcal{P}/\mathcal{C}$ and $\mathcal{C}$ respectively, and for all $i \in V$, there exists a path*[1] *$\tau_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ of finite length*[2] *$L_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ and without loops linking node i to the destination node N.*

The following result shows that the assumption above ensures $\mathbf{\Phi}_{\mathcal{C}}^{(1)}$ to be finite, for any couple of strategies.

[1] a path is a sequence of connected nodes

[2] the length of the path is the number of edges that it is composed of.

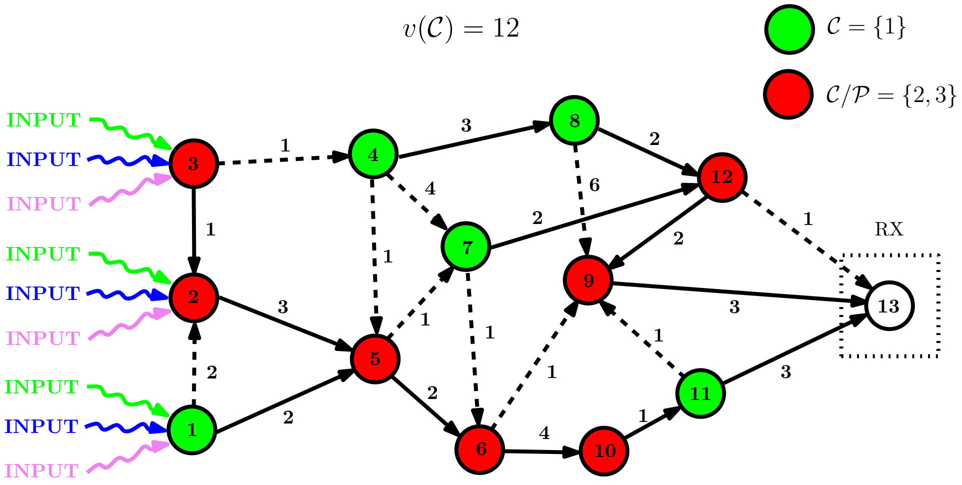

Figure 2.2: Value of singleton coalition $\{1\}$, in the routing model in Figure 2.1. The continuous arrows are the activated links. The costs are specified next to each arrow.

Proposition 2.4.2. *Suppose that assumption 1 holds. Then, for all the pure strategies* $\mathbf{f}_{\mathcal{P}/\mathcal{C}} \in \mathbf{F}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}} \in \mathbf{F}_{\mathcal{C}}$*:*

(i) the path $\tau_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ *is unique;*

(ii) $\mathbf{\Phi}_{\mathcal{C}}^{(1)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}) < +\infty$.

Proof. Let $\tau_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}) = \{V_0 = i, V_1, \dots, V_{L_i} = N\}$ be the nodes crossed by the path τ_i when $\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}$ are fixed. If there existed more than one path linking two nodes then there would exist at least one node in which more than one arc go out of it. This is impossible since the strategies are pure. Then, *(i)* is proved. Therefore, we can say that

$$\begin{cases} p_t(j|V_0 = i, \mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}) = \mathbb{I}(j = V_t), & \forall\, t \in [1; L_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})] \\ p_t(j|V_0 = i, \mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}) = 0, & \forall\, t > L_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}}) \end{cases}$$

where $p_t(j|V_0)$ is the probability that the t-th node crossed by the packets starting in node V_0 is j. Thus, $\forall\, i \in V$, the i-th component of $\Phi_{\mathcal{C}}^{(1)}(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ is bounded by

$$L_i(\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})\, |\, \mathcal{C}\, |\, \max_{k,i,j} c_k(i,j) < +\infty$$

□

If Assumption 1 holds, then Algorithm 2.4.1 can be adapted as follows (see Lemma 2.4.8).

Algorithm 2.4.3. *Set* $\mathcal{C} \subseteq \mathcal{P}$.

1. *Pick a pure routing strategy* $\mathbf{f}_{\mathcal{C}}$ *for coalition* $\mathcal{C}$.

2. *Find the best strategy* $\mathbf{f}_{\mathcal{P}/\mathcal{C}}$ *for coalition* $\mathcal{P}/\mathcal{C}$ *in the optimization problem* $\Gamma_{\mathcal{P}/\mathcal{C}}(\mathbf{f}_{\mathcal{C}})$*, for* $\beta = 1$.

3. *Find the first node controlled by coalition* $\mathcal{C}$ *in which a change of strategy* $\mathbf{f}'_{\mathcal{C}}$ *is an improvement for coalition* $\mathcal{C}$*, for* $\beta = 1$*. If it does not exists, then set* $(\mathbf{f}^*_{\mathcal{P}/\mathcal{C}}, \mathbf{f}^*_{\mathcal{C}}) := (\mathbf{f}_{\mathcal{P}/\mathcal{C}}, \mathbf{f}_{\mathcal{C}})$ *and go to step 4. Otherwise, set* $\mathbf{f}_{\mathcal{C}} := \mathbf{f}'_{\mathcal{C}}$ *and go to step 2.*

4. *Set* $v(\mathcal{C}) = \sum_{i=1}^{n} \mathbf{\Phi}^{(1)}_{\mathcal{C}}(v_i, \mathbf{f}^*_{\mathcal{P}/\mathcal{C}}, \mathbf{f}^*_{\mathcal{C}})$.

We remark that the algorithm 2.4.3 is analogous to the one described by Raghavan and Syed (2002, [78]) when $\beta = 1$ and restricted to the transient case, with the difference that in step 2 the search is not necessarily lexicographic for coalition $\mathcal{P}/\mathcal{C}$. Indeed, at each iteration $\mathcal{P}/\mathcal{C}$ is allowed to find its own temporarily optimal strategy with *any* Markov Decision Process solving method.

2.4.4 Network design

Although the network design is not our purpose, we suggest which steps could be followed in this direction.

A network designer should aim at devising both the routing decisions α_i available to each provider in each node $i \in V$ and the cost of the links $c_k(i,j)$, in order to ensure that each coalition of providers has an interest in not deviating from the global optimum policy $\mathbf{F}^{\text{o}}$. Formally, a network designer should ensure the non-emptiness of the Core of the TU (transferable utility) coalition game (P, v), i.e. that set of cost $\mathbf{g} \in \mathbf{Co}$ that providers can share among themselves through binding agreements, such that

$$\begin{cases} \sum_{k=1}^{P} g_k = v(\mathcal{P}) \\ \sum_{\{k\} \in \mathcal{C}} g_k \leq v(\mathcal{C}), \quad \forall \mathcal{C} \subset \mathcal{P}. \end{cases}$$

We see from the former equation that the Core is globally *efficient* for the network and from the latter that it is also *stable* with respect to the formation of greedy coalitions.

In the following sections we adapt the competitive game between coalition $\mathcal{C}$ and the anti-coalition $\mathcal{P}\backslash\mathcal{C}$, examined just to compute the coalition value $\mathcal{C}$, to three other different scenarios.

2.4.5 Hacker-Provider routing game

The routing competitive game between two conflicting coalitions described in Section 2.4.2 can also be re-interpreted in the framework of the conflicts

between one service provider and one hacker.
There is a set $V_1 \subseteq V$ of vulnerable nodes, where the routing control may be got hold by a hacker. V_0 is the set of nodes in which the routing is handled by a service provider. The set $V_2 = V_0/V_1$ is the set of unattackable nodes among the ones controlled by the service provider. Each link $i \to j$ is assigned $c(i,j) > 0$, that in this case can be also interpreted as a *delay*, i.e. the time that a packet of provider k spends to go from node i to node j. In such a case, let us assume that the nodes are capable to re-direct all the incoming packets as soon as they receive them, without any additional delay due to the buffering. The service provider here wants to find the routing rule that *jointly* minimizes the packet delay $\mathbf{\Phi}^{(1)}$ for all the sources; conversely, the hacker wants to slow down the network.

As in Section 2.4.2, there may be some couple of strategies for the two players for which there exist loops in the network, that cause the packet delay from some sources to be infinite. Note that the hacker can also disrupt some nodes, by forcing a loop on them. Hence, here we deal with the general case of undiscounted Competitive MDPs described in the Appendix. The undiscounted optimal strategies can be computed by the algorithm 2.4.1, in which player 1 is now the hacker which controls nodes V_1, and player 2 is the service provider, which controls nodes V_2.

Note that in this case, in contrast with the coalition game, we are more interested in the computation of the optimal strategies, than in the value of the game at the Nash equilibrium. Indeed, *the optimal strategy for the service provider is the pure routing policy it should adopt in order to minimize the source-wise packet delay in the worst case scenario.* The worst situation for the provider is when the hacker is able to control all the vulnerable nodes V_1 and has at its disposal as many routing policies as possible. Note that the optimal strategies for both players are pure, i.e. the routing policy is deterministic in each node.

2.4.6 Natural disaster

Let us reformulate the model described in Section 2.4.5, where player 1 is now a natural agent that can put out of order some nodes $V_1 \subset V$ of the network, independently of the routing action taken by the service provider in such nodes. This addresses the practical situation in which nodes V_1 are located in areas subject to catastrophic natural phenomena. It is straightforward to see that the computation of the optimal strategies for the service provider boils down to the calculation of a Markov Decision Process uniform optimal solution (see Hordijk at al., 1985 [41]), in which the set of nodes of interest is reduced to V_2, that is the collection of nodes controlled by the service provider.

2.4.7 Epidemic network

In this section we describe our third and last competitive game scenario, inspired by the routing game between conflicting coalitions. We model an epidemic network with N nodes; $N-1$ possibly infected individuals are located in nodes $\{1,\ldots,N-1\}$ respectively. Each individual can infect, with some probability, only one among a subset of other individuals in its neighbourhood. There is a probability $\boldsymbol{\mu}_i$ that the infection process starts from the i-th individual. The infection spread terminates when the virus reaches the healer, located in node N. Hence, there is a probability $\boldsymbol{\mu}_N$ that the epidemic spread is averted. There are two player: player 2, the "good" one, wants to design and force the connections among the individuals such that the lowest expected number of individuals are infected, while player 1 has the opposite goal. The assumption of perfect information still holds, i.e. the set of nodes in which player 1 and player 2 have more than one action available are disjoint.

The formulation of the problem is analogous to the two-player game described in Section 2.4.2, in which the cost of the link (i,j) is 1 for all nodes i,j. The nodes are substituted by the individuals, the destination with the healer, the sources become the first infected entity, the packet routing is replaced by the virus transmission. In this context, we wish $\sum_{i=1}^{N}\mu_i\mathbf{\Phi}^{(1)}(i)$ to represent the average number of infected individuals. Therefore, for each couple of routing strategies, *no loops* in the network are allowed, i.e. we suppose that the Assumption 1 holds. Hence, thanks to Proposition 2.4.2, for every couple of pure stationary strategies $(\mathbf{f},\mathbf{g})$,

$$\sum_{i=1}^{N}\mu_i\mathbf{\Phi}^{(1)}(i,\mathbf{f},\mathbf{g}) \tag{2.12}$$

is actually the expected number of infected individuals.

We can use the algorithm 2.4.3 to find the optimal strategy for the "good" player, who is interested in minimizing the objective function (2.12). If $(\mathbf{f}^*,\mathbf{g}^*)$ are the undiscounted optimal strategies, then the value

$$\sum_{i=1}^{N}\mu_i\mathbf{\Phi}^{(1)}(i,\mathbf{f}^*,\mathbf{g}^*)$$

is the worst-case estimate for player 2 for the expected number of infected individuals.

2.4.8 Conclusions

Several providers share the same network and control the routing in disjoint sets of nodes. There are several information sources and one destination.

By using the framework of Competitive MDPs, we provided algorithms to compute the minimum costs that each coalition of providers can ensure for itself. This helps the optimum design of a network, which should guarantee the existence of an efficient and stable costs partition among the providers. We also modelled situations in which there are two players with conflicting interests, like a hacker against a service provider, or in which a service provider wants to reduce the damages to the network caused by a natural disaster. An epidemic spread network model was shown as well. From a theoretical perspective, we extended some results on uniform optimal strategies in Competitive MDP to the case of undiscounted criterion.

2.4.9 Appendix

Minimax routing game as a Competitive MDP

Indeed, our routing game between $\mathcal{C} \subset \mathcal{P}$ and $\mathcal{P} \backslash \mathcal{C}$ can be interpreted as a zero-sum Competitive MDP with 2 players and perfect information. For this purpose we adopt the same notation utilized in Section 2.1. Player 2 is the coalition $\mathcal{C} \subset \mathcal{P}$, while player 1 is the rest of the providers $\mathcal{P}/\mathcal{C}$. There exist a bijective association between the network nodes V and the states $\mathcal{S}$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the set of states associated to the set of nodes $\bigcup_{\{k\} \in \mathcal{P}/\mathcal{C}} V_k$ and to $\bigcup_{\{k\} \in \mathcal{C}} V_k$, respectively. The network link $i \to j$ is activated if and only if player k selects the action $a_j^{(k)}(s_i)$, where $j \in \alpha_i$, $k : s_i \in \mathcal{S}_k$. The instantaneous reward $r(s_i, a_j^{(k)}(s_i)) = \sum_{\{p\} \in \mathcal{C}} c_p(i,j)$, where k is the player that controls the node i. The transition probability is $p(s_w | s_i, a_j^{(k)}(s_i)) = \mathbb{I}(w = j)$, where $\mathbb{I}$ is the indicator function. Note that $\sum_{s' \in \mathcal{S}} p(s' | s, \mathbf{f}, \mathbf{g}) = 1$, $\forall\, s \in \mathcal{S}/\{s_N\}$ and for each couple of stationary strategies $(\mathbf{f}, \mathbf{g})$. The destination node is a "sink", i.e. $p(s_i | s_N) = 0$, $\forall\, i \in [1; N]$, and no actions are available in it for both players.

Undiscounted criterion with positive rewards

We want to prove that Algorithm 2.4.1 actually computes the value of any coalition. Before, let us state two important Theorems.

Theorem 2.4.4 (Abel's Theorem on power series)**.** *Let the power series* $h(x) = \sum_{n=0}^{\infty} a_n x^n$ *have radius of convergence* r *and still converge for* $x = r$*. Then,* $\lim_{x \uparrow r} h(x) = h(r)$*.*

Theorem 2.4.5 ([46])**.** *Let* $\sum_{k \geq 0} c_k$ *be a divergent series of positive terms. Then*

$$\lim_{x \uparrow 1} \sum_{k \geq 0} x^k c_k = +\infty$$

Now we can we state as follows.

Corollary 2.4.6. *Let $\sum_{k\geq 0} c_k$ be a series of positive terms and $\xi \in \mathbb{R}$. Then*

$$\begin{cases} \lim_{x\uparrow 1} \sum_{k\geq 0} x^k c_k = \xi & \iff \sum_{k\geq 0} c_k = \xi \\ \lim_{x\uparrow 1} \sum_{k\geq 0} x^k c_k = +\infty & \iff \sum_{k\geq 0} c_k = +\infty \end{cases}$$

Proof. For the *if* conditions, see Theorems 2.4.4, 2.4.5.
About the *only if* conditions, we know (see Knopp, 1990 [46]) that a positive term series either converges or diverges to $+\infty$. If $\sum_{k\geq 0} c_k = \xi_1 \neq \xi$, then $\lim_{x\uparrow 1} \sum_{k\geq 0} x^k c_k = \xi_1$ for Theorem 2.4.4. Hence, both the ($\Leftarrow$) relations are proved by contradiction. □

Now we are ready to state the following result.

Lemma 2.4.7. *Suppose that all the instantaneous rewards are non-negative. Let us utilize the extended line of real numbers, i.e. treat $\pm\infty$ as a number ($\pm\infty = \pm\infty$, $-\infty < a \in \mathbb{R} < +\infty$). Then, the uniform optimal strategies are optimal in the undiscounted criterion as well, i.e.*

$$\mathbf{\Phi}^{(1)}(\mathbf{f}, \mathbf{g}^*) \leq \mathbf{\Phi}^{(1)}(\mathbf{f}^*, \mathbf{g}^*) \leq \mathbf{\Phi}^{(1)}(\mathbf{f}^*, \mathbf{g}) \quad \forall\, \mathbf{f}, \mathbf{g} \tag{2.13}$$

Hence, Algorithm 2.4.1 actually computes the value of any coalition.

Proof. By definition, the saddle point relation (2.13) is valid $\forall\, \beta \in [\beta; 1)$ and hence also for the limit $\beta \uparrow 1$. Then, it is still valid for $\beta = 1$ for Corollary 2.4.6. □

Transient games

Definition 2. *Let $p_t(.|s)$ be the transition probability from state s after t steps. A Competitive MDP **transient** if*

$$\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} p_t(s'|s, \mathbf{f}, \mathbf{g}) < +\infty \tag{2.14}$$

for each $s \in \mathcal{S}$ and all pure stationary strategies $\mathbf{f}$ and $\mathbf{g}$.

Note that the Competitive MDP in Section 2.4.3 is transient.

Lemma 2.4.8. *Algorithm 2.4.3 provides the undiscounted optimal strategies for transient Competitive MDPs.*

Proof. In transient Competitive MDPs with bounded instantaneous payoffs, the undiscounted reward is also bounded, for each couple of stationary strategies (see Filar and Vrieze, 1997 [32]). Furthermore, under the transient condition, the uniform optimal strategies are optimal under the undiscounted criterion as well (see Section 2.1). It is straightforward to prove that all the elements, belonging to $F(\mathbb{R})$, of the simplex tableaux built throughout the algorithm 2.4.1 are right continuous in $\rho = 0$ (or, equivalently, left continuous in $\beta = 1$). Therefore, we are allowed to shift the ordered field on which the algorithm works from $F(\mathbb{R})$ to $\mathbb{R}$, with $\beta = 1$. □

Chapter 3

Dynamic Cooperative MDPs

3.1 Cooperative MDPs: Time Consistency, Greedy Players Satisfaction, and Cooperation Maintenance

We deal with multi-agent Markov Decision Processes (MDPs) in which cooperation among players is allowed. Unlike Section 2.4, coalitions may form throughout the game, and a payoff needs to be assigned to the players during the game. We find a Cooperative Payoff Distribution Procedure on MDPs (MDP-CPDP) that distributes in the course of the game the payoff that players would earn in the long run game. We show under which conditions such a MDP-CPDP fulfills a time consistency property, contents greedy players, and strengthen the coalition cohesiveness throughout the game. Finally we refine the concept of Core for Cooperative MDPs, by utilizing the notion of Cooperation Maintenance.

3.1.1 Introduction

In static cooperative game theory, in which only one static game is played, the main challenge is to devise a procedure that shares the total reward earned by the whole community of players among the players themselves, and that complies with an agreeable definition of "fairness" (e.g. Peleg and Sudhölter 2007, [69]). When the interaction among the players is reiterated over time, it is reasonable to assume that the players demand to be rewarded in the course of the game, and the issue of designing such an allocation procedure has drawn much attention in the last few decades, especially in the field of cooperative differential games. Such games address the realistic situation in which the interaction among several players (e.g. countries, firms, business partners etc.) spans a certain period of time, and the environment in which the players operate (commonly called "state") changes according to a differential equation. A contract signed by all the players dictates how to share a certain payoff among the participants during the game. Bulk of the literature on cooperative differential games deals with the design of a payoff distribution procedure fulfilling a sensible time consistency property, under which no coalition of players is enticed to breach the agreement at any of the stage of the game (see Zaccour, 2008 [102] and references therein).

A different situation is considered by repeated cooperative games, which model situations in which the *same* game is repeatedly played over time and players can cooperate and form coalitions throughout the duration of the game. The papers by Oviedo (2000, [66]) and by Kranich, Perea, and Peters (2001, [47]) are the two independent pioneering works in this field.

While the theory of competitive Markov Decision Processes (MDPs), otherwise called non-cooperative stochastic games, has been thoroughly studied (Filar and Vrieze, 1997 [32] for an extensive survey), to the best of the authors' knowledge, there is very little work on cooperative MDPs in the literature. Unlike classic repeated games, in which the same game is played repeatedly over time, in cooperative MDPs several *different* static games follow one another. Unlike differential games, in our model the static games follow a discrete-time Markov chain, whose transition probabilities depend on the players' actions in each state. Players can decide whether to join the grand coalition or, throughout the game, to form coalitions. The payoff earned by a coalition is, under the transferable utility (TU) assumption, shared among its participants. Once a group of players has withdrawn from the grand coalition, it cannot rejoin it later on.

Petrosjan (2006, [73]), in his pioneering work, proposed a time consistent cooperative payoff distribution procedure (CPDP) in cooperative games on finite trees. In this paper we deal with discount cooperative MDPs, in which the payoffs at each stage are multiplied by a discount factor and summed up over time. Our game model is in fact more general than the one by Petrosjan (2006, [73]), since we allow for cycles on the state space and we do not impose the finiteness of the game horizon. We also point out that our model is different from the one proposed by Predtetchinski (2007, [75]), since we assume that the utility of the coalitions is transferable and the probability transitions among the static games does depend on the players' actions in each stage.

The paper is organized as follows. Section 3.1.2 is a short survey on non-cooperative and cooperative multi-agent MDPs. Following the lines of Petrosjan's work, in Section 3.1.3 we propose a stationary stage-wise CPDP for cooperative discounted MDPs (MDP-CPDP). In Section 3.1.4 we prove that the MDP-CPDP satisfies what we call the "terminal fairness property", i.e. the expected discounted sum of payoff allocations belongs to a cooperative solution (i.e. Shapley Value, Core, etc.) of the whole discounted game. In Section 3.1.5 we show that the MDP-CPDP fulfills the time consistency property, which is a crucial one in repeated games theory (e.g. Filar and Petrosjan, 2000 [31]): it suggests that a payoff distribution procedure should respect the terminal fairness property in a sub-game starting from any state, at any time step. In Section 3.1.6 we show that, under some conditions, for all discount factors small enough, also the greedy players having a myopic perspective of the game are satisfied with the MDP-CPDP. In Section 3.1.7 we deal with perhaps the most meaningful attribute for a CPDP, which is the n-tuple step cooperation maintenance property. It claims that, at each stage of the game, the long run reward that each group of players expects to gain by withdrawing from the grand coalition after n step should be less than what it would earn by sticking to the grand coalition forever. In some sense, if such a condition is fulfilled for all integers n's, then no players are enticed to withdraw from the grand coalition. We find that the single step cooperation maintenance property, earliest introduced in a deterministic setting by Mazalov and Rettieva (2010, [59]), is the strongest one among all n's. Furthermore, we give a necessary and sufficient condition, inspired by the celebrated Bondareva-Shapley Theorem (Bondareva, 1963 [22]; Shapley 1967, [83]), for the existence of an MDP-CPDP satisfying the n-tuple step cooperation maintenance property, for any integer n. Inspired by this property, we propose a refinement of the Core solution concept for cooperative MDPs, dubbed as "Cooperation Maintaining solution". Finally, Section 3.1.9 deals with a special case of our model, entailing that the transition probabilities among the states do not depend on the players' strategies.

A lexical remark. We define the "stage" of the game at time t as the random state that the game finds itself in at time t.

Some notation remarks. The ordering relations $<, >$, if referred to vectors, are component-wise, as well as the max and min operators. The entry that lies in the i-th row and in the j-th column of matrix $\mathbf{A}$ is written as $\mathbf{A}_{i,j}$. An equivalent notation for the n-by-m matrix $\mathbf{A}$ is $[\mathbf{A}_{i,j}]_{i=1,j=1}^{n,m}$. The i-th element of column vector $\mathbf{a}$ is denoted by $\mathbf{a}_i$. The expression $\text{val}(\mathbf{A})$ stands for the value (e.g. Filar and Vrieze, 1997 [32]) of the matrix $\mathbf{A}$. Let $\{C_i\}_i$ be a collection of sets; we define the sum set $\sum_i C_i$ as $\{\sum_i c_i : \ c_i \in C_i, \ \forall\, i\}$.

3.1.2 Discounted Cooperative MDPs

In a multi-agent Markov Decision Process (MDP) Γ with $P > 1$ players there is a finite set of states $\mathcal{S} := \{s_1, s_2, \ldots, s_N\}$, and for each state s the set of actions available to the i-th player is denoted by $A_i(s)$, $i = 1, \dots, P$, and $|A_i(s)| := m_i(s)$. To each $(P+1)$-tuple $(s, a_1, \dots, a_P)$, with $a_i \in A_i(s)$, an immediate reward $r_i(s, a_1, \dots, a_P)$ for player $i = 1, \dots, P$ and a transition probability distribution $p(.|s, a_1, \dots, a_P)$ on the state space $\mathcal{S}$ are assigned. Hence, in each state s the static game $\Omega_s \equiv (\mathcal{P}, A_i(s), r_i(s, .))$ is played, and the states succeed one another following a Markov chain controlled by the players' actions.

Let $\mathcal{P} := \{1, \dots, P\}$ be the grand coalition. We assume that any subset of players $\Lambda \subseteq \mathcal{P}$ can withdraw from the grand coalition and form a coalition at stage of the game, and all the players are compelled to play throughout the whole duration of the game. Moreover, once a coalition is formed, it can no longer rejoin the grand coalition in the future.

Let $A_\Lambda(s) := \prod_{i \in \Lambda} A_i(s)$ be the set of actions available to coalition Λ in state s, for all $s \in \mathcal{S}$. A stationary strategy $\mathbf{f}_\Lambda$ for the coalition Λ is a probability distribution on $A_\Lambda(s)$, such that $\mathbf{f}_\Lambda(a|s)$ is the probability that the coalition Λ chooses the action $a \in A_\Lambda(s)$ in state s. We define $\mathbf{F}_\Lambda$ as the set of stationary strategies for coalition $\Lambda \subseteq \mathcal{P}$. Let Λ_1,Λ_2 two disjoint nonempty coalitions. Then, $\mathbf{F}_{\Lambda_1} \cup \mathbf{F}_{\Lambda_2} \subset \mathbf{F}_{\Lambda_1 \cup \Lambda_2}$. If for every $s \in \mathcal{S}$ there exists $a(s)$ such that $\mathbf{f}_\Lambda(a(s)|s) = 1$, then the stationary strategy $\mathbf{f}_\Lambda$ is dubbed "pure".
Let us define the transition probability distribution on the state space $\mathcal{S}$, given the independent strategies $\mathbf{f}_\Lambda \in \mathbf{F}_\Lambda$, $\mathbf{f}_{\mathcal{P}\setminus\Lambda} \in \mathbf{F}_{\mathcal{P}\setminus\Lambda}$, as

$$p(s'|s, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda}) := \sum_{a_\Lambda \in A_\Lambda(s)} \sum_{a_{\mathcal{P}\setminus\Lambda} \in A_{\mathcal{P}\setminus\Lambda}(s)} p(s'|s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda})\, \mathbf{f}_\Lambda(a_\Lambda|s)\, \mathbf{f}_{\mathcal{P}\setminus\Lambda}(a_{\mathcal{P}\setminus\Lambda}|s),$$

for all $s, s' \in \mathcal{S}$. Analogously, let $r_i(s, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda})$ be the expected instantaneous reward for player i in state s. Let

$$r_\Lambda(s, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda}) := \sum_{i \in \Lambda} r_i(s, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda})$$

be the bounded and deterministic reward gained by the coalition Λ in state s. We assume that the rewards are geometrically discounted over time, and $\beta \in [0; 1)$ is the discount factor. We define $\Phi_\Lambda^{(\beta)}(s, .)$ as the expected β-discounted long run reward for coalition $\Lambda \subseteq \mathcal{P}$ when the initial state of the game is s_k:

$$\Phi_\Lambda^{(\beta)}(s, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda}) := \mathbb{E}\left(\sum_{t=0}^{\infty} \beta^t r_\Lambda(S_t, \mathbf{f}_\Lambda, \mathbf{f}_{\mathcal{P}\setminus\Lambda}) \big| S_0 = s \right) \quad \forall\, s \in \mathcal{S},$$

where S_t is the stage of the game at time t. Hence, we can write the vector

$\mathbf{\Phi}_\Lambda^{(\beta)}(.) := [\Phi_\Lambda(s_1,.),\ldots,\Phi_\Lambda(s_N,.)]^T$ as

$$\begin{aligned}\mathbf{\Phi}_\Lambda^{(\beta)}(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda}) &= \sum_{t=0}^{\infty}\beta^t\mathbf{P}^t(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda})\,\mathbf{r}_\Lambda(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda})\\ &= \left[\mathbf{I}-\beta\mathbf{P}^t(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda})\right]^{-1}\mathbf{r}_\Lambda(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda}),\end{aligned} \tag{3.1}$$

where $\mathbf{P}(\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda})$ is the N-by-N transition probability matrix and $\mathbf{r}_\Lambda(.) := [r_\Lambda(s_1,.),\ldots,r_\Lambda(s_N,.)]^T$. Let $\mathbf{f}_{\mathcal{P}}^{(\beta)*}$ be the global optimum strategy for the grand coalition $\mathcal{P}$, i.e.

$$\mathbf{f}_{\mathcal{P}}^{(\beta)*} = \operatorname*{argmax}_{\mathbf{f}_{\mathcal{P}}\in\mathbf{F}_{\mathcal{P}}}\mathbf{\Phi}_{\mathcal{P}}^{(\beta)}(\mathbf{f}_{\mathcal{P}}), \qquad \forall\,\beta\in[0;1), \tag{3.2}$$

where the maximization is component-wise. For simplicity of notation, we will denote $\mathbf{P}^{*(\beta)} := \mathbf{P}(\mathbf{f}_{\mathcal{P}}^{(\beta)*})$, which is the transition probability matrix associated to the global optimal stationary strategy $\mathbf{f}_{\mathcal{P}}^{(\beta)*}$, whose (i,j) element is $p(s_j|s_i,\mathbf{f}_{\mathcal{P}}^{(\beta)*})$.

Let Γ_s be the long run game Γ starting in state $s\in\mathcal{S}$. For any $\beta\in[0;1)$ and for every state s, we assign to each coalition Λ a value $v^{(\beta)}(\Lambda,\Gamma_s)\in\mathbb{R}$. Under the transferable utility (TU) condition, the value of a coalition can be shared in any manner among the members of the coalition itself. Hence, the set of feasible allocations for coalition $\Lambda\subseteq\mathcal{P}$ in the game Γ_s is $\mathcal{V}^{(\beta)}(\Lambda,\Gamma_s)$, where

$$\mathcal{V}^{(\beta)}(\Lambda,\Gamma_s) := \left\{\mathbf{x}\in\mathbb{R}^{|\Lambda|} : \sum_{i\in\Lambda}x_i\le v^{(\beta)}(\Lambda,\Gamma_s)\right\}.$$

It is widely accepted to assign to the empty coalition a null utility, i.e.

$$v^{(\beta)}(\{\emptyset\},\Gamma_s) = 0.$$

Throughout the paper, if not specified, we always consider nonempty coalitions. We consider the value associated to the grand coalition $v^{(\beta)}(\mathcal{P},\Gamma_s)$ to be the biggest achievable discounted sum of reward in the game Γ_s:

$$v^{(\beta)}(\mathcal{P},\Gamma_s) = \Phi_\Lambda^{(\beta)}(s,\mathbf{f}_{\mathcal{P}}^{(\beta)*}).$$

In many applications it makes sense to define the coalition value $v^{(\beta)}(\Lambda,\Gamma_s)$ as the maximum total reward that coalition Λ can ensure for itself in the β-discounted long run game Γ_s (von Neumann and Morgenstern, 1944 [96]), i.e.

$$v^{(\beta)}(\Lambda,\Gamma_s) := \max_{\mathbf{f}_\Lambda\in\mathbf{F}_\Lambda}\ \min_{\mathbf{f}_{\mathcal{P}\setminus\Lambda}\in\mathbf{F}_{\mathcal{P}\setminus\Lambda}}\ \Phi_\Lambda^{(\beta)}(s,\mathbf{f}_\Lambda,\mathbf{f}_{\mathcal{P}\setminus\Lambda}) \tag{3.3}$$

Nevertheless, we will consider the specific value formulation in (3.3) solely in Sections 3.1.6 and 3.1.8. Next we provide some useful definitions and preliminary results.

Definition 3 (Linear combination of games)**.** *Let $\mathcal{V}(\Delta_i, \Lambda)$ be the set of feasible allocations for the coalition $\Lambda \subseteq \mathcal{P}$ in the game Δ_i, for $i = 1, \ldots, N$. The linear combination $\sum_i b_i \Delta_i$ is a game in which the set of feasible allocations for the coalition Λ, $\mathcal{V}(\sum_i b_i \Delta_i, \Lambda)$, equals the Minkowski sum $\sum_i b_i \mathcal{V}(\Delta_i, \Lambda)$.*

Proposition 3.1.1. *Let $\Delta_1, \ldots, \Delta_N$ be N games with transferable utilities. Let $v(\Lambda, \Delta_i)$ be the value of coalition $\Lambda \subseteq \mathcal{P}$ in the game Δ_i. Let $b_i \geq 0$, for all $i = 1, \ldots, N$. Then, $\sum_i b_i \Delta_i$ is a TU game in which the value of the coalition $\Lambda \subseteq \mathcal{P}$ is*

$$v\left(\Lambda, \sum_{i=1}^{N} b_i \Delta_i\right) = \sum_i b_i v(\Lambda, \Delta_i).$$

Proof. Let

$$\widetilde{\mathcal{V}}(\Lambda) := \left\{ \mathbf{x} \in \mathbb{R}^P : \sum_{i:\{i\} \in \Lambda} x_i \leq \sum_i b_i v(\Lambda, \Delta_i) \right\}.$$

We have to prove that, for all $\Lambda \subseteq \mathcal{P}$, $\mathcal{V}(\sum_i b_i \Delta_i, \Lambda) = \sum_i b_i \mathcal{V}(\Delta_i, \Lambda) = \widetilde{\mathcal{V}}(\Lambda)$. Let the real $|\Lambda|$-tuple $\mathbf{c}(i) \in \mathcal{V}(\Delta_i, \Lambda)$, for all i. It is straightforward to see that $\sum_i b_i \mathbf{c}(i) \in \widetilde{\mathcal{V}}(\Lambda)$. Then, $\sum_i b_i \mathcal{V}(\Delta_i, \Lambda) \subseteq \widetilde{\mathcal{V}}(\Lambda)$. Let us fix the real P-tuple $\widetilde{\mathbf{c}} \in \widetilde{\mathcal{V}}(\Lambda)$. We define $I := \{i : b_i > 0\}$. We need to find $\{\mathbf{c}'(i) \in \mathcal{V}(\Delta_i, \Lambda)\}_{i \in I}$ such that $\sum_{i \in I} b_i \mathbf{c}'(i) = \widetilde{\mathbf{c}}$. Let $\mathbf{c}'_j(i) = \widetilde{\mathbf{c}}_j / (|I| b_i)$ for all j such that $\{j\} \notin \Lambda$. To determine the remaining $|I||\Lambda|$ elements $\{\mathbf{c}'_j(i),\ \forall\, i \in I,\ j : \{j\} \in \Lambda\}$, we introduce the following set of inequalities:

$$\begin{cases} \sum_{i \in I} b_i \mathbf{c}'_j(i) = \widetilde{\mathbf{c}}_j & \forall\, j : \{j\} \in \Lambda \\ \sum_{j:\{j\} \in \Lambda} \mathbf{c}'_j(i) \leq v(\Lambda, \Delta_i) & \forall\, i \in I \end{cases} \tag{3.4}$$

Let us prove that (3.4) admits a solution. Let $\epsilon_i \geq 0$, for all $i \in I$, be such that

$$\sum_{i \in I} \epsilon_i = \sum_{i \in I} b_i v(\Lambda, \Delta_i) - \sum_{j:\{j\} \in \Lambda} \widetilde{\mathbf{c}}_j \geq 0 \tag{3.5}$$

We write the following linear system

$$\begin{cases} \sum_{i \in I} b_i \mathbf{c}'_j(i) = \widetilde{\mathbf{c}}_j & \forall\, j : \{j\} \in \Lambda \\ b_i \sum_{j:\{j\} \in \Lambda} \mathbf{c}'_j(i) = b_i v(\Lambda, \Delta_i) - \epsilon_i & \forall\, i \in I \end{cases} \tag{3.6}$$

Evidently, any solution to (3.6) is also a solution to (3.4). Thanks to (3.5), the sum of the first $|\Lambda|$ equations of (3.6) equals the sum of the remaining $|I|$ equations. By discarding the last equation of (3.6) we obtain a linear system with $|\Lambda| + |I| - 1$ linearly independent equations in $|\Lambda||I| > |\Lambda| + |I| - 1$ unknowns. Hence, a solution to (3.6) exists and $\sum_i b_i \mathcal{V}(\Delta_i, \Lambda) \supseteq \widetilde{V}(\Lambda)$. Then, $\sum_i b_i \mathcal{V}(\Delta_i, \Lambda) = \widetilde{\mathcal{V}}(\Lambda)$ and the thesis is proven. □

Still, we could consider the long run game Γ_s as a classic static cooperative game, solely characterized by the set of players $\mathcal{P}$ and the coalition values $v^{(\beta)}$. Therefore we can still assign to it a classic solution concept.

Definition 4 (Terminal cooperative solution)**.** *Set $\beta \in [0;1)$. The terminal cooperative solution $\mathbf{T}^{(\beta)}(\Gamma_s)$ is a set-valued function which represents a static cooperative solution (e.g. Shapley value, Core, etc.) of the long run game Γ_s starting in state s, i.e.*

$$\mathbf{T}^{(\beta)}(\Gamma_s) \ : \ \{v^{(\beta)}(\Lambda, \Gamma_s)\}_{\Lambda \subseteq \mathcal{P}} \ \to \ \mathbb{R}^P, \qquad \forall\, s \in \mathcal{S}.$$

Analogously, we define $\mathbf{T}^{(\beta)}(\sum_i b_i \Gamma_{s_i})$ as the terminal cooperative solution of the cooperative game with coalition values $\{v^{(\beta)}(\Lambda, \sum_i b_i \Gamma_{s_i})\}_{\Lambda \subseteq \mathcal{P}}$.

The terminal cooperative solution $\mathbf{T}^{(\beta)}$ can represent any of the classical cooperative solutions. For example, $\mathbf{T} \equiv \mathbf{Co}$ represents the Core of the β-discounted game Γ_s, that is the set, possibly empty, of the real P-tuples $\mathbf{x}$ satisfying

$$\begin{cases} \sum_{i \in \mathcal{P}} x_i = v^{(\beta)}(\mathcal{P}, \Gamma_s) \\ \sum_{i \in \Lambda} x_i \geq v^{(\beta)}(\Lambda, \Gamma_s), \ \forall\, \Lambda \subset \mathcal{P}. \end{cases} \tag{3.7}$$

A game with nonempty Core is said to be balanced. The strict Core $\mathbf{sCo}^{(\beta)}(\Gamma_s)$ is defined as in (3.7), but with the strict inequality signs.
The terminal cooperative solution $\mathbf{T} \equiv \mathcal{S}h^{(\beta)}(\Gamma_s)$ stands for the Shapley value of the β-discounted game Γ_s, i.e. for all $i = 1, \ldots, P$,

$$\mathcal{S}h_i^{(\beta)}(\Gamma_s) = \sum_{\Lambda \subseteq \mathcal{P}/\{i\}} \frac{|\Lambda|!\,(P-|\Lambda|-1)!}{P!} \left[v^{(\beta)}(\Lambda \cup \{i\}, \Gamma_s) - v^{(\beta)}(\Lambda, \Gamma_s) \right].$$

We finally present a linearity property of the Core and Shapley value.

Proposition 3.1.2. *Let $\Delta_1, \ldots, \Delta_N$ be games with transferable utilities with non empty Cores $\mathbf{Co}(\Delta_1), \ldots, \mathbf{Co}(\Delta_N)$, respectively. Let $b_1, \ldots, b_N$ be non negative coefficients. Then, $\sum_{i=1}^N b_i \mathbf{Co}(\Delta_i) \subseteq \mathbf{Co}(\sum_{i=1}^N b_i \Delta_i)$.*

Proof. Let $x_1(i), \ldots, x_P(i)$ be an allocation belonging to the Core $\mathbf{Co}(\Delta_i)$. Thanks to the linearity property of coalition values shown in Proposition 3.1.1, we can write

$$\begin{aligned} &\sum_{i=1}^N \sum_{k \in \mathcal{P}} b_i x_k(i) = \sum_{i=1}^N b_i v(\mathcal{P}, \Delta_i) = \ v\left(\mathcal{P}, \sum_{i=1}^N b_i \Delta_i\right) \\ &\sum_{i=1}^N \sum_{k \in \Lambda} b_i x_k(i) \geq \sum_{i=1}^N b_i v(\Lambda, \Delta_i) = \ v\left(\Lambda, \sum_{i=1}^N b_i \Delta_i\right), \qquad \forall\, \Lambda \subset \mathcal{P}. \end{aligned}$$

Hence, the thesis is proven. □

Corollary 3.1.3. *For all* $\beta \in [0;1)$, $\sum_{i=1}^{N} b_i \mathcal{S}h^{(\beta)}(\Gamma_{s_i}) = \mathcal{S}h^{(\beta)}(\sum_{i=1}^{N} b_i \Gamma_{s_i})$, *where* $b_i \geq 0, \ \forall\, i$.

Proof. The proof follows straightforward from Proposition 3.1.1 and from the linearity property of the Shapley value. □

3.1.3 Cooperative Payoff Distribution Procedure

In cooperative MDPs, different static games follow one another in time. If we conceive the dynamic game as a whole, the payoff allocation issue boils down to the computation of the terminal cooperative solution $\mathbf{T}^{(\beta)}(\Gamma_s)$, and the players are rewarded a certain amount $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{T}^{(\beta)}(\Gamma_s)$ at the *end* the game. Of course, if the length of game is not finite, the players need to be rewarded throughout the game. Even if the game has a limited duration, though, the players may not be willing to wait until its conclusion before receiving a payoff (e.g. wage earners). Our goal is then to build a connection between static and dynamic cooperative game theory on Markov Decision Processes, by devising a procedure which distributes the terminal solution $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s)$ throughout the game, in each of its stages. With respect to static cooperative game theory, an additional complication here lies in satisfying, or at least being fair with, all the players at each stage of the game, since *coalitions are allowed to form throughout the game unfolding.* Moreover we assume that, *once a coalition has formed, it cannot rejoin the grand coalition later on.*

Remark 6. *All the results presented in the current section, as well as the ones in Sections 3.1.4, 3.1.5, 3.1.7, 3.1.9 can be easily extended to undiscounted transient MDPs, i.e. games for which* $\beta = 1$ *and*

$$\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} p_t(s'|s, \mathbf{f}_{\mathcal{P}}) < \infty, \quad \forall\, s \in \mathcal{S},\ \mathbf{f}_{\mathcal{P}} \in \mathbf{F}_{\mathcal{P}}. \tag{3.8}$$

where $p_t(s'|s) = p(S_t = s'|S_0 = s)$ *is the probability of being in state* s' *at the* t*-th step, knowing that the starting state was* s. *In fact the reader should notice that, mathematically speaking, introducing a discount factor* $\beta \in [0;1)$ *is equivalent to multiplying each transition probability by* β*, which automatically ensures the transient condition (3.8).*

Let us now define the concept of cooperative payoff distribution procedure, which is crucial in this paper.

Definition 5 (CPDP). *The cooperative payoff distribution procedure (CPDP)* $g^{(\beta)} := [g_1^{(\beta)}, \dots, g_P^{(\beta)}]$ *is a recursive function that, for each time step* $t \geq 0$*, associates a real* P*-tuple* $g^{(\beta)}(\mathbf{h}_t)$ *to the past history* $\mathbf{h}_t = [S_0, g^{(\beta)}(\mathbf{h}_0), S_1, \dots, g^{(\beta)}(\mathbf{h}_{t-1}), S_t]$ *of states succession and stage-wise allocations up to time* t.

The following are two alternative interpretations for $g_i^{(\beta)}$:

i) $\beta^t g_i^{(\beta)}(\mathbf{h}_t)$ is the payoff that player $i \in \mathcal{P}$ gains at the stage t of the game, when $\mathbf{h}_t$ is the history of the process;

ii) $g_i^{(\beta)}(\mathbf{h}_t)$ is the payoff that player i obtains at time t when the transition probabilities are discounted by a factor β, i.e. we consider a new distribution $p'(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*}) = \beta p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$, for all $s, s' \in \mathcal{S}$. Hence, $1 - \beta$ is the stopping probability in each state.

Next we provide a definition of stationary CPDP's. Let $\mathcal{H}_t$ the class of state and allocation histories up to time t.

Definition 6 (Stationarity)**.** *Set* $\beta \in [0; 1)$*. A CPDP* $g^{(\beta)}$ *is stationary whenever* $g^{(\beta)}(\mathbf{h}_t) = g^{(\beta)}(S_t = s) := g^{(\beta)}(s)$*, for all* $t \geq 0$ *and* $\mathbf{h}_t \in \mathcal{H}_t$*.*

Hence, a stationary CPDP $g^{(\beta)} : \mathcal{S} \to \mathbb{R}^P$ is a stage-wise payoff distribution law that does not depend on the whole history, but only on the last observable state of the process.

In his pioneering work, Petrosjan (2006, [73]) introduced a CPDP for games on finite trees. Following his lines, we now propose a stationary CPDP for cooperative MDPs (MDP-CPDP) with β-discounted criterion, with $\beta \in [0; 1)$ fixed *a priori*.

Definition 7 (MDP-CPDP)**.** *Set* $\beta \in [0; 1)$*. Select the a terminal cooperative solution* $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{T}^{(\beta)}(\Gamma_s)$*,* $\forall\, s \in \mathcal{S}$*. The cooperative payoff distribution procedure* $\gamma^{(\beta)}$ *on MDP (MDP-CPDP) associated to* $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s)$ *is defined as*

$$\gamma^{(\beta)}(s, \overline{\mathbf{T}}) := \sum_{s' \in \mathcal{S}} \left[\delta_{s,s'} - \beta\, p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\right] \overline{\mathbf{T}}^{(\beta)}(\Gamma_{s'}), \qquad \forall\, s \in \mathcal{S}. \tag{3.9}$$

Throughout the paper, we will not specify the dependence of $\gamma^{(\beta)}$ on $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s)$ when this is clear from the context.
In Section 3.1.4 it will be clear to the reader that not all the stationary CPDP are MDP-CPDP, but only those whose expected β-discounted long run summation is actually a terminal cooperative solution. In the next sections we will study some appealing properties of the MDP-CPDP, defined as in (3.9).

3.1.4 Terminal Fairness

In this section, we let the terminal cooperative solution $\mathbf{T}$ be any of the classic cooperative solution (Core, Shapley value, Nucleolus, etc.). In the following we will propose two desirable properties for a CPDP and we prove that the MDP-CPDP defined in (3.9) fulfills both of them.

Firstly, we wish to guarantee a natural continuity between static cooperative game theory and dynamic payoff allocation. Hence, we require the expected discounted sum of the stage-wise allocations to equal the terminal cooperative solution of the game, as formalized in the following.

Property 1 (Terminal fairness). *Set $\beta \in [0;1)$. The CPDP $g^{(\beta)}$ is said to be terminal fair w.r.t. the terminal cooperative solution $\overline{\mathbf{T}}^{(\beta)}$ whenever $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s)$ is stage-wisely distributed in the course of the game, i.e.*

$$\mathbb{E}\Big[\sum_{t\geq 0} \beta^t g^{(\beta)}(\mathbf{h}_t)|S_0 = s\Big] \in \mathbf{T}^{(\beta)}(\Gamma_s), \quad \forall\, s \in \mathcal{S}.$$

Now we show that the proposed MDP-CPDP can be defined axiomatically, as the only stationary allocation that fulfills the terminal fairness property. Hence, $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ establishes a bijective relation between a terminal cooperative solution $\overline{\mathbf{T}}$ and a stage-wise allocation procedure $\gamma^{(\beta)}$.

Theorem 3.1.4. *The MDP-CPDP $\gamma^{(\beta)}(s,\overline{\mathbf{T}}) \in \mathbb{R}^P$, defined in (3.9) is the unique stationary CPDP that satisfies the terminal fairness property w.r.t. $\overline{\mathbf{T}}^{(\beta)}$, for all $\beta \in [0;1)$.*

Proof. We know that, for all $i \in \mathcal{P}$,

$$\begin{bmatrix} \mathbb{E}[\sum_{t\geq 0} \beta^t \gamma_i^{(\beta)}(S_t)|S_0 = s_1] \\ \vdots \\ \mathbb{E}[\sum_{t\geq 0} \beta^t \gamma_i^{(\beta)}(S_t)|S_0 = s_N] \end{bmatrix} = \Big[\mathbf{I} - \beta \mathbf{P}^{*(\beta)}\Big]^{-1} \begin{bmatrix} \gamma_i^{(\beta)}(s_1) \\ \vdots \\ \gamma_i^{(\beta)}(s_N) \end{bmatrix}.$$

If we substitute (3.9) in the equation above, we find that $\gamma_i^{(\beta)}$ defined in (3.9) satisfies the relation:

$$\mathbb{E}\Big[\sum_{t\geq 0} \beta^t \gamma^{(\beta)}(S_t)|S_0 = s\Big] = \overline{\mathbf{T}}^{(\beta)}(\Gamma_s), \;\; \forall\, s \in \mathcal{S},\;\; i \in \mathcal{P}.$$

Since the matrix $\sum_{t\geq 0}[\beta \mathbf{P}^{*(\beta)}]^t = [\mathbf{I} - \beta \mathbf{P}^{*(\beta)}]^{-1}$ is invertible, then such $\gamma^{(\beta)}$ is also unique. Hence, the thesis is proven. □

In each state s of the game, the grand coalition receives a total payoff $r_{\mathcal{P}}(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$. In principle, only a portion of it could be shared among the players, and accordingly the remaining part is allocated in the following stages of the game. We point out that this procedure would require the presence of an external "regulator" agent, managing the payoff stream. In this work we want to rule out this possibility, thus we demand that, in each state s, the whole amount $r_{\mathcal{P}}(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$ is shared among the players. We call this property *stage-wise efficiency.* In order to ensure such a property surely, we also have to ensure that the instantaneous rewards are deterministic. This is straightforward to obtain, since $\mathbf{f}_{\mathcal{P}}^{(\beta)*}$ can be found in the class of pure policies.

Property 2 (Stage-wise efficiency)**.** *Set $\beta \in [0;1)$. The CPDP $g^{(\beta)}$ is stage-wise efficient whenever $\sum_{i\in\mathcal{P}} g_i^{(\beta)}(s) = \sum_{i\in\mathcal{P}} r_i(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$ for all $s \in \mathcal{S}$, where $\mathbf{f}_{\mathcal{P}}^{(\beta)*}$ is the global optimum pure stationary strategy.*

Theorem 3.1.5. *The MDP-CPDP $\gamma^{(\beta)}$, defined in (3.9), fulfills the stage-wise efficiency property, for all $\beta \in [0;1)$.*

Proof. The global optimum strategy $\mathbf{f}_{\mathcal{P}}^{(\beta)*}$ is pure, since the optimization problem (3.2) that it solves can be formulated as a Markov Decision Process (Puterman, 1994 [77]). Hence, $r_i(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$ is also deterministic, for all $i \in \mathcal{P}$. Let us sum (3.9) over all possible $i \in \mathcal{P}$, for all $s \in \mathcal{S}$, and we obtain:

$$v^{(\beta)}(\mathcal{P}, \Gamma_s) = \sum_{i\in\mathcal{P}} \gamma_i^{(\beta)}(s) + \beta \sum_{s'\in\mathcal{S}} p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\, v^{(\beta)}(\mathcal{P}, \Gamma_{s'}).$$

Since the following is also valid for all $s \in \mathcal{S}$ from the definition of $v^{(\beta)}$:

$$v^{(\beta)}(\mathcal{P}, \Gamma_s) = \sum_{i\in\mathcal{P}} r_i(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*}) + \beta \sum_{s'\in\mathcal{S}} p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\, v^{(\beta)}(\mathcal{P}, \Gamma_{s'}),$$

then, $\sum_{i\in\mathcal{P}} \gamma_i^{(\beta)}(s) = \sum_{i\in\mathcal{P}} r_i(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$, surely. □

It is straightforward to verify that the MDP-CPDP $\gamma^{(\beta)}$ defined in (3.9) also fulfills a *terminal efficiency* property, i.e.

$$\sum_{i\in\mathcal{P}} \mathbb{E}\Big[\sum_{t\geq 0} \beta^t \gamma_i^{(\beta)}(S_t|S_0 = s)\Big] = v^{(\beta)}(\mathcal{P}, \Gamma_s), \qquad \forall\, s \in \mathcal{S}.$$

3.1.5 Time Consistency

Time consistency is a well known concept in dynamic cooperative theory (Filar and Petrosjan, 2000 [31], Zaccour, 2008 [102] and references therein). It captures the idea that the stage-wise allocation must respect the terminal fairness property even from a later starting time of the game, for any possible trajectory of the game up to that instant. In other words, players are never enticed to renegotiate the agreement on CPDP at any intermediate time step, because even if they did, assuming that cooperation has prevailed from the initial date until that instant, then the payoff distribution procedure would remain the same. Let us adopt the convention $\mathbf{h}_{-1} = \emptyset$. The time consistency property can be formalized as follows.

Property 3 (Time consistency)**.** *Set $\beta \in [0;1)$. A CPDP $g^{(\beta)}$ is time consistent w.r.t. a terminal cooperative solution $\mathbf{T}^{(\beta)}$ whenever, for all $t \geq 0$ and for all possible allocation/state histories $\mathbf{h}_t \in \mathcal{H}_t$,*

$$\mathbb{E}\left[\sum_{k=t}^{\infty} \beta^k g^{(\beta)}(S_k, \mathbf{h}_{k-1}) \Big| \mathbf{h}_t\right] \in \beta^t \mathbf{T}^{(\beta)}(\Gamma_{\bar{s}}), \tag{3.10}$$

where $\bar{s}$ is the state at time t of history $\mathbf{h}_t$.

Note that the time consistency property boils down to the terminal fairness property when $t = 0$. In particular, if we choose $\mathbf{T} \equiv \mathbf{Co}$, then the time consistency properties entails that, if a coalition forms at time t, then the expected long run payoff that it receives from time t onwards is not larger than the one it would earn by cooperating, for any t. Formally, for all $t \geq 0$,

$$\sum_{i \in \Lambda} \mathbb{E}\left[\sum_{k=t}^{\infty} \beta^k g_i^{(\beta)}(S_k, \mathbf{h}_{k-1}) \Big| \mathbf{h}_t\right] \geq \beta^t v^{(\beta)}(\Lambda, \Gamma_{\bar{s}}), \quad \forall\, \Lambda \subset \mathcal{P},\ \mathbf{h}_t \in \mathcal{H}_t.$$

In other words, when $\mathbf{T} \equiv \mathbf{Co}$, the time consistency property clears up any coalition's dilemma "*Shall we stick to the grand coalition forever or withdraw now?*" in favor of the first alternative. We will extend further this concept in Section 3.1.7.

Next we extend the definition of time consistency by suggesting that, at any instant t, the expected payoff obtained by the players from time $t+n$ onwards should belong to the terminal solution associated to the stage of the game at time $t+n$.

Property 4 (n-tuple step time consistency)**.** *Set $\beta \in [0;1)$ and let $n \in \mathbb{N}_0$. A CPDP $g^{(\beta)}$ is n-tuple step time consistent w.r.t. a terminal cooperative solution $\mathbf{T}^{(\beta)}$ whenever, for all $t \geq 0$, $\mathbf{h}_t \in \mathcal{H}_t$,*

$$\mathbb{E}\left[\sum_{k=t+n}^{\infty} \beta^k g^{(\beta)}(S_k, \mathbf{h}_{k-1}) \Big| \mathbf{h}_t\right] \in \beta^{t+n} \mathbf{T}^{(\beta)}\left(\sum_{s' \in \mathcal{S}} p_n(s'|S_t = \bar{s}, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\Gamma_{s'}\right),$$

where p_n is the n-step transition probability and $\bar{s}$ is the state at time t of history $\mathbf{h}_t$.

The reader should notice that Property 4 reduces to Property 3 when $n = 0$. Now we are ready to show that the MDP-CPDP fulfills the n-tuple step time consistency property for any value of n. The proof follows from the stationarity of the allocation, the terminal fairness property, and two linearity properties of the Core and of the Shapley value, respectively.

Theorem 3.1.6. *Let $\mathbf{T}$ represent the Shapley Value, or the Core if we suppose that $\mathbf{Co}^{(\beta)}(\Gamma_s)$ is nonempty for any $s \in \mathcal{S}$. The stationary MDP-CPDP $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ is time consistent w.r.t. $\mathbf{T}^{(\beta)}$ for all $\beta \in [0;1)$. Moreover, it satisfies the n-tuple step time consistency property for all $n \in \mathbb{N}_0$ and $\beta \in [0;1)$.*

Proof. Since $\gamma^{(\beta)}$ is stationary, we can rewrite (3.10) as

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\beta^k\gamma^{(\beta)}(S_{t+k})\Big|S_t=\bar{s}\right]\in\mathbf{T}^{(\beta)}(\Gamma_{\bar{s}}). \tag{3.11}$$

Thanks to Theorem 3.1.4, (3.11) holds, hence $\gamma^{(\beta)}$ is time consistent. It is easy to verify that

$$\mathbb{E}\left[\sum_{k=t+n}^{\infty}\beta^k g^{(\beta)}(S_k,\mathbf{h}_{k-1})\Big|\mathbf{h}_t\right]=\beta^{t+n}\sum_{s'\in\mathcal{S}}p_n(s'|S_t=\bar{s},\mathbf{f}_{\mathcal{P}}^{(\beta)*})\mathbf{T}^{(\beta)}(\Gamma_{s'}).$$

Therefore, from Proposition 3.1.2 we claim that, if $\mathbf{T}\equiv\mathbf{Co}$, then

$$\mathbb{E}\left[\sum_{k=t+n}^{\infty}\beta^k g^{(\beta)}(S_k,\mathbf{h}_{k-1})\Big|\mathbf{h}_t\right]\in\beta^{t+n}\mathbf{Co}^{(\beta)}\left(\sum_{s'\in\mathcal{S}}p_n(s'|S_t=\bar{s},\mathbf{f}_{\mathcal{P}}^{(\beta)*})\Gamma_{s'}\right).$$

Moreover, for Corollary 3.1.3 we claim that, if $\mathbf{T}\equiv\mathcal{S}h$, then

$$\mathbb{E}\left[\sum_{k=t+n}^{\infty}\beta^k g^{(\beta)}(S_k,\mathbf{h}_{k-1})\Big|\mathbf{h}_t\right]=\beta^{t+n}\mathcal{S}h^{(\beta)}\left(\sum_{s'\in\mathcal{S}}p_n(s'|S_t=\bar{s},\mathbf{f}_{\mathcal{P}}^{(\beta)*})\Gamma_{s'}\right).$$

Thus (3.11) is verified for $\mathbf{T}\equiv\mathbf{Co}$ and $\mathbf{T}\equiv\mathcal{S}h$, and the thesis is proven. □

3.1.6 Greedy Players Satisfaction

In this section we allow for the presence of greedy players, i.e. players having a myopic perspective of the game and who only look to receive the highest reward in the static game played in the current state. From an allocation procedure design point of view, the most conservative approach is to expect that all the players *might* manifest a greedy behavior, and to construct a CPDP that contents all of them. The most natural way to formalize this property is requiring that the payoff allocation in each state belongs to the Core of its respective static game.

Property 5 (Greedy players satisfaction)**.** *Set $\beta\in[0;1)$. For all $s\in\mathcal{S}$, the CPDP $g^{(\beta)}(s)$ belongs to Core of the stage-wise game Ω_s, i.e. $g^{(\beta)}(s)\in$ $\mathbf{Co}(\Omega_s)$.*

By demanding that MDP-CPDP should fulfill Property 5, we seek to accommodate two apparently contrasting needs. On the one hand, we are trying to allocate a payoff which is globally optimum and in some sense "fair" in the long run game. On the other hand, we need to satisfy potential greedy players, hence the allocation needs to be globally optimum and stable in each static game Ω. The theory of MDPs claims that, in general, our goal cannot be reached for any value of $\beta\in[0;1)$, since the myopic strategy for the grand coalition $\mathcal{P}$ is not in general global optimum when β is sufficiently close to 1. Nevertheless, by letting the discount factor β be sufficiently close to 0, we will show a sufficient condition under which Property 5 holds. For this purpose, in the current section we consider the Shapley value as terminal fair solution, i.e. $\mathbf{T} \equiv \mathcal{S}h$.

Let us *assume* in the current section that the static game in state s, Ω_s, is a cooperative TU game, for all $s \in \mathcal{S}$. Moreover, in this section we suppose that the coalition values $v^{(\beta)}(\Lambda, \Gamma_s), v^{(\beta)}(\Lambda, \Omega_s)$ are the β-discounted values of the two player zero-sum game of coalition Λ against $\mathcal{P}\backslash\Lambda$ in the games Γ_s and Ω_s respectively. This classic formulation was originally devised by von Neumann and Morgenstern (1944, [96]). Of course, $v^{(0)}(\Lambda, \Gamma_s) = v(\Lambda, \Omega_s)$.

Condition 1 (max-min coalition values)**.** *The coalition value $v^{(\beta)}(\Lambda, \Gamma_s)$ is computed as the max-min expression in (3.3), for all $\Lambda \subseteq \mathcal{P},\ s \in \mathcal{S}$. The analogous expression holds for $v(\Lambda, \Omega_s)$.*

Lemma 3.1.7. *There exists a pure strategy $\underline{\mathbf{f}}^*_{\mathcal{P}} \in \mathbf{F}_{\mathcal{P}}$ and $\beta^* > 0$ such that $\underline{\mathbf{f}}^*_{\mathcal{P}}$ is optimal for all $\beta \in [0; \beta^*)$.*

Proof. The global optimization problem is a Markov Decision Process (MDP) having $\Phi^{(\beta)}_{\mathcal{P}}$ as discounted reward. Take a strictly decreasing sequence $\{\beta_k\}$ such that $\lim_{k\to\infty} \beta_k = 0$. Since both the actions and the states have a finite cardinality, then there exists a pure strategy $\underline{\mathbf{f}}^*_{\mathcal{P}}$ and an infinite subsequence of $\{\beta_k\}$, namely $\{\beta_{n_k}\}$, with $n_k < n_{k+1}\ \forall\, k$, such that $\underline{\mathbf{f}}^*_{\mathcal{P}}$ is optimal for all the discount factors $\{\beta_{n_k}\}$. Fix a pure strategy $\mathbf{f}_{\mathcal{P}} \in \mathbf{F}_{\mathcal{P}}$. Then

$$y^{(\beta_{n_k})}(s, \mathbf{f}_{\mathcal{P}}) := \Phi^{(\beta_{n_k})}_{\mathcal{P}}(s, \underline{\mathbf{f}}^*_{\mathcal{P}}) - \Phi^{(\beta_{n_k})}_{\mathcal{P}}(s, \mathbf{f}_{\mathcal{P}}) \geq\ 0, \qquad \forall\, k \in \mathbb{N}. \tag{3.12}$$

It is easy to see that $y^{(\beta)}$ is a continuous rational function in $\beta \in (0; 1)$. Then, either it is identically zero for all $\beta \in (0; 1)$ or $y^{(\beta)} = 0$ in a finite number of points in the interval $(0; 1)$. Hence, for (3.12), there exists $\beta^*(s, \mathbf{f}_{\mathcal{P}}) > 0$ such that $y^{(\beta)}(s, \mathbf{f}_{\mathcal{P}}) \geq 0$, for all $\beta \in (0; \beta^*(s, \mathbf{f}_{\mathcal{P}}))$. Take $\beta^* = \min_{s, \mathbf{f}_{\mathcal{P}}} \beta^*(s, \mathbf{f}_{\mathcal{P}}) > 0$.

Since $\Phi^{(\beta)}_{\mathcal{P}}(s, \underline{\mathbf{f}}^*_{\mathcal{P}})$ is also right-continuous in β at $\beta = 0$, then $\underline{\mathbf{f}}^*_{\mathcal{P}}$ is also optimal for $\beta = 0$. Hence the thesis is proven. □

Let us define Θ_s as the affine space:

$$\Theta_s :\ \left\{ \mathbf{x} \in \mathbb{R}^P :\ \sum_{i\in\mathcal{P}} x_i = \sum_{i\in\mathcal{P}} r_i(s, \underline{\mathbf{f}}^*_{\mathcal{P}}) \right\}, \tag{3.13}$$

where $\underline{\mathbf{f}}^*_{\mathcal{P}}$ is the global optimal strategy for all discount factors sufficiently close to 0, i.e.

$$\exists\, \beta^* > 0\ :\ \underline{\mathbf{f}}^*_{\mathcal{P}} = \operatorname*{argmax}_{\mathbf{f}_{\mathcal{P}}\in\mathbf{F}_{\mathcal{P}}} \mathbf{\Phi}^{(\beta)}_{\mathcal{P}}(\mathbf{f}_{\mathcal{P}}) \quad \forall\, \beta \in [0; \beta^*). \tag{3.14}$$

Corollary 3.1.8. *For any $s \in \mathcal{S}$, $\gamma^{(\beta)}(s)$ belongs to the affine space Θ_s, for all β sufficiently close to 0.*

Proof. The proof follows straightforward from Theorem 3.1.5 and from Lemma 3.1.7. □

Next we present a useful result.

Lemma 3.1.9. *Let $\mathbf{T} \equiv \mathcal{S}h$. Under Condition 1, $\lim_{\beta\downarrow 0}\gamma^{(\beta)}(s) = \mathcal{S}h^{(0)}(\Gamma_s) \equiv \mathcal{S}h(\Omega_s)$.*

Proof. Let us rewrite (3.9) as

$$\gamma^{(\beta)}(s, \mathcal{S}h) = \sum_{s'\in\mathcal{S}} \left[\delta_{s,s'} - \beta\, p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\right] \mathcal{S}h^{(\beta)}(\Gamma_{s'}),\ \forall\, s \in \mathcal{S}.$$

It is sufficient to prove that $\lim_{\beta\downarrow 0}\mathcal{S}h^{(\beta)}(\Gamma_s) = \mathcal{S}h^{(0)}(\Gamma_s)$, $\forall\, s \in \mathcal{S}$. Since each component of the vector $\mathcal{S}h^{(\beta)}(\Gamma_s)$ is a linear combination of the discounted values $\{v^{(\beta)}(\Lambda, \Gamma_s)\}_{\Lambda\subseteq\mathcal{P}}$, then we only need to show that

$$\lim_{\beta\downarrow 0} v^{(\beta)}(\Lambda, \Gamma_s) = v^{(0)}(\Lambda, \Gamma_s) = v(\Lambda, \Omega_s),\ \forall\, s \in \mathcal{S},\ \Lambda \subseteq \mathcal{P}.$$

Firstly, let us recall the relation (Filar and Vrieze, 1997 [32])

$$|\operatorname{val}(\mathbf{B}) - \operatorname{val}(\mathbf{C})| \le \max_{i,j} |\mathbf{B}_{i,j} - \mathbf{C}_{i,j}| \tag{3.15}$$

where $\mathbf{B}, \mathbf{C}$ are matrices with the same size. We know from Filar and Vrieze (1997, [32]) that

$$\begin{aligned} v^{(\beta)}(\Lambda, \Gamma_s) = \ \operatorname{val}\Bigg(\Big[\sum_{i\in\Lambda} r_i(s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda}) + \dots \\ + \beta \sum_{s'\in\mathcal{S}} p(s'|s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda})\, v^{(\beta)}(\Lambda, \Gamma_{s'})\Big]_{a_\Lambda=1, a_{\mathcal{P}\setminus\Lambda}=1}^{m_\Lambda(s), m_{\mathcal{P}\setminus\Lambda}(s)}\Bigg), \end{aligned} \tag{3.16}$$

where $a_\Lambda \in A_\Lambda(s)$ and $a_{\mathcal{P}\setminus\Lambda} \in A_{\mathcal{P}\setminus\Lambda}(s)$. Thus, from (3.15,3.16) we can say that, for all $\Lambda \subseteq \mathcal{P}$,

$$\begin{aligned} |v^{(\beta)}(\Lambda, \Gamma_s) - v^{(0)}(\Lambda, \Gamma_s)| &\le \max_{a_\Lambda, a_{\mathcal{P}\setminus\Lambda}} \left|\beta \sum_{s'\in\mathcal{S}} p(s'|s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda})\, v^{(\beta)}(\Lambda, \Gamma_{s'})\right| \\ &\le \frac{\beta}{1-\beta} M \end{aligned}$$

where $M = \max_{s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda}} |r_\Lambda(s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda})|$. Fix $\epsilon > 0$. Set $\delta = \epsilon/(M+\epsilon)$. Then, for all $\beta \in [0;\delta)$ we have $|v^{(\beta)}(\Lambda, \Gamma_s) - v^{(0)}(\Lambda, \Gamma_s)| < \epsilon$. Hence, $v^{(\beta)}(\Lambda, \Gamma_s)$ is right-continuous in β at $\beta = 0$ for all $s \in \mathcal{S}$, $\Lambda \subseteq \mathcal{P}$. □

Let us formulate an additional condition, on the strict convexity of static games, which holds only in the current section.

Condition 2 (Stage-wise strict convexity)**.** *The static games* $\{\Omega_s\}_{s\in\mathcal{S}}$ *are strictly convex, i.e.* $v(\Lambda_1 \cup \Lambda_2, \Omega_s) + v(\Lambda_1 \cap \Lambda_2, \Omega_s) > v(\Lambda_1, \Omega_s) + v(\Lambda_2, \Omega_s)$*, for all* $\Lambda_1, \Lambda_2 \subseteq \mathcal{P}$*,* $s \in \mathcal{S}$*.*

We know from Shapley (1971, [84]) that, if Condition 2 holds, then the Core of Ω_s is $(P-1)$-dimensional for any $s \in \mathcal{S}$, i.e. the affine hull of $\mathbf{Co}(\Omega_s)$ coincides with Θ_s in (3.13). Note that, in general, the affine hull of $\mathbf{Co}(\Omega_s)$ could be a proper subset of Θ_s.

Corollary 3.1.10. *Suppose that the stage-wise strict convexity Condition 2 holds. Then, for all* $s \in \mathcal{S}$*,*

i) the Shapley value of Ω_s *lies in the relative interior of* $\mathbf{Co}(\Omega_s)$*;*

ii) the interior of $\mathbf{Co}(\Omega_s)$ *relative to* Θ_s *coincides with the strict Core* $\mathbf{sCo}(\Omega_s)$*.*

Proof. For the proof of *i)*, see Shapley (1971, [84]). The proof of *ii)* is straightforward. □

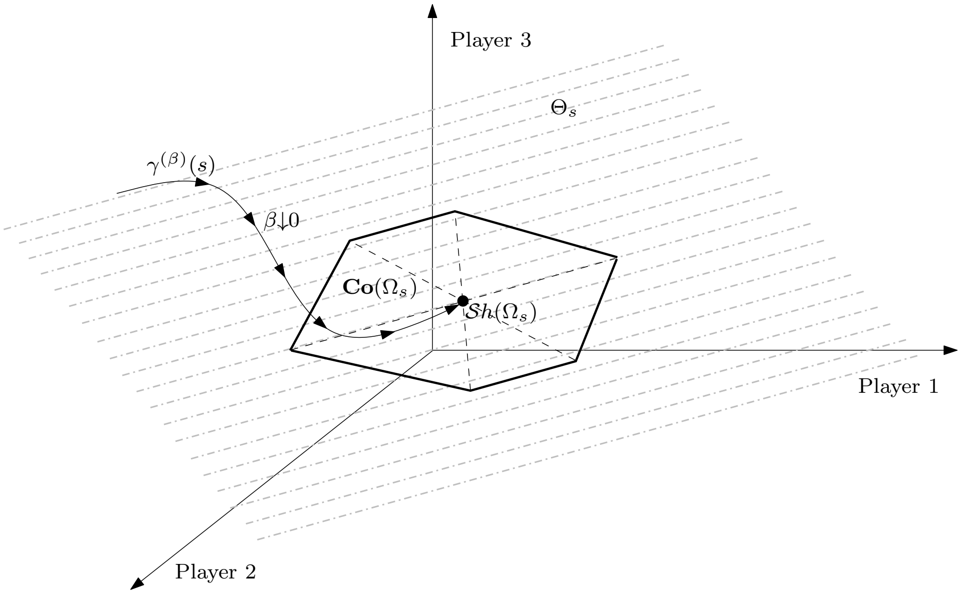

Figure 3.1: $\gamma^{(\beta)}(s)$ when $\beta \downarrow 0$, for a 3-player stochastic game in which Ω_s is strictly convex. Note that, in order to ensure Property 5, the affine hull of $\mathbf{Co}(\Omega_s)$ must coincide with Θ_s.

Finally, we are ready to show under which conditions the MDP-CPDP fulfills the greedy players satisfaction property.

Theorem 3.1.11. *Under Conditions 1 and 2, the greedy players satisfaction property is verified by $\gamma^{(\beta)}(\mathcal{S}h^{(\beta)})$ for all discount factors β sufficiently close to* 0.

Proof. Fix $s \in \mathcal{S}$. We know from Corollary 3.1.10 that $\mathcal{S}h(\Omega_s)$ lies in the relative interior of $\mathbf{Co}(\Omega_s)$. The affine hull of $\mathbf{Co}(\Omega_s)$ coincides with the hyperplane Θ_s for Condition 2. Moreover, from Corollary 3.1.8 we know that, for all $s \in \mathcal{S}$, $\gamma^{(\beta)}(s)$ belongs to the affine space Θ_s for all $\beta \in [0, \beta^*)$, where β^* is defined as in (3.14). Hence, for Lemma 3.1.9 we can say that for all $\epsilon > 0$ there exists $\delta_s \in (0, \beta^*)$ such that

$$\forall\, \beta \in [0; \delta_s),\ \gamma^{(\beta)}(s) \in [B_{\delta_s} \cap \Theta_s] \subseteq \mathbf{Co}(\Omega_s),$$

where B_{δ_s} is the ball belonging to $\mathbb{R}^P$ having radius of δ_s. Take $\delta = \min_{s \in \mathcal{S}} \delta_s$. The thesis is proven. □

Hence, under Condition 2, for all $\beta \in [0; \delta)$, all the greedy players are content with payoff allocation procedure, since the MDP-CPDP belongs to the Core of each static game Ω_s, for all $s \in \mathcal{S}$.

3.1.7 Cooperation Maintenance

The (single step) cooperation maintenance property was first introduced by Mazalov and Rettieva (2010, [59]), who employed it in a deterministic fish war setting. Such a property is very desirable, since it helps to preserve the cooperation agreement throughout the game. Indeed it suggests that the long run payoff that each coalition expects to earn by deviating in the next stage of the game should be not smaller than the payoff that the coalition receives by deviating in the current stage. In this section we will adapt and apply this property to our cooperative MDP model. For simplicity, we restrict the following definitions to stationary CPDP's.

Property 6 (Single step cooperation maintenance)**.** *Set $\beta \in [0; 1)$. The stationary CPDP $g^{(\beta)}$ satisfies, in any state $s \in \mathcal{S}$ and for each coalition $\Lambda \subset \mathcal{P}$,*

$$\sum_{i \in \Lambda} g_i^{(\beta)}(s) + \beta v^{(\beta)} \left(\Lambda, \sum_{s' \in \mathcal{S}} p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\, \Gamma_{s'} \right) \geq v^{(\beta)}(\Lambda, \Gamma_s). \tag{3.17}$$

In other words, Property 6 claims that each coalition has always an incentive to postpone the moment in which it will withdraw from the grand coalition, under the condition that, once a coalition $\Lambda \subset \mathcal{P}$ is formed, it can no longer rejoin the grand coalition in the future. By induction, we can say that the cooperation maintenance property enforces the grand coalition agreement throughout the game.

We point out that the transition probabilities in (3.17) are invariant with respect to a change of strategy by Λ, which can only withdraw at the following time step.

n-tuple step cooperation maintenance

Intuitively, Property 6 sorts out a coalition's dilemma "*Shall we withdraw from the grand coalition in one time step or now?*" in favor of the first option, at *any* stage of the game. It is natural to extend this property to a setting in which a coalition investigates the benefit of withdrawing in a later stage of the game. In other words, if a coalition faces the dilemma "*Shall we withdraw from the grand coalition in n time steps or now?*", we suggest that a CPDP should always persuade the coalition to defer the decision of defecting, for any integer n.

Property 7 (n-tuple step cooperation maintenance)**.** *Set $\beta \in [0;1)$. Let $n \in \mathbb{N}_0$. The stationary CPDP $g^{(\beta)}$ satisfies the n-tuple step cooperation maintenance property whenever, for any initial state $s \in \mathcal{S}$ and for each coalition $\Lambda \subset \mathcal{P}$,*

$$\sum_{t=0}^{n-1} \beta^t p_t(s'|s,\mathbf{f}_{\mathcal{P}}^{(\beta)*}) \sum_{i\in\Lambda} g_i^{(\beta)}(s') + \ldots$$
$$\beta^n v^{(\beta)}\left(\Lambda, \sum_{s'\in\mathcal{S}} p_n(s'|s,\mathbf{f}_{\mathcal{P}}^{(\beta)*})\Gamma_{s'}\right) \geq v^{(\beta)}(\Lambda, \Gamma_s).$$

Next we show a necessary and sufficient condition for the existence of an MDP-CPDP $\gamma^{(\beta)}$ satisfying the n-tuple step cooperation maintenance property, for $n \geq 1$. Before this, a notation remark. We denote $\mathbf{v}^{(\beta)}(\Lambda,\Gamma)$ as

$$\mathbf{v}^{(\beta)}(\Lambda,\Gamma) := \left[v^{(\beta)}(\Lambda,\Gamma_{s_1}),\ \ldots,\ v^{(\beta)}(\Lambda,\Gamma_{s_N})\right]^T, \qquad \forall\,\Lambda \subseteq \mathcal{P}.$$

Theorem 3.1.12. *Let $n \in \mathbb{N}_0$, $\beta \in [0;1)$. The set of MDP-CPDP's satisfying the n-tuple step cooperation maintenance property is nonempty if and only if the vectors*

$$\left[\mathbf{I} - \left[\beta\mathbf{P}^{*(\beta)}\right]^n\right] \mathbf{v}^{(\beta)}(\Lambda,\Gamma) :=\ \widetilde{\mathbf{v}}^{(\beta,n)}(\Lambda,\Gamma), \qquad \Lambda \subseteq \mathcal{P} \tag{3.18}$$

are component-wisely balanced, i.e. for every function $\alpha : 2^P/\{\emptyset\} \to [0;1]$ such that:

$$\forall\, i \in \mathcal{P} : \sum_{\substack{\Lambda\subseteq\mathcal{P}:\\ \Lambda\ni i}} \alpha(\Lambda) = 1,$$

the following condition holds:

$$\sum_{\Lambda\subseteq\mathcal{P}} \alpha(\Lambda)\widetilde{\mathbf{v}}_k^{(\beta,n)}(\Lambda,\Gamma) \leq \widetilde{\mathbf{v}}_k^{(\beta,n)}(\mathcal{P},\Gamma), \qquad 1 \leq k \leq N,$$

where $\widetilde{\mathbf{v}}_k^{(\beta,n)}(\Lambda,\Gamma)$ is the k-th component of $\widetilde{\mathbf{v}}^{(\beta,n)}(\Lambda,\Gamma)$.

Proof. Let us rewrite (3.9) as:

$$\boldsymbol{\gamma}_i^{(\beta)}(\overline{\mathbf{T}}) = \left[\mathbf{I} - \beta \mathbf{P}^{*(\beta)}\right] \overline{\mathbf{T}}_i^{(\beta)}, \quad \forall\, i \in \mathcal{P} \tag{3.19}$$

where $\boldsymbol{\gamma}_i^{(\beta)}(.) = [\gamma_i^{(\beta)}(s_1,.),\ldots,\gamma_i^{(\beta)}(s_N,.)]^T$ and $\overline{\mathbf{T}}_i^{(\beta)} = [\overline{\mathbf{T}}_i^{(\beta)}(\Gamma_{s_1}),\ldots,\overline{\mathbf{T}}_i^{(\beta)}(\Gamma_{s_N})]^T$. Thanks to Proposition 3.1.1, by applying twice the well known formula for matrix geometric series:

$$\sum_{k=0}^{n-1} \left[\beta \mathbf{P}^{*(\beta)}\right]^k = \left[\mathbf{I} - \beta \mathbf{P}^{*(\beta)}\right]^{-1} \left[\mathbf{I} - \left[\beta \mathbf{P}^{*(\beta)}\right]^n\right],$$

we can reformulate Property 7 as

$$\begin{cases} \left[\mathbf{I} - \left[\beta \mathbf{P}^{*(\beta)}\right]^n\right] \sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} \geq \left[\mathbf{I} - \left[\beta \mathbf{P}^{*(\beta)}\right]^n\right] \mathbf{v}^{(\beta)}(\Lambda,\Gamma), \quad \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases} \tag{3.20}$$

Since the matrix $\mathbf{I} - [\beta \mathbf{P}^{*(\beta)}]^n$ is invertible for any $n \in \mathbb{N}$, then we can equivalently rewrite (3.20) as

$$\begin{cases} \sum_{i\in\Lambda} \widetilde{\overline{\mathbf{T}}}_i^{(\beta,n)} \geq \widetilde{\mathbf{v}}^{(\beta,n)}(\Lambda,\Gamma), \ \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \widetilde{\overline{\mathbf{T}}}_i^{(\beta,n)} = \widetilde{\mathbf{v}}^{(\beta,n)}(\mathcal{P},\Gamma) \end{cases} \tag{3.21}$$

where

$$\widetilde{\overline{\mathbf{T}}}_i^{(\beta,n)} = \left[\mathbf{I} - \left[\beta \mathbf{P}^{*(\beta)}\right]^n\right] \overline{\mathbf{T}}_i^{(\beta)}.$$

Since the relations in the systems of inequalities in (3.21) are componentwise, for the Bondareva-Shapley Theorem (Bondareva, 1963 [22]; Shapley 1967, [83]) the thesis is proven. □

The reader should notice that, in the limit for $n \to \infty$, the result of Theorem 3.1.12 coincides (component-wisely) with the Bondareva-Shapley Theorem for static cooperative games.

Next we show an intuitive result which reinforces the importance of the single step cooperation maintenance property. If an MDP-CPDP satisfies the n-tuple step property for $n = 1$, then it also fulfills it for all integers n. In this case, for any coalition, the worst decision between defecting at the current stage and at *any* future stage happens to be the former one.

Theorem 3.1.13. *Let $\beta \in [0;1)$. If the MDP-CPDP $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ satisfies the single step cooperation maintenance property, then it satisfies the n-tuple step cooperation maintenance property, for all $n > 1$.*

Proof. Since $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ satisfies the single step cooperation maintenance property, then we can write

$$\begin{cases} \beta \mathbf{P}^{*(\beta)} \left[\sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda,\Gamma) \right] \geq \sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda,\Gamma), & \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases} \tag{3.22}$$

By iteratively left multiplying by the nonnegative matrix $\beta\mathbf{P}^{*(\beta)}$ both sides of the first expression in (3.22), then we obtain for each coalition $\Lambda \subset \mathcal{P}$:

$$\sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda,\Gamma) \leq \beta \mathbf{P}^{*(\beta)} \left[\sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda,\Gamma) \right] \leq$$
$$\left[\beta \mathbf{P}^{*(\beta)}\right]^2 \left[\sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda,\Gamma) \right] \leq \ldots$$

Hence, the thesis is proven. □

Cooperation Maintaining solution

In the following we prove that if an MDP-CPDP $\gamma^{(\beta)}$ fulfills the single step cooperation maintenance property, then the discounted sum of allocations for each player, when s is the initial state, belongs to the Core of the game Γ_s, i.e. $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{Co}^{(\beta)}(\Gamma_s)$, for all $s \in \mathcal{S}$.

Corollary 3.1.14. *Set $\beta \in [0;1)$. If an MDP-CPDP $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ satisfies the single step cooperation maintenance property, then*

$$\mathbb{E}\left[\sum_{t\geq 0} \beta^t \gamma^{(\beta)}(S_t) \middle| S_0 = s \right] \in \mathbf{Co}^{(\beta)}(\Gamma_s), \quad \forall\, s \in \mathcal{S}. \tag{3.23}$$

Proof. Since $\gamma^{(\beta)}$ satisfies Property 6, then (3.20) is verified, with $n = 1$. By left multiplying each set of inequalities in (3.20) by the nonnegative matrix $(\mathbf{I} - \beta\mathbf{P}^{*(\beta)})^{-1}$, we obtain the following expressions:

$$\begin{cases} \sum_{i\in\Lambda} \overline{\mathbf{T}}_i^{(\beta)} \leq \mathbf{v}^{(\beta)}(\Lambda,\Gamma), & \forall\, \Lambda \subset \mathcal{P}, \\ \sum_{i\in\mathcal{P}} \overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases} \tag{3.24}$$

Thanks to Theorem 3.1.4, we can say that the relations in (3.24) are equivalent to (3.23), hence the thesis is proven. □

Interestingly, Corollary 3.1.14 suggests that the cooperation maintenance property might be considered as a refinement of the concept of the Core of a long run game. In Section 3.1.7 we will show that it is actually a proper refinement. Therefore, it is worth coining a new terminal cooperative solution for cooperative MDPs, that we dub *Cooperation Maintaining solution*, grounded on the cooperation maintenance property.

Definition 8 (Cooperation Maintaining solution)**.** *Let $\beta \in [0;1)$. The Cooperation Maintaining solution $\mathbf{CM}^{(\beta)}(\Gamma)$ is the set of long run allocations $\{\mathbf{x}_i \in \mathbb{R}^N\}_{i=1,\dots,P}$ such that*

$$\begin{cases} \left[\mathbf{I} - \beta \mathbf{P}^{*(\beta)}\right] \sum_{i\in\Lambda} \mathbf{x}_i \geq \left[\mathbf{I} - \beta \mathbf{P}^{*(\beta)}\right] \mathbf{v}^{(\beta)}(\Lambda, \Gamma), \quad \forall \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \mathbf{x}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P}, \Gamma). \end{cases}$$

We point out that a classic terminal cooperative solution, such as Core, Shapley value etc., can be defined just for a specific a long run game Γ_s, for some $s \in \mathcal{S}$, by computing the coalition values $v^{(\beta)}(., \Gamma_s)$. Therefore, a classic terminal solution is a vector in $\mathbb{R}^P$. Instead, the Cooperation Maintaining solution involves the computation of all coalition values $v^{(\beta)}(., \Gamma_s)$, for all $s \in \mathcal{S}$, and a solution point is a collection of P vectors belonging to $\mathbb{R}^N$. Of course, a Cooperation Maintaining solution point can be expressed as a collection of N vectors in $\mathbb{R}^P$, and either of the two definitions can be used, at one's convenience.

By collecting the results of this section, we enumerate the properties of the Cooperation Maintaining solution in the following Corollaries.

Corollary 3.1.15. *The Cooperation Maintaining solution $\mathbf{CM}^{(\beta)}(\Gamma)$ is a nonempty set if and only if the modified coalition values $\{\widetilde{\mathbf{v}}_k^{(\beta,1)}(\Lambda, \Gamma)\}_{\Lambda\subseteq\mathcal{P}}$, defined as in (3.18), are balanced, for $k = 1, \dots, N$.*

Corollary 3.1.16. *Let us assume that $\mathbf{CM}^{(\beta)}(\Gamma)$ is nonempty. Then $\cup_{s\in\mathcal{S}} \overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{CM}^{(\beta)}(\Gamma)$ if and only if the MDP-CPDP $\gamma^{(\beta)}(., \overline{\mathbf{T}})$ satisfies the n-tuple step cooperation maintenance property, for all $n \in \mathbb{N}$.*

Corollary 3.1.17. *For all $\beta \in [0, 1)$, $\mathbf{CM}^{(\beta)}(\Gamma) \subseteq \cup_{s\in\mathcal{S}} \mathbf{Co}^{(\beta)}(\Gamma_s)$.*

The Cooperation Maintaining solution is a proper refinement of the Core

It is natural to ask whether the converse of Corollary 3.1.17 is true, i.e. whether trivially $\mathbf{CM}^{(\beta)}(\Gamma) \equiv \cup_{s\in\mathcal{S}} \mathbf{Co}^{(\beta)}(\Gamma_s)$ or the Cooperation maintaining concept is a proper refinement of the Core. In this section we will show that $\mathbf{CM}^{(\beta)}(\Gamma) \neq \cup_{s\in\mathcal{S}} \mathbf{Co}^{(\beta)}(\Gamma_s)$, by finding an allocation $\overline{\mathbf{T}}^{(\beta)}$ such that $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{Co}^{(\beta)}(\Gamma_s)$ for all $s \in \mathcal{S}$, but $\cup_{s\in\mathcal{S}} \overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \notin \mathbf{CM}^{(\beta)}(\Gamma_s)$. Hence, the Cooperation maintaining solution is a proper refinement of the Core solution concept for cooperative MDPs.

Let us devise the counterexample. We consider a cooperative MDP with two players ($P = 2$), four states ($N = 4$), and with perfect information, i.e. in each state at most one player has more than one action available. Player 1 controls states (s_1, s_2), and the remaining states (s_3, s_4) are controlled by player 2. Let the discount factor $\beta = 0.8$. The immediate rewards for each player and the transition probabilities for each state/action pair are shown in Table 3.1.7.

	(s,a)	r_1	r_2	$p(s_1\|s,a)$	$p(s_2\|s,a)$	$p(s_3\|s,a)$	$p(s_4\|s,a)$
pl. 1	(s_1,a_1)	1	3	0.1	0.4	0.1	0.4
	(s_1,a_2)	2	1	0.4	0.1	0.1	0.3
	(s_1,a_3)	1	0	0.4	0.2	0.4	0.1
	(s_2,a_4)	2	1	0.1	0	0.4	0.4
	(s_2,a_5)	3	1	0.2	0.2	0.2	0.5
	(s_2,a_6)	4	3	0.2	0	0.2	0.3
pl. 2	(s_3,a_7)	5	1	0.3	0.6	0.4	0.1
	(s_3,a_8)	1	3	0.3	0.4	0.2	0
	(s_3,a_9)	2	6	0.3	0.3	0.1	0
	(s_4,a_{10})	0	1	0.5	0	0.1	0.1
	(s_4,a_{11})	2	2	0.1	0.3	0.5	0.2
	(s_4,a_{12})	3	0	0.1	0.5	0.3	0.6

Table 3.1: Immediate rewards and transition probabilities for each player, state, and strategy.

In this case, the vector values of the coalitions $\{1\}$, $\{2\}$ and $\mathcal{P} = \{1,2\}$, rounded off to the second decimal, are

$$\mathbf{v}^{(0.8)}(\{1\}) \approx \begin{bmatrix} 8.73 \\ 10.03 \\ 7.34 \\ 7.16 \end{bmatrix}, \quad \mathbf{v}^{(0.8)}(\{2\}) \approx \begin{bmatrix} 9.57 \\ 8.65 \\ 10.93 \\ 11.23 \end{bmatrix}, \quad \mathbf{v}^{(0.8)}(\{1,2\}) \approx \begin{bmatrix} 33.08 \\ 30.78 \\ 33.77 \\ 30.83 \end{bmatrix}.$$

where for simplicity of notation we write $\mathbf{v}^{(\beta)}(.)$ instead of $\mathbf{v}^{(\beta)}(.,\Gamma)$. Since the coalition values are component-wisely superadditive by construction, then $\mathbf{Co}^{(0.8)}(\Gamma_s)$ for the two-player case always exists, for all $s \in \mathcal{S}$. Let us select:

$$\overline{\mathbf{T}}_1^{(0.8)} = \mathbf{v}^{(0.8)}(\{1\}) + \begin{bmatrix} 0.7 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left[\mathbf{v}^{(0.8)}(\{1,2\}) - [\mathbf{v}^{(0.8)}(\{1\}) + \mathbf{v}^{(0.8)}(\{2\})]\right]$$

$$\overline{\mathbf{T}}_2^{(0.8)} = \mathbf{v}^{(0.8)}(\{2\}) + \begin{bmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \left[\mathbf{v}^{(0.8)}(\{1,2\}) - [\mathbf{v}^{(0.8)}(\{1\}) + \mathbf{v}^{(0.8)}(\{2\})]\right].$$

Thus, we obtain

$$\overline{\mathbf{T}}_1^{(0.8)} \approx \begin{bmatrix} 19.07 & 14.87 & 10.44 & 19.60 \end{bmatrix}^T$$
$$\overline{\mathbf{T}}_2^{(0.8)} \approx \begin{bmatrix} 14.01 & 15.91 & 23.32 & 11.23 \end{bmatrix}^T.$$

We find that:

$$\begin{aligned}
\widetilde{\overline{\overline{\mathbf{T}}}}_1^{(0.8)}(s_2) &\approx 2.92 < \widetilde{\mathbf{v}}_2^{(0.8,1)}(\{1\}) \approx 3.65 \\
\widetilde{\overline{\overline{\mathbf{T}}}}_1^{(0.8)}(s_3) &\approx -0.75 < \widetilde{\mathbf{v}}_3^{(0.8,1)}(\{1\}) \approx 0.51 \\
\widetilde{\overline{\overline{\mathbf{T}}}}_2^{(0.8)}(s_1) &\approx 0.48 < \widetilde{\mathbf{v}}_1^{(0.8,1)}(\{2\}) \approx 1.61 \\
\widetilde{\overline{\overline{\mathbf{T}}}}_2^{(0.8)}(s_4) &\approx 0.90 < \widetilde{\mathbf{v}}_4^{(0.8,1)}(\{2\}) \approx 3.00.
\end{aligned}$$

Therefore, the converse of Corollary 3.1.17 is not true, $\mathbf{CM}^{(\beta)}(\Gamma) \neq \cup_{s\in\mathcal{S}}\mathbf{Co}^{(\beta)}(\Gamma_s)$, and the Cooperation Maintaining solution is a proper refinement of the Core. On the other hand, it is interesting to observe that in this example, by randomly generating vectors $\overline{\mathbf{T}}^{(0.8)}(\Gamma_s) \in \mathbf{Co}^{(0.8)}(\Gamma_s)$, in about the 99.45% of the cases $\overline{\mathbf{T}}^{(0.8)}(\Gamma_s) \in \mathbf{CM}^{(0.8)}(\Gamma_s)$ as well, for all $s \in \mathcal{S}$.

3.1.8 Strictly convex static games

In the same spirit of Section 3.1.6, we now show that the sole strict convexity Condition 2 on the static games ensures the existence of an MDP-CPDP satisfying the cooperation maintenance property for all discount factors β small enough. As in Section 3.1.6, we assume that Condition 1 holds, i.e. the coalition values are computed *'a la* von Neumann and Morgenstern.

Theorem 3.1.18. *Suppose that Conditions 1,2 hold. Then,* $\gamma^{(\beta)}(., \mathcal{S}h)$ *satisfies the single step cooperation maintenance property for all* β *close enough to 0.*

Proof. Thanks to the linearity property of coalition values (see Proposition 3.1.1) we can reformulate Property 6 as

$$\sum_{i\in\Lambda}\gamma_i^{(\beta)}(s,\mathcal{S}h) \geq \sum_{s'\in\mathcal{S}}\left[\delta_{s,s'} - \beta p(s'|s,\mathbf{f}_{\mathcal{P}}^{(\beta)*})\right] v^{(\beta)}(\Lambda,\Gamma_{s'}), \qquad \forall\,\Lambda\subset\mathcal{P},\ s\in\mathcal{S}.$$

From (3.9), considering $\mathbf{T} \equiv \mathcal{S}h$,

$$\sum_{i\in\Lambda}\gamma_i^{(\beta)}(s,\mathcal{S}h) = \sum_{s'\in\mathcal{S}}\left[\delta_{s,s'} - \beta p(s'|s,\mathbf{f}_{\mathcal{P}}^{(\beta)*})\right]\sum_{i\in\Lambda}\mathcal{S}h_i^{(\beta)}(\Gamma_{s'}).$$

By hypothesis, for all $s \in \mathcal{S}$ the Shapley value $\mathcal{S}h(\Omega_s) = \mathcal{S}h^{(0)}(\Gamma_s)$ belongs to the strict Core $\mathbf{sCo}(\Omega_s)$ for all β sufficiently close to 0. Hence, by right continuity of the Shapley value and of coalition values in $\beta = 0$ (see proof of Lemma 3.1.9), we conclude that, for all β sufficiently close to 0,

$$\sum_{s'\in\mathcal{S}}\left[\delta_{s,s'} - \beta p(s'|s,\underline{\mathbf{f}}_{\mathcal{P}}^{*})\right]\left[\sum_{i\in\Lambda}\mathcal{S}h_i^{(\beta)}(\Gamma_{s'}) - v^{(\beta)}(\Lambda,\Gamma_{s'})\right] \geq 0, \quad \forall\, s\in\mathcal{S},$$

where $\underline{\mathbf{f}}_{\mathcal{P}}^{*}$ is the optimal strategy for grand coalition for all β sufficiently small, as in (3.14). Hence, the thesis is proven. □

3.1.9 Transition probabilities not depending on the actions

In this final section we deal with a special case of our model, entailing that the Markov process among the states is endogenous, i.e. players' strategies do not influence the transition probabilities among the states. This is formalized as follows.

Condition 3. *The probabilities of transition among the states do not depend on the players' actions, i.e.* $p(s'|s, a_1, \dots, a_P) = p(s'|s)$*, for all* $a_i \in A_i(s)$ *and for each* $s, s' \in \mathcal{S}$*.*

As in Sections 3.1.6 and 3.1.8, we consider the static games $\{\Omega_s\}_s$ to possess transferable utilities $\{v(\Lambda, \Omega_s)\}_{s\in\mathcal{S},\Lambda\subseteq\mathcal{P}}$. Nevertheless, we no longer impose the max-min Condition 1 on the coalition values. This model is equivalent to the one of Predtetchinski (2007, [75]), except for the TU assumption.
Now we show that, under Condition 3, the allocation problem simplifies considerably. In fact, the balancedness of each static game is a sufficient condition to ensure the existence of an MDP-CPDP satisfying Properties 5, 6, and 7.

Theorem 3.1.19. *Suppose that the static games* $\{\Omega_s\}_{s\in\mathcal{S}}$ *are balanced. Then, for all* $\beta \in [0;1)$*, there exists an MDP-CPDP* $\gamma^{(\beta)}(., \overline{\mathbf{T}})$ *such that the following properties are jointly met:*

i) $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{Co}^{(\beta)}(\Gamma_s)$*, for all* $s \in \mathcal{S}$*;*

ii) $\gamma^{(\beta)}(., \overline{\mathbf{T}})$ *fulfills the greedy player satisfaction property;*

iii) $\cup_{s\in\mathcal{S}}\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbf{CM}^{(\beta)}(\Gamma)$*, i.e.* $\gamma^{(\beta)}(., \overline{\mathbf{T}})$ *fulfills the n-tuple step cooperation maintenance property, for* $n \in \mathbb{N}$*.*

Proof. From the hypothesis, there exists $\{\boldsymbol{\gamma}_i^{(\beta)} \in \mathbb{R}^N\}_{i=1,\dots,P}$ such that

$$\begin{cases} \sum_{i\in\Lambda} \boldsymbol{\gamma}_i^{(\beta)} \geq \ \mathbf{v}(\Lambda, \Omega) \quad \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \boldsymbol{\gamma}_i^{(\beta)} = \ \mathbf{v}(\mathcal{P}, \Omega). \end{cases} \tag{3.25}$$

From the linearity property of coalition values (see Proposition 3.1.1) we claim that

$$\mathbf{v}^{(\beta)}(\Lambda, \Gamma) = \ \left[\mathbf{I} - \beta\mathbf{P}\right]^{-1} \mathbf{v}(\Lambda, \Omega) \quad \forall\, \Lambda \subseteq \mathcal{P}, \tag{3.26}$$

where $\mathbf{v}(\Lambda, \Omega) := \ [v(\Lambda, \Omega_{s_1}), \dots, v(\Lambda, \Omega_{s_N})]^T$. Thus, by left multiplying the expressions in (3.25) by the nonnegative matrix $(\mathbf{I} - \beta\mathbf{P})^{-1}$ we obtain

$$\begin{cases} \sum_{i\in\Lambda} \overline{\mathbf{T}}_i \geq \ \mathbf{v}^{(\beta)}(\Lambda, \Gamma) \quad \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \overline{\mathbf{T}}_i = \ \mathbf{v}^{(\beta)}(\mathcal{P}, \Gamma) \end{cases}$$

Hence, *i)* and *ii)* are proven by the construction of $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$. By plugging (3.26) in (3.25), we can write

$$\begin{cases} \sum_{i\in\Lambda} \gamma_i^{(\beta)} \geq & \left[\mathbf{I}-\beta\mathbf{P}\right] \mathbf{v}^{(\beta)}(\Lambda,\Gamma) \quad \forall\, \Lambda \subset \mathcal{P} \\ \sum_{i\in\mathcal{P}} \gamma_i^{(\beta)} = & \left[\mathbf{I}-\beta\mathbf{P}\right] \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases}$$

which coincides with the definition of the single step cooperation maintenance property. For Theorem 3.1.13, *iii)* is proven. Thus the thesis follows. □

Not surprisingly, Condition 3 simplifies considerably the allocation procedure issue at hand. Indeed, it is sufficient to prove the balancedness of the static games to ensure both the cooperation maintenance property and the greedy players satisfaction property. We recall that, in the general case in which the transition probabilities do depend on the players' actions, the hypothesis of stage-wise balancedness does not even imply property *ii)* of Theorem 3.1.19 for β sufficiently high, as pointed out in Section 3.1.6. Moreover, Theorem 3.1.19 suggests that, under Condition 3 and if the static games are balanced, it is convenient to devise a stage-wise allocation in a bottom-up fashion, i.e. by first allocating $\gamma^{(\beta)}(s) \in \mathbf{Co}(\Omega_s)$ in each state s, and then computing the terminal solution $\overline{\mathbf{T}}^{(\beta)}$, which turns out to belong to $\mathbf{Co}^{(\beta)}(\Gamma_s)$, in all states.

We also remark that the converse of property *i)* of Theorem 3.1.19 is not true. Indeed, it is possible to find a terminal cooperative solution $\overline{\mathbf{T}}^{(\beta)}$ belonging to the Core of the long run games Γ_s, for all $s \in \mathcal{S}$, whose associated MDP-CPDP $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ lies outside the Core of at least one static games Ω_s.

We conclude by providing a result for the Shapley value allocation procedure. The proof follows straightforward from Corollary 3.1.3 and equation (3.26).

Corollary 3.1.20. *Let $\beta \in [0;1)$. Let $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) \in \mathbb{R}^P$ be a terminal cooperative solution, for all $s \in \mathcal{S}$. Under Condition 3, $\gamma^{(\beta)}(s,\overline{\mathbf{T}}) = \mathcal{S}h(\Omega_s)$, for all $s \in \mathcal{S}$, if and only if $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) = \mathcal{S}h^{(\beta)}(\Gamma_s)$, for all $s \in \mathcal{S}$.*

3.1.10 Conclusions

This paper deals with Cooperative Markov Decision Processes (MDPs), in which sub-coalition of players may form throughout the game. Thus it is crucial to enforce at the beginning of the game an agreement that no player has interest to breach at any time step. Hence we proposed a payoff allocation procedure, called MDP-CPDP, distributing a cooperative solution, associated with the long run game, in each state of the MDP. Such an

MDP-CPDP is the only stationary allocation fulfilling a terminal fairness property, it is stage-wise efficient, and it is time consistent, i.e. the agreement stipulated at the beginning of the game holds throughout the game. We found sufficient conditions under which the MDP-CPDP also contents greedy players, having a myopic perspective of the game, for all discount factors sufficiently small. We studied a cooperation maintenance property, which is crucial since it enforces the cohesiveness of the grand coalition throughout the game. This property allowed us to define a new cooperative solution, dubbed Cooperation Maintaining solution, which is a refinement of the concept of Core for MDPs. We finally considered a simpler model with an endogenous Markov chain, in which the MDP-CPDP satisfies all the cited properties under more relaxed constraints.

3.2 Dynamic Rate Allocation in Markovian Quasi-Static Multiple Access Channels

We deal with multiple access channels in which the channel coefficients follow a quasi-static Markov process on a finite set of states. In the corresponding MDP model, the transition probabilities among the (channel) states do not depend on the players (users). Hence, under this perspective, the model is simpler than the one in Section 3.1. On the other hand, the rewards (rates) cannot be distributed in any manner in each state, but only within a feasibility (capacity) region. In other words, we drop the TU assumption of Section 3.1 and we deal with Non Transferable Utility (NTU) games. We first show how to allocate the rates in a global optimal fashion, in each state of the process. We give a sufficient condition under which the optimal rates adhere to some fairness criteria, holding in a time consistent fashion. Then, we borrow two concepts from dynamic cooperative game theory, i.e. the Time Consistent Core and the Cooperation Maintaining sets, which measure users' satisfaction with the assigned rate. We show that the sets of rates fulfilling these properties coincide, and they also coincide with the set of global optimal rate allocations.

3.2.1 Introduction

The concepts of user fairness and satisfaction have received significant attention in previous years. These notions will play an increasingly crucial role in future networks, due to the paradigm shift that we are witnessing, from fully centralized with dumb terminals to distributed networks with rational users able to pool resources with each other.

In the literature, the notion of fair and satisfactory rate allocation has been dealt with under manifold perspectives in static Gaussian or ergodically fading Multiple Access Channels (MAC). In (Shum and Sung, 2006 [86]), the fairness of a rate allocation in a Gaussian MAC is related to the economical concept of Lorenz order, used for measuring disparity in income distributions. Such fair allocation always exists, it is Pareto optimal, and also solution of a Nash bargaining problem with zero disagreement pay-off allocation. In the following (Shum and Sung, 2006 [87]), the authors show the existence of a unique rate allocation which is max-min and proportional fair. The results in [86, 87] are extended to the general framework of α-fairness (Mo and Walrand, 2000 [62]) in (Altman et al., 2009 [5]). For MAC's with polymatroid regions, all α-fair rate allocations collapse into a single point, which is max-min and proportional fair, too. An analysis of rate allocations in the context of constrained games points out that the normal Nash equilibrium (Rosen, 1965 [79]) also coincides with the α-fair and Pareto optimal allocations.

On the other hand, the issue of users satisfaction is addressed by Cooperative Game Theory (CGT) with non-transferable utility (NTU) (see Peleg and Sudhölter, 2007 [69] for an overview), which provides powerful tools to derive efficient and stable allocations in a setting in which the users can cooperate to reach a common goal. In (Madiman, 2008 [54]), a Gaussian MAC is studied with an approach not strictly game-theoretical, but some tools borrowed from CGT are utilized to characterize the capacity region. In (La and Anantharam, 2004 [50]), the authors expressed the rate allocation problem in static Gaussian MAC with jamming in a cooperative game-theoretical setting.They found a satisfactory rate allocation fulfilling the newly introduced concept of envy-free. The envy-free allocation exists, is unique and Pareto optimal, but in general it does not coincide with the α-fair solution. In this contribution we study and extend for the first time the concepts of optimal, fair, and satisfactory rate allocations to a *dynamic* scenario, described by a Gaussian MAC where the channel evolves quasi-statically, according to a Homogeneous Markov Chain (HMC) on a finite state space. The structure of the paper ramifies into two main sections. The former is Section 3.2.3, in which we discuss the design of optimal and fair allocations in the dynamic process. The latter is Section 3.2.4, in which we find the rate allocations which are satisfactory throughout the process, according to a Dynamic Cooperative Game Theory (DCGT) formulation. We study a bottom-up (Section 3.2.3) and a top-down procedure (Section 3.2.3) to allocate a global optimal rate in each state of the HMC. The former prescribes to allocate first the static allocations and derive next the long-run ones; conversely, the latter suggests to select first the long-run rate allocations. Though the top-down procedure would be more useful since the user have a long-run perspective, it is not always feasible, and we offer a remedy to this. In Section 3.2.4 we provide a sufficient condition under which it there exists a rate allocation which is fair, i.e. max-min, proportional, and α-fair, both state-wisely and in the long-run process. Most importantly, the fairness property of such allocation is Time Consistent, i.e. it is fair throughout the process, from each of its intermediate steps onwards. Conversely, a fair allocation always exists in the static case (Altman et al., 2009 [5]), (Maddah-Ali, 2009 [53]). In Section 3.2.4 we introduce a game formulation with jamming users similar to the one in (La and Anantharam, 2004 [50]), but in a dynamic scenario. We then characterize the set of global optimal allocations as satisfactory too, since it coincides with the set of rates for which two crucial DCGT properties hold. These properties are the (Time Consistent) Core, introduced in (Petrosjan, 1977 [72]), and the Cooperation Maintenance property (Mazalov and Rettieva, 2010 [59]). Such concepts formulate in two different, but equally appealing, manners the concept for which all users should find the allocation, in a sense, acceptable throughout the game.

3.2.2 System Model

We consider a wireless system in which K terminals attempt to send information to a single receiver or base station. Let $\mathcal{P} = \{1, \dots, K\}$ be the set of all users. Each user k has a power constraint P_k. We assume a quasi-static channel, i.e. the channel coefficients can be considered constant for the whole duration of a codeword. Thus, the t-th signal block received by the unique receiver, for $t \in \mathbb{N}_0$, can be written as

$$\mathbf{y}[t] = \sum_{k=1}^{K} h^{(k)}[t]\, \mathbf{x}^{(k)}[t] + \mathbf{w}[t]$$

where $\mathbf{x}^{(k)}[t]$ is the codeword of user k, $h^{(k)}[t]$ is the complex channel coefficient for user k at time step t, and $\mathbf{w}[t]$ is zero mean white Gaussian noise with variance N_0. We assume that the set of channel coefficients $\{h^{(1)}, h^{(2)}, \dots, h^{(K)}\}$ is finite and it follows a discrete time Homogeneous Markov chain (HMC), which can change state at every new codeword. In other words, if S_t is the channel state at time step t, where

$$S_t := \Big[h^{(1)}[t], \dots, h^{(K)}[t]\Big],$$

then the random process $\{S_t,\ t \geq 0\}$ is a HMC. We define $\mathcal{S}$ as the set of all the N possible states of the HMC. Let $\mathbf{P}$ be its N-by-N transition probability matrix, such that

$$P_{i,j} = \ \mathrm{prob}\,(S_{t+1} = s_j \,|\, S_t = s_i)\,, \quad \forall\, t \geq 0,\ s_i, s_j \in \mathcal{S}.$$

We point out that the codeword length is supposed to be very long, such that the conditions of applicability of the Shannon Capacity (i.e. infinite codeword) are practically satisfied. This assumption is widely applied in quasi-static channels (see e.g. Katz and Shamai, 2005 [44]).

Markovian feasibility region

In each channel state, we consider a Gaussian MAC scenario, in which K users communicate with a single receiver. By relying on the classic quasi-static approximation assumption (see e.g. Katz and Shamai, 2005 [44]), we can compute the capacity rate region for all users in state s as the polymatroid $\mathcal{R}(\mathcal{P}, s)$ with rank function $g_{(\mathcal{P})}$ (Tse and Viswanath, 2004 [94]):

$$\mathcal{R}(\mathcal{P}, s) = \left\{ \mathbf{r} \in \mathbb{R}^K : \sum_{k \in \mathcal{T}} r_k \leq g_{(\mathcal{P})}(\mathcal{T}, s),\ \forall\, \mathcal{T} \subseteq \mathcal{P} \right\}$$

$$g_{(\mathcal{P})}(\mathcal{T}, s) := C\bigg(\sum_{k \in \mathcal{T}} |h^{(k)}(s)|^2 P_k, N_0 \bigg), \quad \forall\, \mathcal{T} \subseteq \mathcal{P},$$

where $C(a, b) = \log_2(1 + a/b)$. When considering the channel dynamics, an HMC evolves on a finite set of channel states $\mathcal{S} = \{s_1, \ldots, s_N\}$. Since we consider the channel to be constant during a codeword, the transition among states occurs at the end of each coherence period of the channel.
We allocate a rate to each user in each of the state of the Markov chain. We assume that the rate assigned in state at time t, $S_t \in \mathcal{S}$, depends only on the value of S_t, and not on the past history of state/allocations up to time t. In this sense, we say that the dynamic allocation is *stationary*, and we call $r_k(s)$ the rate assigned to user k in state s. We also assume that the length of the communication is finite, but of unknown duration. Typically (e.g. Section 3.1), this situation is dealt with by considering a probability $1 - \beta$ that, at any time step, the communication terminates. Then, the expected sum of the rates assigned to user k over the sequence of channel states equals

$$r_k(\Gamma_s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \beta^t \, r_k(S_t)\right), \tag{3.27}$$

where Γ_s is the Markov process starting at time 0 in state s. We anticipate that the choice of β is irrelevant to all our results. By recalling the relation $\sum_{t\geq 0} \beta^t \mathbf{P}^t = (\mathbf{I}-\beta\mathbf{P})^{-1}$, we can write (3.27) in the following matricial form:

$$\begin{bmatrix} \mathbf{r}(\Gamma_{s_1}) \\ \vdots \\ \mathbf{r}(\Gamma_{s_N}) \end{bmatrix} = (\mathbf{I} - \beta\mathbf{P})^{-1} \begin{bmatrix} \mathbf{r}(s_1) \\ \vdots \\ \mathbf{r}(s_N) \end{bmatrix}, \tag{3.28}$$

where $\mathbf{r}(s) := [r_1(s), r_2(s), \ldots, r_K(s)]$ and $\mathbf{r}(\Gamma_s)$ is defined similarly. By defining $\Psi := (\mathbf{I}-\beta\mathbf{P})^{-1}$ and utilizing a compact matrix notation, we rewrite (3.28) as

$$[\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}} = \Psi \; [\mathbf{r}(s)]_{s\in\mathcal{S}} \tag{3.29}$$

Remark 7. *Expression (3.29) defines an application from the set of stationary state-wise rate allocations to the set of feasible long-run rates. In Section 3.2.3 we will show that, in general, the application is* not invertible, *since multiplying a set of long-run allocations by Ψ^{-1} does not always produce feasible state-wise allocations.* □

It is natural to define the long-run rate region $\mathcal{R}(\mathcal{P}, \Gamma_s)$ as the set of all rates $\mathbf{r}(\Gamma_s)$ that can be written as the long-run expected sum of stationary state-wise rate allocations, as in (3.28). We now give a convenient expression for $\mathcal{R}(\mathcal{P}, \Gamma_s)$, which follows from (Herzog and Hibi, 2010 [39]), p. 241, Theorem 12.1.5, claiming that the sum of polymatroids is still a polymatroids whose rank function is the sum of the rank functions of the summands.

Lemma 3.2.1. *For any $s_j \in \mathcal{S}$, the long-run rate feasibility region $\mathcal{R}(\mathcal{P}, \Gamma_{s_j})$ is a polymatroid with rank function:*

$$g_{(\mathcal{P})}(\mathcal{T}, \Gamma_{s_j}) = \sum_{n=1}^{N} \nu_n(s_j)\, g_{(\mathcal{P})}(\mathcal{T}, s_n), \quad \forall \mathcal{T} \subseteq \mathcal{P},$$

where $\boldsymbol{\nu}(s_j)$ is the j-th row of the matrix Ψ. □

Relevance to LTE systems

In LTE systems, the statistics of the channel are estimated at regular intervals and used for resource allocation. Under the common assumption of fast fading Gaussian channel in additive Gaussian noise, in each period t the state of the HMC is given by the channel distribution, completely characterized by its second-order statistics. The rate region in absence of instantaneous knowledge of the channel at the transmitter is still a polymatroid, with rank function $\mathbb{E}_h[g_{(\mathcal{P})}(\mathcal{T}, s)]$, as shown in [80]. Since the results presented in the following strongly rely on the polymatroid structure of the rate region in each state of the HMC, then they also hold for LTE systems. Hence, our general results in particular address the issue of allocating the rate to users in a MAC LTE system at each feed-back time interval, so that optimality, fairness, and the users' satisfaction is preserved throughout the communication.

3.2.3 Optimal and fair rate allocation design

In this section we address the issue of allocating the rate to all users during the transmission process, in each state of the channel Markov chain. For a classic result on polymatroids (see e.g. Herzog and Hibi, 2010 [39]), we know that the dominant facet, or simply facet, $\mathcal{M}(\mathcal{R}(\mathcal{P}, s))$ of the rate region $\mathcal{R}(\mathcal{P}, s)$ is maximum sum-rate, i.e.

$$\mathcal{M}(\mathcal{P}, s) := \mathcal{M}(\mathcal{R}(\mathcal{P}, s)) = \underset{\mathbf{r} \in \mathcal{R}(\mathcal{P}, s)}{\operatorname{argmax}} \sum_{k \in \mathcal{P}} r_k. \tag{3.30}$$

Similarly, the facet $\mathcal{M}(\mathcal{P}, \Gamma_s)$ is maximum sum-rate in the long-run process Γ_s. Hence, the global optimum rate design solution would be that both the state-wise and the long-run rate allocations belong to the facets $\mathcal{M}(\mathcal{P}, s)$ and $\mathcal{M}(\mathcal{P}, \Gamma_s)$, for all $s \in \mathcal{S}$. So, we will restrict our focus on the allocations inside $\mathbf{M}$, defined in the following.

Definition 10 ($\mathcal{M}$). $\mathbf{M}$ *is the set of stationary state-wise allocations belonging to the dominant facets of both state-wise and long-run feasibility regions,* *i.e.*

$$\begin{aligned}\mathbf{M} := \Big\{ \{\mathbf{r}(s)\}_{s\in\mathcal{S}} \colon \mathbf{r}(s) \in \mathcal{M}(\mathcal{P}, s),& \\ \mathbf{r}(\Gamma_s) \in \mathcal{M}(\mathcal{P}, \Gamma_s),\, \forall\, s \in \mathcal{S} \Big\},& \\ \text{where } [\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}} = \Psi\,[\mathbf{r}(s)]_{s\in\mathcal{S}}.& \qquad \Box\end{aligned}$$

Now, we will investigate two different approaches to select an allocation in **M**. The first, called bottom-up procedure (Section 3.2.3), is the most natural one, and it prescribes to select a set of state-wise allocations in $\mathcal{M}(\mathcal{P}, s)$, for all $s \in \mathcal{S}$, and then to derive the set of associated long-run allocations via multiplication by Ψ. Conversely, the second approach, dubbed top-down (Section 3.2.3), would be more useful, but unfortunately it is not always feasible. It suggests to select first the long-run allocations, in $\mathcal{M}(\mathcal{P}, \Gamma_s)$, for all $s \in \mathcal{S}$, and then to multiply by Ψ^{-1} to obtain the state-wise allocations. Clearly, the choice over the adopted procedure depends on the priority that the designer gives to the state-wise/long-run allocation. By adopting the top-down procedure, one embraces a long-run perspective of the process, by preferring to adhere to a specific fairness selection criterion in the long-run process, rather than in the state-wise one. Clearly, the best scenario would consist in being fair in each state, in the long-run process, and from each intermediate step onwards. A sufficient condition to attain this will be provided in Section 3.2.3.

BOTTOM-UP DESIGN: From single-stage to long-run allocations

In this section we investigate the feasibility of our first procedure to select an allocation in **M**. It is called *bottom-up* rate allocation approach, and it consists in selecting a set of stage-wise allocations belonging to the dominant facet of each state-wise feasibility region. Then, we need to compute the respective long-run allocations and check whether they belong to the dominant facets of the feasibility region of the respective long-run processes. By a linearity argument, it is easy to see that the facet $\mathcal{M}(\mathcal{P}, \Gamma_s)$ is obtained as the Minkowski sum $\sum_{n=1}^{N} \nu_n(s)\mathcal{M}(\mathcal{P}, s_n)$. Therefore, if the state-wise allocations all belong to the dominant facet in the respective states, then their expected long-run sum also lies in the dominant facet of the long-run process. Then, the bottom-up procedure always produces stationary allocations belonging to **M**.

Proposition 3.2.2 (Bottom-up allocation procedure)**.** *Select s set of state-wise rate allocations* $\{\mathbf{r}(s) \in \mathcal{M}(\mathcal{P}, s)\}_{s\in\mathcal{S}}$. *Then, their associated long-run allocations* $[\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}} = \Psi[\mathbf{r}(s)]_{s\in\mathcal{S}}$ *belong to the respective long-run dominant facets, i.e.* $\mathbf{r}(\Gamma_s) \in \mathcal{M}(\mathcal{P}, \Gamma_s)$, *for all* $s \in \mathcal{S}$. □

Then, the first positive result of Proposition 3.2.2 is that there exist allocations belonging to the dominant facet of both state-wise and long-run processes, jointly, i.e. $\mathbf{M}$ is non-empty. Secondly, it is easy to find them, since it suffices to select a rate allocation on the dominant facet of $\mathcal{R}(\mathcal{P}, s)$, for all $s \in \mathcal{S}$. Finally, as a by-product of Proposition 3.2.2, we are allowed to simplify the definition of $\mathbf{M}$ as:

$$\mathbf{M} \equiv \Big\{ \{\mathbf{r}(s)\}_{s\in\mathcal{S}} \text{ s.t. } \mathbf{r}(s) \in \mathcal{M}(\mathcal{P}, s), \; \forall\, s \in \mathcal{S} \Big\}.$$

TOP-DOWN DESIGN: From long-run to single-stage allocations

The bottom-up procedure always produces feasible allocations, but it is not what really concerns us. Indeed, the users are endowed with a long-term perspective of the communication process, hence one may wish to select first a set of long-run allocations in $\{\mathcal{M}(\mathcal{P}, \Gamma_s)\}_{s\in\mathcal{S}}$ which adhere to a certain criterion in the respective long-run processes (e.g. a fairness criterion, as in Section 3.2.3). Then, the state-wise rate allocations $\{\mathbf{r}(s)\}_{s\in\mathcal{S}}$ are obtained via multiplication by Ψ^{-1}. Unfortunately this method, dubbed *top-down*, does not always produces feasible stationary state-wise allocations. We interpret this fact by saying that the linear application defined by Ψ in (3.29) is not always invertible in the space of feasible stationary allocations. In Example 3.2.1 we show an instance of the described scenario.

Example 3.2.1. *Set $\beta = 0.8$, $N_0 = 0.1\,W$. Consider two users, with power constraints $P_1 = P_2 = 2\,W$. Consider two states. In s_1, $|h^{(1)}(s_1)|^2 = 0.1$, $|h^{(2)}(s_1)|^2 = 0.2$. In s_2, $|h^{(1)}(s_2)|^2 = 0.15$, $|h^{(2)}(s_2)|^2 = 0.15$. The transition probability matrix is $\mathbf{P} = [0.8\ 0.2;\ 0.3\ 0.7]$. Choose the optimal allocations in the long-run process*

$$\mathbf{r}(\Gamma_{s_1}) = [0.5843;\ \ 1.1109] \in \mathcal{M}(\mathcal{P}, \Gamma_{s_1}) \text{ bits/s/Hz}$$
$$\mathbf{r}(\Gamma_{s_2}) = [0.8270;\ \ 0.8682] \in \mathcal{M}(\mathcal{P}, \Gamma_{s_2}) \text{ bits/s/Hz.}$$

The corresponding state-wise allocations, through Ψ^{-1}, are both not feasible, because

$$\mathbf{r}(s_1) \cong [0.0780;\ \ 0.2610] \notin \mathcal{R}(\mathcal{P}, s_1)$$
$$\mathbf{r}(s_2) \cong [0.2236;\ \ 0.1154] \notin \mathcal{R}(\mathcal{P}, s_2) \qquad \square$$

Remark 8. *One may argue that there is no need to select the whole set of long-run allocations $\{\mathbf{r}(\Gamma_s)\}_{s\in\mathcal{S}}$, but only the one corresponding to the actual initial state. Indeed, since the channel state S_0 at time 0 is known, one could select $\mathbf{r}(\Gamma_{S_0})$ according to the desired criterion and then compute the state-wise allocations by choosing one solutions among the infinite possible of the* *equation*

$$\mathbf{r}(\Gamma_{S_0}) = \sum_{n=1}^{N} \nu_n(S_0)\,\mathbf{r}(s_n).$$

Finally, the remaining long-run allocations are automatically computed by re-inverting the relation, as $\Psi[\mathbf{r}(s)]_{s\in\mathcal{S}}$*. Of course, in this way there is no control over the long-run allocations* $\mathbf{r}(\Gamma_s)$*, with* $s \neq S_0$*.*
On the other hand, thanks to the stationarity of the pay-off allocation, the long-run sub-process starting at time $T > 0$ *is precisely the* β^T*-scaled version of* Γ_{S_T}*, i.e.*

$$\mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}(S_t) \;\middle|\; \mathbf{h}(T)\right) = \beta^T \mathbf{r}(\Gamma_{S_T}),$$

where $\mathbf{h}(T)$ *is the history of state/allocations from time 0 up to time* T*. Therefore, jointly choosing the long-run allocations* $\mathbf{r}(\Gamma_s)$ *for* all *states* $s \in \mathcal{S}$ *is equivalent to assign the long-run allocations that each user obtains in each sub-process from any intermediate time step* $T \geq 0$ *onwards.* □

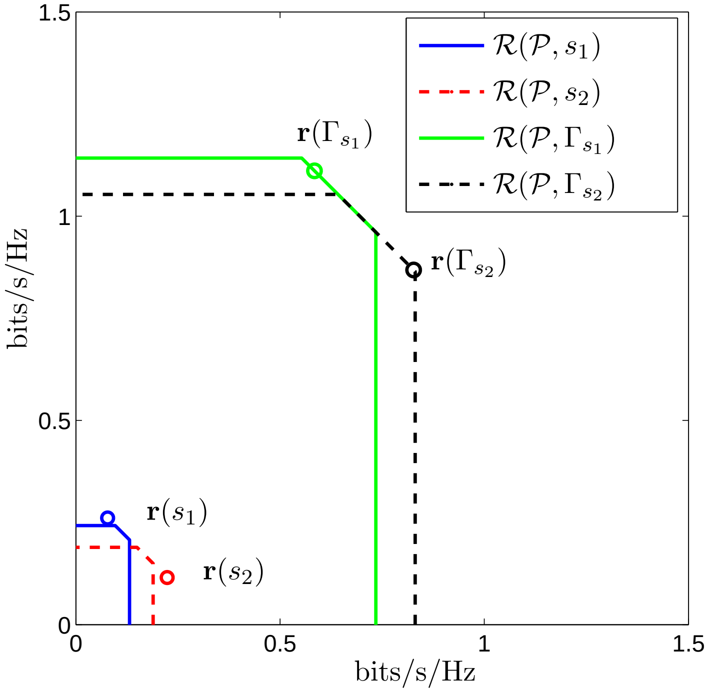

Figure 3.2: Example 3.2.1. $\mathbf{r}(\Gamma_s) \in \mathcal{M}(\mathcal{P}, \Gamma_s)$, for $s = s_1, s_2$, but $\mathbf{r}(s) \notin \mathcal{R}(\mathcal{P}, s)$, for $s = s_1, s_2$, where $[\mathbf{r}(s)]_{s\in\mathcal{S}} = \Psi^{-1}[\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}}$.

Example 3.2.1 seems to discourage a top-down allocation procedure. Indeed in general, if one chooses a set of long-run allocations, there is no guarantee that the allocation is actually feasible, since the associated stationary state-wise allocation might be not feasible. Of course, this does not rule out the possibility to carry out a top-down allocation procedure successfully. Indeed, in Theorem 3.2.4 we will present a top-down procedure guaranteeing the feasibility of the associated state-wise rate allocations. Before, let us introduce a classic result on polymatroids (see Edmonds, 2003 [30]). Let $\mathcal{R}$ be a polymatroid on the ground set $\{1, \ldots, K\}$, with rank function g. Let $\Pi(K)$ be the set of permutations of $\{1, \ldots, K\}$. The facet $\mathcal{M}(\mathcal{R})$ has at most $K!$ extreme points, and each of them has an explicit characterization as a function of the rank function g. Indeed, $\mathbf{w}$ is a vertex of $\mathcal{M}(\mathcal{R})$ if and only if there exists a permutation π of $\{1, \ldots, K\}$ such that, for all $k = 1, \ldots, K$,

$$\mathbf{w}_k = \; g(\{\pi_1, \ldots, \pi_{k-1}, \pi_k\}) - g(\{\pi_1, \ldots, \pi_{k-1}\}) := \; \mathbf{w}_k(\pi).$$

Proposition 3.2.3. *Let $a_n \geq 0$, for $n = 1, \ldots, N$. Let $\mathcal{R}_1, \ldots, \mathcal{R}_N$ be N polymatroids on the ground set $\{1, \ldots, K\}$. Let $\mathcal{R} = \sum_{n=1}^{N} a_n \mathcal{R}_n$. Let $\mathbf{w}(\pi)(n)$ be the vertex of the facet $\mathcal{M}(\mathcal{R}_n)$ associated to the permutation $\pi \in \Pi(K)$. Let $\mathbf{w}(\pi)$ be a vertex of $\mathcal{M}(\mathcal{R})$. Then,*

$$\mathbf{w}(\pi) = \sum_{n=1}^{N} a_n \, \mathbf{w}(\pi)(n), \quad \forall\, \pi \in \Pi(N). \qquad \square$$

Proposition 3.2.3 claims that the vertex of the facet $\mathcal{M}(\mathcal{R})$ associated to the permutation π can be decomposed into the sum of the vertices associated to the same π of each facet $\mathcal{M}(\mathcal{R}_n)$, $n = 1, \ldots, N$. Then, our idea is to choose *one* set of convex coefficients, valid for any $s \in \mathcal{S}$, and to define the set of long-run allocations $\{\mathbf{r}(\Gamma_s) \in \mathcal{M}(\mathcal{P}, \Gamma_s)\}_{s \in \mathcal{S}}$ as the same convex combination of the vertices of the respective dominant facets. The associated state-wise allocations are then obtained as the *same* convex combination of the vertices of the respective state-wise dominant facets, hence they are feasible and optimal.

Theorem 3.2.4 (Top-down allocation procedure)**.** *Choose a set of convex coefficients $\{c(\pi)\}_{\pi \in \Pi(K)}$, such that $c(\pi) \geq 0$ and $\sum_{\pi \in \Pi(K)} c(\pi) = 1$. Let $\mathbf{w}(\pi)(\Gamma_s)$ be the vertex of $\mathcal{M}(\mathcal{P}, \Gamma_s)$ associated to the permutation π. Compute the set of long-run allocations as*

$$\mathbf{r}(\Gamma_s) = \sum_{\pi \in \Pi(K)} c(\pi) \, \mathbf{w}(\pi)(\Gamma_s), \quad \forall\, s \in S.$$

Then,

$$[\mathbf{r}(s)]_{s \in S} = \; \Psi^{-1} \, [\mathbf{r}(\Gamma_s)]_{s \in S}$$

is a set of feasible state-wise rate allocations, and moreover $\mathbf{r}(s) \in \mathcal{M}(\mathcal{P}, s)$, for all $s \in S$. □

Proof. Let us write

$$\begin{bmatrix}\mathbf{r}(s_1)\\ \vdots \\ \mathbf{r}(s_N)\end{bmatrix} = \Psi^{-1}\begin{bmatrix}\sum_{\pi\in\Pi(K)} c(\pi)\mathbf{w}(\pi)(\Gamma_{s_1})\\ \vdots \\ \sum_{\pi\in\Pi(K)} c(\pi)\mathbf{w}(\pi)(\Gamma_{s_N})\end{bmatrix}$$

$$= \sum_{\pi\in\Pi(K)} c(\pi)\Psi^{-1}\begin{bmatrix}\mathbf{w}(\pi)(\Gamma_{s_1})\\ \vdots \\ \mathbf{w}(\pi)(\Gamma_{s_N})\end{bmatrix}.$$

For Proposition 3.2.3, we can say that

$$\begin{bmatrix}\mathbf{r}(s_1)\\ \vdots \\ \mathbf{r}(s_N)\end{bmatrix} = \sum_{\pi\in\Pi(K)} c(\pi)\begin{bmatrix}\mathbf{w}(\pi)(s_1)\\ \vdots \\ \mathbf{w}(\pi)(s_N)\end{bmatrix}.$$

Hence, the thesis is proven. □

The top-down allocation procedure provided in Theorem 3.2.4 is not the only possible of course, but it leads to an intuitive remark. Each vertex $\mathbf{w}(\pi)(s)$ can be achieved by letting the receiver decode sequentially, in the reverse order of π, the signals coming from each user in channel state $s \in S$, and by considering the signals not decoded yet as Gaussian noise (e.g. see Tse and Viswanath, 2004 [94]). Therefore, any rate allocation on $\mathcal{M}(\mathcal{P}, s)$ can be achieved by time sharing such decoding configurations, and *the time-sharing procedure is independent of the state s*.
We suggest an interesting future research, which may study how to optimize the convex coefficients $c(\pi)$ to make the resulting long-run allocations globally close to the set of long-run allocations fulfilling a certain criterion, e.g. the fairness criterion that we will present in the next section.

FAIR ALLOCATION DESIGN: being fair throughout the process

In this section we deal with a fairness criterion to select an allocation rate inside $\mathbf{M}$. In the static channel case, it is possible to find rate allocations which are fair, under plenty of different criteria (see the Introduction). In the dynamic case, the definition of fairness is much more demanding, and not always there exist allocations fulfilling it. Firstly, we demand an allocation to be fair in the long-run process, since users are endowed with a long-term perspective of the transmission process. Then, the top-down procedure would be best, because it would guarantee the rate allocations to be fair in the long-run. On the other hand, in Section 3.2.3 we showed that this approach not always produces feasible stationary rate allocations. Secondly, we demand that an allocation respects the fairness criterion not only from the beginning of the transmission onwards, but throughout it, i.e. it should be Time Consistent. Thirdly, we wish that the rate allocation is also fair in each state of the HMC. We will see that these three conditions are not generally satisfied, however we provide a sufficient condition for them to hold.

Fairness criteria: a review Let us first introduce the fairness criteria that we will utilize in the next section. In the literature, three fair allocations have been extensively studied: α-fair, max-min fair, and proportional fair allocations. We now provide their general definition, by considering a general rate feasibility region $\mathcal{R}$.

Definition 11 (max-min fairness)**.** *An allocation* $\mathbf{r}^{(\mathrm{MM})}$ *is max-min fair whenever no user* j *with rate* $\mathbf{r}_j^{(\mathrm{MM})}$ *can yield resources to a user* i *with* $\mathbf{r}_i^{(\mathrm{MM})} < \mathbf{r}_j^{(\mathrm{MM})}$ *without violating feasibility in* $\mathcal{R}$*.* □

Definition 12 (α-fairness)**.** *Let us assume that each user* k *possesses a utility function on rate,* $u^{(\alpha)}(r_k) = r_k^{1-\alpha}/[1-\alpha]$*. The* α*-fair allocation* $\mathbf{r}^{(\alpha\mathrm{F})}$*, with* $\alpha \geq 0$*, is defined as*

$$\mathbf{r}^{(\alpha\mathrm{F})} = \underset{\mathbf{r}\in\mathcal{R}}{\operatorname{argmax}} \sum_{k=1}^{K} u^{(\alpha)}(r_k). \qquad \square$$

Definition 13 (proportional fairness)**.** *The proportional fair allocation* $\mathbf{r}^{(\mathrm{PF})}$ *coincides with the* α*-fair allocation when* $\alpha \to 1$*, i.e.*

$$\mathbf{r}^{(\mathrm{PF})} = \underset{\mathbf{r}\in\mathcal{R}}{\operatorname{argmax}} \prod_{k=1}^{K} r_k. \qquad \square$$

We point out that, in general, the α-fair allocation is also max-min fair for $\alpha \uparrow \infty$ and proportional fair for $\alpha \to 1$.
If we consider the long-run process Γ_s, then in Definitions 11, 12, and 13 we should interpret $\mathcal{R} \equiv \mathcal{R}(\mathcal{P}, \Gamma_s)$, while in channel state s, $\mathcal{R} \equiv \mathcal{R}(\mathcal{P}, s)$.
In the special case in which the feasibility region is a polymatroid, which is our case both in Γ_s and in state s, for all $s \in \mathcal{S}$, then the three fair allocations coincide.

Theorem 3.2.5 ([5])**.** *If the feasibility region is a polymatroid* $\mathcal{R}$*, then max-min, proportional, and* α*-fair allocations coincide for all* $\alpha \geq 0$*, and moreover belong to the facet* $\mathcal{M}(\mathcal{R})$ *i.e.*

$$\mathbf{r}^{(\mathrm{MM})} = \mathbf{r}^{(\mathrm{PF})} = \mathbf{r}^{(\alpha\mathrm{F})} := \mathbf{r}^{(\mathrm{F})} \in \mathcal{M}(\mathcal{R}). \qquad \square$$

For Theorem 3.2.5, the three mentioned fair solutions coincide both in the long-run process Γ_s and in state s, for all $s \in \mathcal{S}$. Therefore, we can generally refer to them as *fair allocations*, and we call $\mathbf{r}^{(\mathrm{F})}(\Gamma_s)$ the fair allocation in the long-run process Γ_s, and $\mathbf{r}^{(\mathrm{F})}(s)$ the fair allocation in state s. Moreover, a fair allocation belongs to the dominant facet of the associated feasibility region, hence it is a proper criterion to select a set of allocations in $\mathbf{M}$.

Fair allocation design Finally, we are ready to deal with the design of fair rate allocations on quasi-static channels. We will show under which conditions it is possible to allocate a rate which is *fair* (i.e. max-min, proportional, and α-fair at the same time) both in each state and in the long-run process, and which is fair throughout the game, from each intermediate step, i.e. it is Time Consistent. More formally, we look for a sufficient condition for which the following holds:

$$\begin{cases} \Psi^{-1}\,[\mathbf{r}^{(\mathrm{F})}(\Gamma_s)]_{s\in\mathcal{S}} = \;[\mathbf{r}^{(\mathrm{F})}(s)]_{s\in\mathcal{S}} \\ \Psi\,[\mathbf{r}^{(\mathrm{F})}(s)]_{s\in\mathcal{S}} = \;[\mathbf{r}^{(\mathrm{F})}(\Gamma_s)]_{s\in\mathcal{S}}. \end{cases} \tag{3.31}$$

We stress that property (3.31) is crucial, mainly for three reasons, that we list below.

- The top-down procedure may fail, hence if we choose $\{\mathbf{r}^{(\mathrm{F})}(\Gamma_s)\}_{s\in\mathcal{S}}$, not necessarily it is feasible among the stationary allocations, i.e. in general it may happen that

$$\exists\, s \in \mathcal{S} \;:\; \mathbf{r}(s) \notin \mathcal{R}(\mathcal{P}, s),$$
$$\text{with } [\mathbf{r}(s)]_{s\in\mathcal{S}} = \Psi^{-1}\,[\mathbf{r}^{(\mathrm{F})}(\Gamma_s)]_{s\in\mathcal{S}}.$$

- Though the bottom-up procedure always produces feasible allocations, if the allocation is fair in each state, then not necessarily it is also fair in the long-run processes. Indeed, it may happen that

$$\exists\, s \in \mathcal{S} \;:\; \mathbf{r}(\Gamma_s) \neq \mathbf{r}^{(\mathrm{F})}(\Gamma_s),$$
$$\text{with } [\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}} = \Psi\,[\mathbf{r}^{(\mathrm{F})}(s)]_{s\in\mathcal{S}} \tag{3.32}$$

  As an example, in Figure 3.3 we show an instance in which (3.32) is verified.

- Most importantly, if relation (3.31) holds, then the fairness property of the rate allocation is *Time Consistent* (see Theorem 3.2.6).

The Time Consistency of fair allocations claims that the fairness criteria that induces to enforce a certain rate allocation at time 0 should be consistent in time, at steps $T > 0$ as well. More formally, at each time step T, the

β-discounted sum of allocations that each user obtains from time T onwards should be fair in the long-run process Γ_{S_T}.

Theorem 3.2.6. *If condition (3.31) holds, then the fairness of the stationary rate allocation* $\{\mathbf{r}^{(\mathrm{F})}(s)\}_{s\in\mathcal{S}}$ *is* Time Consistent, *i.e. for all* $T \in \mathbb{N}_0$,

$$\mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}^{(\mathrm{F})}(S_t) \;\Big|\; \mathbf{h}(T)\right) = \beta^T \mathbf{r}^{(\mathrm{F})}(\Gamma_{S_T}),$$

where $\mathbf{h}(T)$ *is the history of states/rate allocations up to time* T. □

Proof. Thanks to the stationarity of the rate allocations, we claim

$$\begin{aligned}\mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}^{(\mathrm{F})}(S_t) \;\Big|\; \mathbf{h}(T)\right) &= \mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}^{(\mathrm{F})}(S_t) \;\Big|\; S_T\right) \\ &= \beta^T\, \mathbb{E}\left(\sum_{t=0}^{\infty} \beta^t \mathbf{r}^{(\mathrm{F})}(S_{t+T}) \;\Big|\; S_T\right) \\ &= \beta^T\, \mathbf{r}^{(\mathrm{F})}(\Gamma_{S_T}). \qquad (3.33)\end{aligned}$$

where (3.33) comes from condition (3.31). Hence, the thesis is proven. □

After presenting the appealing properties of condition (3.31), we wish to find a sufficient condition for (3.31) to hold. For this purpose, it is useful to present first an algorithm, first studied in (Maddah-Ali, 2009 [53]), that produces the fair allocation in a general polymatroid $\mathcal{R}$ with rank function g. Of course, it can be utilized to compute the fair allocation in any state-wise and long-run process.

Algorithm 3.2.7 ([53])**.** *Set* $q := 1$*. Set* $\mathcal{P}' := \mathcal{P}$, $g' := g$.

1) Compute

$$\mathcal{T}^*_{(q)} = \operatorname*{argmin}_{\mathcal{T}\subseteq\mathcal{P}'} \frac{g'(\mathcal{T})}{|\mathcal{T}|}, \quad \mathbf{r}^{(\mathrm{F})}_k = \frac{g'(\mathcal{T}^*_{(q)})}{|\mathcal{T}^*_{(q)}|}, \quad \forall\, k \in \mathcal{T}^*_{(q)}.$$

2) If $\mathcal{T}^*_{(q)} = \mathcal{P}'$*, then stop. The rate allocation* $\mathbf{r}^F$ *is fair for* $\mathcal{R}$*. Otherwise, set* $q := q+1$, $\mathcal{P}' := \mathcal{P}'\backslash\mathcal{T}^*_{(q)}$,

$$g'(\mathcal{T}) := g'(\mathcal{T}\cup\mathcal{T}^*_{(q)}) - g'(\mathcal{T}^*_{(q)}), \quad \forall\, \mathcal{T} \subseteq \mathcal{P}',$$

and return to step 1). □

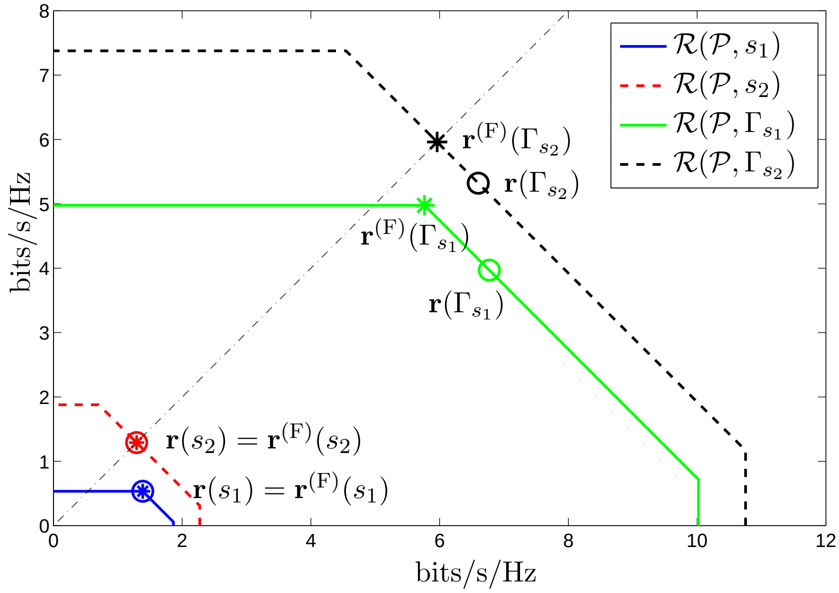

Figure 3.3: Example of situation in (3.32) with two users and two states, in which the state-wise allocations are fair in the respective channel states but the relative long-run allocations are not fair in the respective long-run processes. The allocations indicated with the asterisk are fair, while the circle describes the actual computed allocations.

Finally, we are ready to show a condition that ensures the existence of a rate allocation design which is fair both in each state and in every long-run process, as described in (3.31), and for which the fairness criterion is Time Consistent, as shown in Theorem 3.2.6.

Theorem 3.2.8 (SC existence fair allocations)**.** *Let* $\overline{\mathcal{T}}(s) = [\mathcal{T}^*_{(1)}(s),\ldots,\mathcal{T}^*_{(q(s))}(s)]$ *be the sequence computed in the iterations of step 1, Algorithm 3.2.7, applied to channel state* s*. Suppose that*

$$\exists\ \overline{\mathcal{T}} = \overline{\mathcal{T}}(s), \quad \forall\, s \in \mathcal{S},$$

i.e. $\overline{\mathcal{T}}(s)$ *does not depend on* s*. Then, condition (3.31) holds.* □

Proof. At step 1 of the first iteration of Algorithm 3.2.7 applied to the process Γ_s, we obtain

$$\mathcal{T}^*_{(1)}(\Gamma_s) = \underset{\mathcal{T}\subseteq\mathcal{P}}{\operatorname{argmin}} \frac{\sum_{n=1}^{N} \nu_n(s)\, g_{(\mathcal{P})}(\mathcal{T}, s_n)}{|\mathcal{T}|} = \mathcal{T}^*_{(1)}.$$

Hence, we can compute the fair allocation for the set of users $\mathcal{T}^*_{(1)}$ as $r_k^{\mathrm{F}}(\Gamma_s) = \sum_{n=1}^{N} \nu_n(s) r_k^{\mathrm{F}}(s_n)$, for all $k \in \mathcal{T}^*_{(1)}$. Then, at step 2, the update of the rank function:

$$g'_{(\mathcal{P})}(\mathcal{T}, \Gamma_s) = \sum_{n=1}^{N} \nu_n(s)\, g'_{(\mathcal{P})}(\mathcal{T}, s_n), \quad \forall\, \mathcal{T} \subseteq \mathcal{P} \backslash \mathcal{T}^*_{(1)}$$

preserves the linearity property of the rank function also in the next iteration. Hence, by induction, the thesis is proven. □

3.2.4 Optimal and Satisfactory allocations: A game-theoretical approach

In (3.30) we have defined the set of global optimum rate region $\mathbf{M}$, as the set of stationary state-wise allocations belonging to the dominant facets of both state-wise and long-run feasible rate regions. In this section we are going to provide two further characterizations of $\mathbf{M}$, in cooperative game-theoretical terms. We will show indeed that $\mathbf{M}$, besides being global optimum, also "satisfies" all the users throughout the game. Hence, in our case, CGT is utilized as a mathematical tool which allows us to define and quantify the users' satisfaction with the assigned rate allocation.

CORE characterization of M

Generally speaking, Static Cooperative Game Theory (SCGT) with non-transferable utility (NTU) studies one-shot interactions among different players who can collaborate with each other by coordinating the respective strategies. It is assumed that grand coalition $\mathcal{P}$, composed by all the players, is formed, and the main challenge consists in devising a pay-off allocation for each player, according to some pre-defined criteria. To this aim, the typical procedure in SCGT consists in investigating the *potential* scenario in which a sub-coalition (or simply, coalition) $\mathcal{C} \subset \mathcal{P}$ of players withdraws from the grand coalition and no longer coordinates its actions with the excluded players; then, the set of pay-off allocations that $\mathcal{C}$ can earn on its own is computed (see Peleg and Sudhölter, 2007 [69] for a thorough survey).
Let us then translate these preliminary few concepts into our scenario. We first consider the static process in state s, that we call *static game*. For the static game case we adopt the same model as in (La and Anantharam, 2004 [50]). In our situation, the players are the users, and the grand coalition is the set of transmitting users $\mathcal{P}$. We say that a coalition of users $\mathcal{C}_{\mathcal{J}} := \mathcal{P} \backslash \mathcal{J} \subset \mathcal{P}$ forms when its members share the respective codes with the receiver, which can then decode the signals transmitted by $\mathcal{C}_{\mathcal{J}}$. For us, the pay-off for a player is the assigned transmission rate. The SCGT literature provides several ways to compute the set of rate allocations achievable by each subset of users $\mathcal{C}_{\mathcal{J}}$. One of the most utilized is the max-min method, originally introduced by von Neumann and Morgenstern (1944) in [96], suggesting that the set of feasible allocations $\mathcal{R}(\mathcal{C}_{\mathcal{J}}, s)$ should be defined as

the *set of rate allocations that $\mathcal{C}_{\mathcal{J}}$ can achieve whatever is the transmission strategy employed by the remaining user $\mathcal{J}$.* Then, we need to take into account the *worst* possible scenario for $\mathcal{C}_{\mathcal{J}}$, i.e. when the users in $\mathcal{J}$ do not allow joint decoding and jam the network, and investigate the set of rates $\mathcal{R}(\mathcal{C}_{\mathcal{J}}, s)$ that the users in $\mathcal{C}_{\mathcal{J}}$ can achieve in this hypothetical worst-case scenario. When the users in $\mathcal{J}$ jam, they sum coherently the respective signals and transmit with an overall power:

$$\Lambda(\mathcal{J}, s) = \left(\sum_{k \in \mathcal{J}} |h^{(k)}(s)| \sqrt{P_k} \right)^2 .$$

In this worst-case scenario, in (La and Anantharam, 2004 [50]) it is shown that, among $\mathcal{C}_{\mathcal{J}}$, only the users $\widehat{\mathcal{A}}_{\mathcal{J}}$ whose associated received power level is high enough to overwhelm the jamming signal can communicate, i.e.

$$\widehat{\mathcal{A}}_{\mathcal{J}}(s) := \left\{ k \in \mathcal{C}_{\mathcal{J}} : \ |h^{(k)}(s)|^2 P_k > \Lambda(\mathcal{J}, s) \right\}.$$

Then, $\mathcal{R}(\mathcal{C}_{\mathcal{J}}, s)$ is a polymatroid with rank function (La and Anantharam, 2004 [50]):

$$g_{(\mathcal{C}_{\mathcal{J}})}(\mathcal{T}, s) := C\left(\sum_{k \in \mathcal{T}} |h^{(k)}(s)|^2 \widetilde{P}_k, \Lambda(\mathcal{J}, s) + N_0 \right),$$

where $\widetilde{P}_k = P_k$ for $k \in \widehat{\mathcal{C}}_{\mathcal{J}}(s)$ and $\widetilde{P}_k = 0$ for all $k \in \mathcal{C}_{\mathcal{J}} \backslash \widehat{\mathcal{C}}_{\mathcal{J}}(s)$.
Now, let us consider the feasibility region $\mathcal{R}(\mathcal{C}_{\mathcal{J}}, \Gamma_s)$ for a coalition $\mathcal{C}_{\mathcal{J}}$ in the long-run process (or game) Γ_s. Similarly to the static case, it is still defined in the max-min fashion, as the set of long-run rate allocations that the users $\mathcal{C}_{\mathcal{J}}$ can guarantee, whatever is the transmission strategy adopted by $\mathcal{J}$, throughout the process. Therefore, we have to consider the worst-case scenario in which $\mathcal{J}$ jams during the whole process Γ_s and, analogously to Lemma 3.2.1, we claim that $\mathcal{R}(\mathcal{C}_{\mathcal{J}}, \Gamma_s)$ is a polymatroid with rank function:

$$g_{(\mathcal{C}_{\mathcal{J}})}(\mathcal{T}, \Gamma_{s_j}) = \sum_{n=1}^{N} \nu_n(s_j)\, g_{(\mathcal{C}_{\mathcal{J}})}(\mathcal{T}, s_n), \quad \forall \mathcal{T} \subseteq \mathcal{C}_{\mathcal{J}}.$$

Our goal is now to further characterize $\mathbf{M}$, and we achieve this via the definition of the Core set for NTU cooperative games. The Core is the set of rate allocations that no coalition $\mathcal{C}_{\mathcal{J}} \subset \mathcal{P}$ can improve upon when the remaining users $\mathcal{J}$ jam. Let us define formally the Core in the static game in state s. We say that a rate allocation for the grand coalition $\mathbf{r} \in \mathcal{R}(\mathcal{P}, s)$ is *blocked* by the coalition $\mathcal{C}_{\mathcal{J}} \subseteq \mathcal{P}$ whenever there exists $\mathbf{r}' \in \mathcal{R}(\mathcal{C}_{\mathcal{J}}, s)$ such that $\mathbf{r}'_k > \mathbf{r}_k$ for all $k \in \mathcal{C}_{\mathcal{J}}$. In other words, the rate allocation $\mathbf{r}$ is unacceptable by the set of users in $\mathcal{C}_{\mathcal{J}}$.

Definition 14. *The Core* $\mathbf{Co}(s)$ *is the set of unblocked rate allocations in* $\mathcal{R}(\mathcal{P}, s)$.

Remark 9. *We can intuitively define the* Core as the set of all "acceptable" rates for all users*: indeed, if an allocation does not belong to the Core, at least a subset of users is dissatisfied with it, because they can all attain a better rate allocation even when the remaining users do not participate to the transmission and jam.* □

Additionally, please note that an allocation in $\mathbf{Co}(s)$ is also not blocked by the grand coalition $\mathcal{P}$, and since $\mathcal{R}(\mathcal{P}, s)$ is a polymatroid, it follows that it is a region with maximum sum-rate, i.e. $\mathbf{Co}(s) \subseteq \mathcal{M}(\mathcal{P}, s)$, for all $s \in \mathcal{S}$.

The Core $\mathbf{Co}(\Gamma_s)$ in the long-run game Γ_s is defined analogously to the static case. We remark that it coincides with the set of long-run allocations that are acceptable for each subset of users *at the beginning of the long-run game.* This definition of $\mathbf{Co}(\Gamma_s)$ relates to SCGT, in which the coalition structure holds steady throughout the game and players do not change their preference over the rate allocations over time. This is a naïve perspective though, since the channel is dynamic. Hence, we demand that a stationary rate allocations is not only "acceptable" for each coalition at the beginning of the game, but also throughout the game. This property is called, in dynamic CGT, *Time Consistency* of the Core (Petrosjan, 1977 [72]). The philosophy behind this definition is analogous to the Time Consistency of Fair allocations, in Theorem 3.2.6. Hence, if the Core property of an allocation is Time Consistent, then at *each* time step, if any coalition faces the dilemma "*do we withdraw now or we cooperate forever?*", it always prefers the second option. Therefore, we will focus our attention on the allocations in $\mathbf{Co}$, defined as follows, and we will prove that $\mathbf{Co} = \mathbf{M}$.

Definition 15 (**Co**)**.** **Co** *is the set of stationary state-wise allocations belonging to the Core of each static game, and that belong to the Core of long-run games in a Time Consistent fashion throughout the game, i.e.*

$$\mathbf{Co} := \Bigg\{ \{\mathbf{r}(s)\}_{s\in\mathcal{S}} \colon \mathbf{r}(s) \in \mathbf{Co}(s), \\ \mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}(S_t) \;\Bigg|\; \mathbf{h}(T) \right) \in \; \beta^T \mathbf{Co}(\Gamma_{S_T}), \; \forall T \in \mathbb{N}_0 \Bigg\},$$

where $\mathbf{h}(T)$ *be the history of states/rate allocations up to time* T. □

Hence, **Co** is the set of stationary allocations that are maximum sum-rate, hence optimum for the global network, and that are "acceptable" for each subset of users, in both static and long-run games, throughout the game. Hence, we can already claim that $\mathbf{Co} \subseteq \mathbf{M}$. Let us show that $\mathbf{Co} = \mathbf{M}$.

In [50], La and Anatharam computed the Core of the static game by relying on SCGT with transferable utilities (TU). Their approach is not completely rigorous, since the rate cannot be shared in any manner among the users, but only within the capacity region. Nevertheless, NTU cooperative game theory yields the same result as (La and Anantharam, 2004 [50]), as we show next.

Theorem 3.2.9. *The Core* $\mathbf{Co}(s)$ *coincides with the facet* $\mathcal{M}(\mathcal{P}, s)$ *of the feasibility region* $\mathcal{R}(\mathcal{P}, s)$ *for the grand coalition.* □

Proof. Is is known (e.g. Edmonds, 2003 [30]) that all the points in $\mathcal{M}(\mathcal{P}, s)$ solve the linear program $\max_{\mathbf{r}\in\mathcal{R}(\mathcal{P},s)} \sum_{k\in\mathcal{P}} r_k$. Hence, all the points in $\mathcal{M}(\mathcal{P}, s)$ are efficient for $\mathcal{P}$. Moreover, in (La and Anantharam, 2004 [50]) it is shown that, for all $\mathbf{r} \in \mathcal{M}(\mathcal{P}, s)$,

$$\sum_{k\in\mathcal{C}_\mathcal{J}} r_k \geq g_{(\mathcal{C}_\mathcal{J})}(\mathcal{C}_\mathcal{J}, s), \quad \forall \mathcal{C}_\mathcal{J} \subset \mathcal{P}.$$

Hence, we can say that, for all $\mathbf{r} \in \mathcal{M}(\mathcal{P}, s)$, there exists no allocation belonging to $\mathcal{M}(\mathcal{C}_\mathcal{J}, s)$ that dominates $\mathbf{r}$ for coalition $\mathcal{C}_\mathcal{J}$. Since any rate allocations belonging to $\mathcal{R}(\mathcal{C}_\mathcal{J}, s)$ is dominated by a rate allocation in $\mathcal{M}(\mathcal{C}_\mathcal{J}, s)$, then $\mathcal{M}(\mathcal{P}, s) \subseteq \mathbf{Co}(s)$. If $\mathbf{r} \notin \mathcal{M}(\mathcal{P}, s)$, either it is not feasible or it is not efficient for $\mathcal{P}$. Then, $\mathcal{M}(\mathcal{P}, s) = \mathbf{Co}(s)$. □

In the light of Theorem 3.2.9 and Lemma 3.2.1, we can easily provide an expression for $\mathbf{Co}(\Gamma_s)$ as well.

Corollary 3.2.10. *The Core* $\mathbf{Co}(\Gamma_s)$ *of the long-run game* Γ_s *coincides with the facet* $\mathcal{M}(\mathcal{P}, \Gamma_s)$. □

Now, we are ready to prove that $\mathbf{M} = \mathbf{Co}$.

Theorem 3.2.11. *The set of stationary state-wise rate allocations* $\mathbf{M}$ *coincides with* $\mathbf{Co}$*, i.e.* $\mathbf{M} = \mathbf{Co}$. □

Proof. We know that $\mathbf{Co} \subseteq \mathbf{M}$. We have to prove that $\mathbf{M} \subseteq \mathbf{Co}$. For Theorem 3.2.9, if $\{\mathbf{r}(s)\}_{s\in\mathcal{S}} \in \mathbf{M}$, then $\{\mathbf{r}(s) \in \mathbf{Co}(s)\}_{s\in\mathcal{S}}$. Then, we just need to prove that, if $\{\mathbf{r}(s)\}_{s\in\mathcal{S}} \in \mathbf{M}$, then the Core is Time Consistent in the long-run game. Similarly to the proof of Theorem 3.2.6, we claim that for all $T \in \mathbb{N}_0$,

$$\mathbb{E}\left(\sum_{t=T}^{\infty} \beta^t \mathbf{r}(S_t) \;\middle|\; \mathbf{h}(T)\right) = \beta^T\, \mathbf{r}(\Gamma_{S_T}),$$

where $[\mathbf{r}(\Gamma_s)]_{s\in\mathcal{S}} = \Psi[\mathbf{r}(s)]_{s\in\mathcal{S}}$ is the set of the associated long-run allocations. Thanks to Proposition 3.2.2, $\mathbf{r}(\Gamma_{S_T}) \in \mathbf{Co}(\Gamma_{S_T})$. Hence, the thesis is proven. □

Thanks to Theorem 3.2.11, the set of stationary state-wise rate allocations $\mathbf{M}$ gains further significance. Not only $\mathbf{M}$ is the maximal sum-rate region, but it also coincides with the set of rates which are "acceptable" both in the long-run and in the static games, under the definition of Core. Moreover, the Core criterion is Time Consistent, hence such rates are acceptable throughout the game.
In the next section we provide a second characterization of $\mathbf{M}$, based on a Cooperation Maintenance property.

COOPERATION MAINTENANCE characterization of M

In this section we show that, by exploiting a crucial concept in DCGT, called Cooperation Maintenance property, we are able to provide a further characterization to the set $\mathbf{M}$ of the maximum sum-rate stationary state-wise allocations. The property that we are going to define is an adaptation to our NTU scenario of the Cooperation Maintenance property defined in (Mazalov and Rettieva, 2010 [59]) and in Section 3.1. It claims that, at each time step, the maximum sum-rate that coalition $\mathcal{C}_{\mathcal{J}}$ expects to obtain if it withdraws (without any chance of joining back) from the grand coalition in one step should be not smaller than what $\mathcal{C}_{\mathcal{J}}$ obtains if it withdraws (still, without a second thought) at the current step.

Remark 10. *When we say, in a game-theoretical jargon, that a coalition $\mathcal{C}_{\mathcal{J}}$ is enticed to* withdraw *from the grand coalition, we actually mean that it is* dissatisfied *with its assigned rate, because, even in the* worst-case *scenario in which $\mathcal{J}$ jams, $\mathcal{C}_{\mathcal{J}}$ could achieve a better allocation. Hence, like in Section 3.2.4, we will utilize Game Theory as a tool to measure users' satisfaction with the assigned rate.* □

The set of allocations for which the Cooperation Maintenance property holds is called **CM**.

Definition 16 (**CM**)**.** *The set of (first step) Cooperation Maintaining allocations* **CM** *is the set of stationary state-wise rate allocations $\{\mathbf{r}(s) \in \mathcal{M}(\mathcal{P}, s)\}_{s\in\mathcal{S}}$ such that, for all coalitions $\mathcal{C}_{\mathcal{J}} \subseteq \mathcal{P}$ and at each time step*

$T \in \mathbb{N}_0$,

$$\sum_{k\in\mathcal{C}_\mathcal{J}} r_k(S_T)+\beta\sum_{s'\in\mathcal{S}} p(s'|S_T)\Big[\max_{\mathbf{r}(\Gamma_{s'})\in\mathcal{R}(\mathcal{C}_\mathcal{J},\Gamma_{s'})}\sum_{k\in\mathcal{C}_\mathcal{J}} r_k(\Gamma_{s'})\Big] \geq \max_{\mathbf{r}(\Gamma_{S_T})\in\mathcal{R}(\mathcal{C}_\mathcal{J},\Gamma_{S_T})}\sum_{k\in\mathcal{C}_\mathcal{J}} r_k(\Gamma_{S_T}). \tag{3.34}$$

□

The intuition behind the definition of **CM** is that, if a coalition faces the dilemma "*do we withdraw now or in one step?*", it should prefer the second option, at any instant. In this way, by induction, no coalition is ever enticed to withdraw and the grand coalition is cohesive throughout the game.

It follows from Definition 16 that **CM** $\subseteq$ **Co**. Also, it is not difficult to show that, if the (first step) Cooperation Maintenance property holds, then the n-tuple step Cooperation Maintenance property also holds (see Section 3.1 for a more general case), i.e. if a coalition faces the dilemma "*do we withdraw now or in n steps?*", it prefers the second option. For $n \uparrow \infty$, such property suggests that whenever a coalition faces the dilemma "*do we withdraw now or cooperate forever?*", then it prefers to stick with the grand coalition forever. Not surprisingly, this notion coincides with the Time Consistency property of the Core that any allocation in **M** possesses, as illustrated in Theorem 3.2.11.
We remark that, in more general settings, **CM** is smaller than the set of the stationary distributions belonging to the Core of long-run games (see Section 3.1). So, the definition of **CM** requires a "higher level of satisfaction" for the players than the Core. We now state that actually, in our scenario, **M** = **CM**. Through this result, we provide a second dynamic characterization of the set **M**.

Theorem 3.2.12. *The maximum sum-rate set of stationary state-wise allocations* **M** *coincides with the Cooperation Maintaining set* **CM**, *i.e.* **M** = **CM**. □

Proof. For Proposition 3.2.2, **CM** $\subseteq$ **M**. Conversely, if an allocation $\{\mathbf{r}(s)\}_{s\in\mathcal{S}} \in$ **M**, then it also belongs to **Co**. So, $\sum_{k\in\mathcal{C}_\mathcal{J}} r_k(s) \geq g_{(\mathcal{C}_\mathcal{J})}(\mathcal{C}_\mathcal{J}, s)$, for all $\mathcal{C}_\mathcal{J} \subseteq \mathcal{P}$, $s \in \mathcal{S}$. Then, thanks to Lemma 3.2.1, we can say that for all $\mathcal{C}_\mathcal{J} \subseteq \mathcal{P}$, $s \in \mathcal{S}$:

$$\sum_{k\in\mathcal{C}_\mathcal{J}} \begin{bmatrix} r_k(s_1) \\ \vdots \\ r_k(s_N) \end{bmatrix} \geq \Psi^{-1} \begin{bmatrix} g_{(\mathcal{C}_\mathcal{J})}(\mathcal{C}_\mathcal{J},\Gamma_{s_1}) \\ \vdots \\ g_{(\mathcal{C}_\mathcal{J})}(\mathcal{C}_\mathcal{J},\Gamma_{s_N}) \end{bmatrix},$$

which is an expression equivalent to (3.34). Hence, **M** $\subseteq$ **CM** and the thesis is proven. □

Therefore, in this section we have provided two game-theoretical characterizations for the global optimal set of allocations $\mathbf{M}$, i.e.

$$\mathbf{M} = \mathbf{Co} = \mathbf{CM}.$$

Hence, $\mathbf{M}$ coincides with the set of rates $\mathbf{Co}$ which are acceptable for all coalitions throughout the game, and with the set of rates $\mathbf{CM}$ that make the grand coalition cohesive at every step of the game.

3.2.5 Conclusions

In this paper we considered a quasi-static Markovian multiple access channel. We allocate the rate for each user in each channel state. We focus on the set $\mathbf{M}$ of allocations which are maximum sum-rate, both in each state and in the long run process. In Section 3.2.3 we investigate two allocation procedures, namely bottom-up and top-down. Though the latter is more useful under a long-run perspective, it does not always produce feasible allocations. Theorem 3.2.4 offers a remedy for this. In Section 3.2.3 we demand the existence of an allocation which is fair both in each state and in the long-run process. Moreover, we ask that the fairness property is Time Consistent. Theorem 3.2.8 provides a sufficient condition for this.
In Section 3.2.4 we provide two further characterizations of $\mathbf{M}$, by utilizing two different concepts in dynamic cooperative game theory, which can be considered as a measure of users' satisfaction. Firstly, in Theorem 3.2.11 we claim that $\mathbf{M}$ coincides with the Core set $\mathbf{Co}$ of allocations which are, in a sense, "acceptable" for all the users, both in the static and in the long run game, in a time consistency fashion. Secondly, in Theorem 3.2.12 we state that $\mathbf{M}$ also coincides with the set of Cooperation Maintaining allocations $\mathbf{CM}$ that makes the coalition of all players cohesive throughout the game. Therefore, all allocations in $\mathbf{CM}$ are both global optimal and satisfy the users throughout the process, according to the criteria defined by $\mathbf{Co}$ and $\mathbf{CM}$.

3.3 Confidence intervals for the Shapley-Shubik power index in Markovian games

We consider a simple Markovian game, in which several states succeed each other over time, following an exogenous discrete-time Markov chain. In each state, a different simple static game is played by the same set of players. We investigate the approximation of the Shapley-Shubik power index in the Markovian game (SSM). We prove that an exponential number of queries on coalition values is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Motivated by this, we propose and study three randomized approaches to compute a confidence interval for SSM. They rest upon two different assumptions, static and dynamic, about the process through which the estimator agent learns the coalition values. Such approaches can also be utilized to compute confidence intervals for the Shapley value in any Markovian game. The proposed methods require a number of queries which is polynomial in the number of players in order to achieve a polynomial accuracy.

3.3.1 Introduction

Cooperative game theory is a powerful tool to analyse, predict, and influence the interactions among several players capable to stipulate deals and form subcoalitions in order to pursue a common interest. Under the assumption that the grand coalition, comprising all the players, is formed, it is a delicate issue to share the payoff earned by the grand coalition among its participants.

Introduced by Lloyd S. Shapley in his seminal paper [82], the Shapley value is one of the best known payoff allocation rules in a cooperative game with transferable utility (TU). It is the only allocation procedure fulfilling three reasonable conditions of symmetry, additivity and dummy player compensation (see [82] for details), under a superadditive assumption on the coalition values. The significance of the Shapley value is witnessed by the breadth of its applications, spanning from pure economics [9] to Internet economics [13, 52, 89], politics [10], and telecommunications [43].

The concept of Shapley value was successfully applied to simple games [85], in which the coalition values are binary. In this case, the Shapley value is commonly referred to as Shapley-Shubik power index. A specific instance of simple games are weighted voting games, in which each player possesses a different amount of resources and a coalition is effective, i.e. its value is 1, whenever the sum of the resources shared by its participants is higher than a certain quota; otherwise, its value is 0. The Shapley-Shubik index proves to be particularly suitable to assess *a priori* the power of the members of a legislation committee, and has many applications to politics (see [91] for an overview).

The computation of the Shapley value for each player $j = 1, \dots, P$ involves the assessment of the increment of the value of a coalition brought on by the presence of player j, over all 2^{P-1} possible coalitions. Hence, it is clear that the complexity of Shapley value in the number of players P is a crucial issue. Mann and Shapley himself [56] were the first to suggest to adopt a Monte-Carlo procedure to approximate the Shapley-Shubik index. They first proposed a very simple algorithm, randomly generating a succession of players' permutations and evaluate the incremental value of player j with respect to the coalition formed by its preceding players in each permutation. The Shapley-Shubik index is approximated as the average of such increments. Then they empirically showed that the "cycling scheme" described below is characterized by a smaller variance. First, a target player is singled out, and the remaining players are placed in a random order. Then, this order is put through all of its cyclic permutations, and the target player is inserted in each position in each permutation. Thus, $P(P-1)$ permutations are generated, and for each of them the incremental value of player j with respect to the coalition formed by its preceding players is assessed. For this cycling approach, deriving a confidence interval for the Shapley-Shubik index seems to be a hard task. Hence, Bachrach *et al.* adopted in [14] the first Monte-Carlo procedure described above to compute a confidence interval. This approximation method, presented for simple games, can be easily generalized to any game.

The bulk of the literature on cooperative games focuses on static games. However, politics or economics is more like a process of continuing negotiation and bargaining. This motivates the introduction of dynamic cooperative game theory (see e.g. [31], [48]). In this work we consider that the game is not played one-shot but rather over an infinite horizon: there exists a finite set of static cooperative games that come one after the other, following a discrete-time homogeneous Markov process. We call this interaction model repeated over time as Markovian game. Our Markovian game model arises naturally in all situations in which several individuals keep interacting and cooperating over time, and an exogenous Markov process influences the value of each coalition, and consequently also the power of each player within coalitions. A very similar model, but with non transferable utilities, was considered in [75]. Our model can also be viewed as a particular case of the cooperative Markov decision process described in [11], or in [73], in which the transition probabilities among the states do not depend on the players' actions.

We take into account the average and the discount criterion to compute the payoff earned by each player in the long-run Markovian game. In this article we extend the approach by Bachrach *et al.* in [14] to compute a confidence interval for Shapley value in Markovian games. In [14], the authors considered a simple static game and proved that any deterministic algorithm which approximates one component of the Banzhaf index with accuracy better than $c/\sqrt{P}$, where $c > 0$ and P is the number of players, needs $\Omega(2^P/\sqrt{P})$ queries. Hence, when P grows large, it is crucial to find a suitable way to approximate the power index with a manageable number of queries. Hence, in [14] a confidence interval for Banzhaf index and Shapley-Shubik power index in simple games has been developed, based on Hoeffding's inequality. In this article we assume that the estimator agent knows the transition probabilities among the states. We first show that it is still beneficial to utilize a randomized approach to approximate the Shapley-Shubik index in simple Markovian games (SSM) for a number of players P sufficiently high. Then, we propose three methods to compute a confidence interval for the SSM, that also apply to the Shapley value of *any* Markovian game. Then, we will essentially demonstrate that, asymptotically in the number of steps of the Markov chain and by exploiting the Hoeffding's inequality, the estimator agent does not need to have access to the coalition values in all the states at the same time. Indeed, it suffices for the estimator agent to learn the coalition values in each state along the course of the game to "well" approximate SSM.

Let us overview the content of this article. We provide some useful definitions, background results, and motivations of our dynamic model in Section 3.3.2. In Section 3.3.3 we motivate the significance of our Markovian model. In Section 3.3.4 we study the trade-off between complexity and accuracy of deterministic algorithms approximating SSM. An exponential number of queries is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Motivated by this, we propose three different randomized approaches to compute a confidence interval for SSM. Their complexity does *not* even depend on the number of players. Such approaches also hold for the classic Shapley value of any cooperative Markovian game (ShM). In Section 3.3.5 we provide the expression of our first confidence interval, SCI, which relies on the static assumption that the estimator agent has access to the coalition values in all the states at the same time, even before the Markov process initiates. Although SCI relies on an impractical assumption, it is still a valid benchmark for the performance of the approaches yielding the confidence intervals described in Sections 3.3.6 and 3.3.6, dubbed DCI1 and DCI2 respectively. DCI1 and DCI2 also hold under the more realistic dynamic assumption that the estimator agent learns the value of coalitions along the course of the game. In Section 3.3.6 we propose a straightforward way to optimize the tightness of DCI1. In Section 3.3.7 we compare the three proposed approaches in terms of tightness of the confidence interval. Finally, in Section 3.3.8 we provide a trade-off between complexity and accuracy of our randomized algorithm, holding for any cooperative Markovian game.

We remark that the extension of our approaches to Banzhaf index [16] is straightforward.

Some notation remarks. If $\mathbf{a}$ is a vector, then $\mathbf{a}_i$ is its i-th component. If A is a random variable (r.v.), then A_t is its t-th realization. Given a set S, $|S|$ is its cardinality. The expression $b^{(s)}$ indicates that the quantity b, standing possibly for Shapley value, Shapley-Shubik index, coalition value, feasibility region etc., is related to the static game played in state s. The expression $\Pr(B)$ stands for the probability of event B. The indicator function is written as $\mathbb{I}(.)$. With some abuse of terminology, we will refer to a confidence interval or to the approach utilized to compute it without distinction.

3.3.2 Markovian Model and Background results

In this article we consider cooperative Markovian games with transferable utility (TU). Let P be the number of players and let $\mathcal{P} = \{1, \ldots, P\}$ be the grand coalition of all players. We have a finite set of states $S = \{s_1, \ldots, s_{|S|}\}$. In state s, each coalition $\Lambda \subseteq \mathcal{P}$ can ensure for itself the value $v^{(s)}(\Lambda)$, that can be shared in any manner among the players under the TU assumption. Hence, in each state $s \in S$ the game $\Psi^{(s)} \equiv (\mathcal{P}, v^{(s)})$ is played. Let $\mathcal{V}^{(s)}(\Lambda)$ be the half-space of all feasible allocations for coalition Λ in the TU game $\Psi^{(s)}$, i.e. the set of real $|\Lambda|$-tuple $\mathbf{a} \in \mathbb{R}^{|\Lambda|}$ such that $\sum_{i=1}^{|\Lambda|} \mathbf{a}_i \leq v^{(s)}(\Lambda)$. We suppose that the coalition values are superadditive, i.e.

$$v^{(s)}(\Lambda_1 \cup \Lambda_2) \geq \ v^{(s)}(\Lambda_1) + v^{(s)}(\Lambda_2), \qquad \forall\, \Lambda_1, \Lambda_2 \subseteq \mathcal{P},\ \Lambda_1 \cap \Lambda_2 = \emptyset.$$

The succession of the states follows a discrete-time homogeneous Markov chain, whose transition probability matrix is $\mathbf{P}$. Let $\mathbf{x}^{(s)} \in \mathbb{R}^P$ be a payoff allocation among the players in the single stage game $\Psi^{(s)}$. Under the β-discounted criterion, where $\beta \in [0;1)$, the discounted allocation in the Markovian dynamic game Γ_{s_k}, starting from state s_k, can be expressed as

$$\sum_{t=0}^{\infty} \beta^t \mathbb{E}\left(\mathbf{x}^{(S_t)}\right) = \sum_{i=1}^{|S|} \boldsymbol{\nu}_i^{(\beta)}(s_k)\, \mathbf{x}^{(s_i)}$$

where S_t is the state of Markov chain at time t and $\boldsymbol{\nu}^{(\beta)}(s_k)$ is the k-th row of the nonnegative matrix $(\mathbf{I} - \beta\mathbf{P})^{-1}$. We stress that β can be interpreted as the probability that the game terminates, at any step. Under the average criterion, if the transition probability matrix $\mathbf{P}$ is *irreducible*, then the allocation in the long-run game Γ_{s_k} can be written as

$$\limsup_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}\left(\mathbf{x}^{(S_t)}\right) = \sum_{i=1}^{|S|} \boldsymbol{\pi}_i\, \mathbf{x}^{(s_i)}$$

where $\boldsymbol{\pi}$ is the stationary distribution of the matrix $\mathbf{P}$.

We define $\mathcal{V}(\Lambda, \Gamma_s)$ as the set of feasible allocations in the long-run game Γ_s for coalition Λ, coinciding with the Minkowski sum:

$$\mathcal{V}(\Lambda, \Gamma_s) \equiv \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, \mathcal{V}^{(s_i)}(\Lambda).$$

where $\boldsymbol{\sigma}_i(s) \equiv \boldsymbol{\nu}_i^{(\beta)}(s)$ if the β-discounted criterion is adopted, and $\boldsymbol{\sigma}_i(s) \equiv \boldsymbol{\pi}_i$ under the average criterion.

Proposition 3.3.1 ([11]). *$\mathcal{V}(\Lambda, \Gamma_s)$ is equivalent to the set $\mathcal{A}$ of real $\mathbb{R}^{|\Lambda|}$-tuples $\mathbf{a}$ such that $\sum_{i=1}^{|\Lambda|} \mathbf{a}_i \leq v(\Lambda, \Gamma_s)$, where $v(\Lambda, \Gamma_s) = \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, v^{(s_i)}(\Lambda)$, for all $s \in S$, $\Lambda \subseteq \mathcal{P}$.*

Thanks to Proposition 3.3.1, it is legitimate to define $v(\Lambda, \Gamma_s)$ as the value of coalition $\Lambda \subseteq \mathcal{P}$ in the long-run game Γ_s. Let us define the Shapley value in static games [82].

Definition 17. *The* Shapley value *$\mathcal{S}h^{(s)}$ in the static game played in state $s \in S$ is a real P-tuple whose j-th component is the payoff allocation to player j:*

$$\mathcal{S}h_j^{(s)} = \sum_{\Lambda \subseteq \mathcal{P}/\{j\}} \frac{|\Lambda|!(P-|\Lambda|-1)!}{P!} \big[v^{(s)}(\Lambda \cup \{j\}) - v^{(s)}(\Lambda)\big].$$

Now, we are ready to define the Shapley value in the Markovian game Γ_s, $\mathrm{ShM}(\Gamma_s)$, that can be expressed, thanks to Proposition 3.3.1 and to the standard linearity property of the Shapley value, as

$$\mathrm{ShM}_j(\Gamma_s) = \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, \mathcal{S}h_j^{(s_i)}, \quad \forall\, s \in S,\ 1 \leq j \leq P. \tag{3.35}$$

In the next sections we will exploit Hoeffding's inequality [40] to derive basic confidence intervals for the Shapley value of Markovian games.

Theorem 3.3.2 (Hoeffding's inequality). *Let $A_1, \ldots, A_n$ be n independent random variables, where $A_i \in [a_i, b_i]$ almost surely. Then, for all $\epsilon > 0$,*

$$\Pr\left(\sum_{i=1}^{n} A_i - \mathbb{E}\left[\sum_{i=1}^{n} A_i\right] \geq n\,\epsilon\right) \leq\ 2\exp\left(-\frac{2n^2\,\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

In this work, several results are shown in the case of *simple Markovian games*. They are Markovian games with transferable utility in which the value of each coalition in each state can only take on binary values, i.e. 0 and 1. Simple games model winning/losing situations, in which winning coalitions have unitary value. The Shapley value applied to simple static games is commonly referred to as Shapley-Shubik power index (SS). We define SSM as the Shapley value in simple Markovian games. Of course, the relation between SS and SSM is analogous to expression (3.35).

We say that player i is *critical* for coalition $\Lambda \subseteq \mathcal{P}\backslash\{i\}$ in state s if $v^{(s)}(\Lambda \cup \{i\}) - v^{(s)}(\Lambda) = 1$.

3.3.3 Motivations of the Markovian model

Many interaction situations among different individuals are not one-shot, but continue over time. Moreover, the environment in which interactions take place is dynamic, and this may influence the negotiation power of each individual. Under these assumptions, the value of each coalition varies over time. In economics, clear examples of this situation are the continuing bargaining among countries, firms, or management unions. This pragmatic reasoning spurred the research on dynamic cooperative games in the last decade (see e.g. [31], [48]). Our Markovian model is a specific instance of a dynamic cooperative game, in which the evolution of the coalition values over time follows an exogenous Markov chain on a finite state space. A concrete example of our model, in which the coalition values are not bound to be binary though, can be found in [12], where a wireless multiple access channel is considered, and several users attempt to transmit to a single receiver. The value of a coalition of users is computed as the maximum sum-rate achievable by the coalition when the remaining players threaten to jam the network. The state of the system is represented by the channel coefficients, whose evolution over time follows a Markov chain, a classic assumption in wireless communications.

Our Markovian scenario can be seen as a natural extension to a dynamic context of static situations with some uncertainty in the model. For example, let us consider games with agent failure (see e.g. [60, 70, 71]), in which each player may withdraw from the game with a certain probability. The dynamic version of this game can be modelled via a very simple Markov chain, where the probability of reaching a state where a certain subset Λ of players survives only depends on Λ and not on the current state. Interestingly, the approach utilized in [14] to approximate the Shapley value in static games has been adapted to a cooperative game with failures in [15]. In [25], a coalition formation scenario with uncertainty is considered, in which the state of the system accounts for the stochastic outcome of the collaboration among agents. Though our model does not consider coalition formation, a simple Markov chain can still be used to extend the scenario in [25] to a dynamic context, in which the transition probabilities still do not depend on the starting state.

It is also worth clarifying the meaning of the Shapley value ShM on Markovian games, defined as in (3.35). Classically, in static games the Shapley value has a two-fold interpretation. It can be thought of either as a measure of agents' power or as a binding agreement the agents make regarding the sharing of the revenue earned by the grand coalition. The first interpretation still holds in Markovian games, where $\mathrm{ShM}_j(\Gamma_s)$ is the expected power of agent j in the long-run game Γ_s. The second interpretation is sensible only when the value of the grand coalition is deterministic; since $v(\mathcal{P}, \Gamma_s)$ is an expected revenue, this second interpretation fails to hold in the Markovian game. Nevertheless, we can still view ShM under a revenue sharing perspective. Suppose indeed that the rewards at each state are deterministic. We see from (3.35) that $\mathrm{ShM}_j(\Gamma_s)$ equals the long-run expected payoff for player j, if in each state s the deterministic revenue $\mathcal{S}h_j^{(s)}$ is assigned to player j. Therefore, $\{\mathcal{S}h_j^{(s)}\}_{s\in\mathcal{S}}$ can be seen as the deterministic distribution of $\mathrm{ShM}(\Gamma_s)$ along the course of the dynamic game, for any initial state $s \in \mathcal{S}$. Moreover, it is straightforward to see that such distribution procedure is time consistent, i.e. if the state at time $t \geq 0$ is S_t, then $\beta^t \mathrm{ShM}_j(\Gamma_{S_t})$ is the long-run expected revenue for player j from time t onwards. For a detailed discussion on this topic, in a more complex model in which the transition probabilities depend on the players' actions, we refer to [11].

3.3.4 Complexity of deterministic algorithms

Since the *exact* computation of the Shapley value - or, equivalently, of the Shapley-Shubik index - involves the calculation of the incremental asset brought by a player to each coalition, then its complexity is proportional to the number of such coalitions, i.e. 2^{P-1}, under oracle access to the characteristic function. In this section we evaluate the complexity of any *deterministic* algorithm which *approximates* the Shapley-Shubik index in a simple Markovian game.
Before starting our analysis, let us introduce some ancillary concepts. We mean by *game instance* a specific Markovian game. In this paper, we implicitly assume that all the algorithms considered - deterministic or randomized - aim at approximating the Shapley value for player j, without loss of generality. Let us clarify our notion of "query".

Definition 18. *A* query *of an algorithm - deterministic or randomized - consists in the evaluation of the marginal contribution of player j to a coalition $\Lambda \subseteq \mathcal{P}\backslash\{i\}$, i.e. $v(\Lambda \cup \{i\}) - v(\Lambda)$.*

Now we define the accuracy of a deterministic algorithm.

Definition 19. *Let us assume that the Shapley-Shubik index for player j in the simple Markovian game Γ_s is $\mathrm{SSM}_j(\Gamma_s) = a$. Let ALG be a* deterministic *algorithm employing q queries. We say that ALG has an* accuracy *of at least $d > 0$ with q queries whenever, for all the game instances, ALG always answers $\mathrm{SSM}_j(\Gamma_s) \in [a - d; a + d]$.*

We will first show that an exponential number of queries is necessary in order to achieve a polynomial accuracy for any deterministic algorithm aiming to approximate the Shapley-Shubik index in the static case. This is an extension of Theorem 3 in [14] to the Shapley-Shubik index, and its proof is in Appendix 3.3.10.

Theorem 3.3.3. *Any deterministic algorithm computing one component of the Shapley-Shubik index in simple static game in state s requires $\Omega(2^P/\sqrt{P})$ queries to achieve an accuracy of at least $1/(2P)$, for all $s \in S$.*

We remark that Theorem 2 does not apply to weighted voting games, for which it is possible to exploit the weight/quota structure to achieve a lower complexity (e.g. see [20, 45, 57, 58, 74, 90]).

Finally, we are ready to derive a trade-off between the accuracy and the complexity of a deterministic algorithm approximating the Shapley-Shubik index in a simple Markovian game, as a function of the number of players P.

Corollary 3.3.4. *There exists $c > 0$ such that any deterministic algorithm approximating one component of the Shapley-Shubik index in the simple Markovian game Γ_s requires $\Omega(2^P/\sqrt{P})$ queries to achieve an accuracy of at least c/P, for all $s \in S$.*

The results of the current section clearly discourage from computing exactly or even approximating SSM with a deterministic algorithm when the number of players P is large. Motivated by this, in the next sections we will direct our attention towards *randomized* approaches to construct confidence intervals for SSM, whose complexity does not even depend on P.

3.3.5 Randomized static approach

In this section we will propose our first approach to compute a confidence interval for the Shapley value in Markovian games. The expression of the confidence interval that we will propose holds for the Shapley value of *any* Markovian game (ShM). Nevertheless, in the following sections we will provide some results holding specifically for the Shapley-Shubik index in the particular case of simple Markovian games (SSM). Let us first define our performance evaluator for a randomized algorithm.

Definition 20. *Let $1 - \delta$ be the probability of confidence. The* accuracy of a randomized algorithm *is the length of the confidence interval produced by the randomized algorithm to approximate SSM.*

In parallel, the reader learns the notion of accuracy of a deterministic algorithm from Definition 19. Throughout the paper we suppose that the transition probability matrix $\mathbf{P}$ is known by the estimator agent. In this section we also assume that the value of all coalitions in each single stage games are available *off-line* to the estimator agent.

Assumption 2. *The estimator agent has access to all the coalition values in each state:*

$$\{v^{(s)}(\Lambda),\ \forall\,\Lambda \subseteq \mathcal{P},\ s \in S\}$$

at the same time, before the Markovian game starts.

It is clear that, under Assumption 2, the estimator agent can perform an off-line randomized algorithm to approximate ShM.

Remark 11. *Assumption 2 seems to be impractical for the intrinsic dynamics of the model we consider. Nevertheless, the randomized approach based on Assumption 2 that we propose next (SCI) will prove to be an insightful performance benchmark for two methods (DCI1 and DCI2) described in Section 3.3.6, based on a more realistic dynamic assumption.*

First, let us find a formulation of the Shapley value in the Markovian game which is suitable for our purpose. Let X be the set of all the permutations of $\{1, \dots, P\}$. Let $\mathcal{C}_\chi(j)$ be the coalition of all the players whose index precedes j in the permutation $\chi \in X$, i.e.

$$\mathcal{C}_\chi(j) \equiv \{i : \chi(i) < \chi(j)\}. \tag{3.36}$$

We can write the Shapley value of the Markovian game Γ_s, both for the discount and for the average criterion, as

$$\begin{aligned}
\mathrm{ShM}_j(\Gamma_s) &= \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, \mathcal{S}h_j^{(s_i)} \\
&= \frac{1}{P!} \sum_{\chi \in X} \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s) \big[v^{(s_i)}(\mathcal{C}_\chi(j) \cup \{j\}) - v^{(s_i)}(\mathcal{C}_\chi(j))\big] \\
&= \mathbb{E}_\chi \left[\sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s) \big[v^{(s_i)}(\mathcal{C}_\chi(j) \cup \{j\}) - v^{(s_i)}(\mathcal{C}_\chi(j))\big] \right],
\end{aligned}$$

where $\mathbb{E}_\chi$ is the expectation over all the permutations $\chi \in X$, each having the same probability $1/P!$.

We now propose our first algorithm to compute a confidence interval for $\mathrm{ShM}_j(\Gamma_s)$, for each player j and initial state s. For each query, labeled by the index $k = 1, \dots, m$, let us select independently over a uniform distribution on X a permutation χ_k of $\{1,\dots,P\}$. Let us define $Z(j)$ as the random (over $\chi \in X$) variable

$$\begin{aligned} Z(j) &\equiv \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\big[v^{(s_i)}(\mathcal{C}_\chi(j)\cup\{j\}) - v^{(s_i)}(\mathcal{C}_\chi(j))\big] \\ &= v(\mathcal{C}_\chi(j)\cup\{j\},\Gamma_s) - v(\mathcal{C}_\chi(j),\Gamma_s) \end{aligned} \tag{3.37}$$

and let $Z_k(j)$ be the k-th realization of $Z(j)$. We remark that $Z(j)$ implies the computation of $|S|$ queries, one in each state. Thanks to Hoeffding's inequality, we can write that, for all $\epsilon > 0$,

$$\Pr\left(\left|\frac{1}{m}\sum_{k=1}^{m} Z_k(j) - \mathrm{ShM}_j(\Gamma_s)\right| \geq \epsilon\right) \leq \ 2\exp\left(-\frac{2m\,\epsilon^2}{[\overline{y}-\underline{y}]^2}\right)$$

where

$$\begin{aligned} \overline{y} &= \max_{\mathcal{C}\subseteq\mathcal{P}} \ \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\big[v^{(s_i)}(\mathcal{C}\cup\{j\}) - v^{(s_i)}(\mathcal{C})\big], \\ \underline{y} &= \min_{\mathcal{C}\subseteq\mathcal{P}} \ \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\big[v^{(s_i)}(\mathcal{C}\cup\{j\}) - v^{(s_i)}(\mathcal{C})\big]. \end{aligned}$$

We remark that, in the case of simple games, $\big[\overline{y}-\underline{y}\big]^2 \leq \big[\sum_{i=1}^{|S|}\boldsymbol{\sigma}_i(s)\big]^2$. Now we are ready to propose our first confidence interval, based on Assumption 2.

Static Confidence Interval 1 (SCI). *Let $1 \leq j \leq P$, $s \in S$. Fix an integer n and set $\delta \in (0;1)$. Then, with probability of confidence $1-\delta$, $\mathrm{ShM}_j(\Gamma_s)$ belongs to the confidence interval*

$$\left[\frac{1}{m}\sum_{k=1}^{m} Z_k(j) - \epsilon(m,\delta)\ ;\ \frac{1}{m}\sum_{k=1}^{n} Z_k(j) + \epsilon(m,\delta)\right],$$

where

$$\epsilon(m,\delta) = \sqrt{\frac{[\overline{y}-\underline{y}]^2\log(2/\delta)}{2m}}. \tag{3.38}$$

In the case of simple games, (3.38) becomes

$$\epsilon(m,\delta) = \sqrt{\frac{\Big[\sum_{i=1}^{|S|}\boldsymbol{\sigma}_i(s)\Big]^2\log(2/\delta)}{2m}}\,. \tag{3.39}$$

Under the average criterion, (3.39) can be written as $\epsilon(m,\delta) = \sqrt{\log(2/\delta)/[2m]}$.

Not surprisingly, the confidence interval SCI is analogous to the one found in [14] for static games. Indeed, the intrinsic dynamics of the game is surpassed by Assumption 2, for which the estimator has global knowledge of all the coalition values, even before the Markov process initiates. Therefore, from the estimator agent's point of view, there exists no conceptual difference between the approach in [14] and SCI, except for the complexity, which increases by a factor $|S|$ in the dynamic game.

3.3.6 Randomized dynamic approaches

In this section we will propose two methods to compute a confidence interval for SSM, for which Assumption 2 on global knowledge of coalition values is no longer necessary. Indeed, the reader will notice that their conception naturally arises from the assumption that the estimator agent learns the coalition values in each single stage game while the Markov chain process unfolds, as formalized below.

Assumption 3. *The state in which the estimator agent finds itself at each time step follows the same Markov chain process of the Markovian game itself. The estimator agent has local knowledge of the game that is being played, i.e. at step $t \geq 0$ the estimator agent has access only to the coalition values associated to the static game in the current state S_t.*

Remark 12. *The approaches described in this section can also be employed under Assumption 2. Indeed, any algorithm requiring the query on coalition values separately in each state can also be run under a static assumption.*

In the following we still assume that the transition probability matrix $\mathbf{P}$ is known by the estimator agent. As in Section 3.3.5, the randomized approaches that we are going to introduce hold for the Shapley value of *any* Markovian game.

First dynamic approach

We propose our first randomized approach to compute a confidence interval for ShM, holding both under the static Assumption 2 and under the dynamic Assumption 3. Let $\chi \in X$ be, as in Section 3.3.5, a random permutation uniformly distributed on the set $\{1, \dots, P\}$. Let us define $Y^{(s_i)}(j)$ as the random (over $\chi \in X$) variable associated to state s_i:

$$Y^{(s_i)}(j) \equiv \; v^{(s_i)}(\mathcal{C}_\chi(j) \cup \{j\}) - v^{(s_i)}(\mathcal{C}_\chi(j)). \tag{3.40}$$

Our dynamic approach suggests to sample the r.v. $Y^{(s_i)}(j)$ n_i times in state s_i. Let $n = \sum_{i=1}^{|S|} n_i$ be the total number of queries. We can still exploit

Hoeffding's inequality to say that, for all $\epsilon' > 0$,

$$\Pr\left(\left|\sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i(s)}{n_i} \sum_{t=1}^{n_i} Y_t^{(s_i)}(j) - \mathrm{ShM}_j(\Gamma_s)\right| \geq n\,\epsilon'\right) \leq \ \ldots$$
$$2\exp\left(-\frac{2[n\,\epsilon']^2}{\sum_{i=1}^{|S|} \boldsymbol{\sigma}_i^2(s)[\overline{x}(i) - \underline{x}(i)]^2/n_i}\right)$$

where, for all $i = 1, \ldots, |S|$,

$$\overline{x}(i) = \max_{\mathcal{C} \subseteq \mathcal{P}} \ v^{(s_i)}(\mathcal{C} \cup \{j\}) - v^{(s_i)}(\mathcal{C})$$
$$\underline{x}(i) = \min_{\mathcal{C} \subseteq \mathcal{P}} \ v^{(s_i)}(\mathcal{C} \cup \{j\}) - v^{(s_i)}(\mathcal{C})$$

We notice that, in the case of simple games, $\overline{x}(i) = 1$ and $\underline{x}(i) = 0$ for all $i = 1, \ldots, |S|$. Now set $\widetilde{\epsilon} = n\,\epsilon'$. Now we are ready to propose our second confidence interval for $\mathrm{ShM}_j(\Gamma_s)$, the first one holding under Assumption 3.

Dynamic Confidence Interval 1 (DCI1). *Let $1 \leq j \leq P$, $s \in S$. Fix the number of queries n and set $\delta \in (0;1)$. Then, with probability of confidence $1 - \delta$, $\mathrm{ShM}_j(\Gamma_s)$ belongs to the confidence interval*

$$\left[\sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i(s)}{n_i} \sum_{t=1}^{n_i} Y_t^{(s_i)}(j) - \widetilde{\epsilon}(n,\delta) \ ; \ \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i(s)}{n_i} \sum_{t=1}^{n_i} Y_t^{(s_i)}(j) + \widetilde{\epsilon}(n,\delta)\right],$$

where

$$\widetilde{\epsilon}(n,\delta) = \sqrt{\frac{\log(2/\delta)}{2} \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i^2(s)}{n_i} [\overline{x}(i) - \underline{x}(i)]^2}. \tag{3.41}$$

In the case of simple games, (3.41) becomes

$$\widetilde{\epsilon}(n,\delta) = \sqrt{\frac{\log(2/\delta)}{2} \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i^2(s)}{n_i}}\,. \tag{3.42}$$

Optimal sampling strategy In this section we focus exclusively on *simple Markovian games*. It is interesting to investigate the optimum number of times n_i^* that the variable $Y^{(s_i)}(j)$ should be sampled in each state s_i, in order to minimize the length of the confidence interval DCI1, keeping the confidence probability fixed. We notice that, by fixing $1 - \delta$, we can find the optimal values for $n_1, \ldots, n_{|S|}$ by setting up the following integer programming problem:

$$\begin{cases} \min\limits_{n_1,\ldots,n_{|S|}} \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i^2(s)[\overline{x}^2(i) - \underline{x}^2(i)]/n_i \\ \sum_{i=1}^{|S|} n_i = n, \qquad n_i \in \mathbb{N} \end{cases} \tag{3.43}$$

Remark 13. *If the static Assumption 2 holds, then the computation of the optimum values $n_1^*, \dots, n_{|S|}^*$ in (3.43) is the only information we need to maximize the accuracy of DCI1, since the sampling is done off-line. Otherwise, if Assumption 3 holds, the estimator does not know in advance the succession of states hit by the process, hence it is crucial to plan a sampling strategy of the variable $Y^{(s_i)}(j)$ along the Markov chain. Of course, a possible strategy would be, when n is fixed, to sample n_i^* times the variable $Y^{(s_i)}(j)$ only the first time the state s_i is hit, until all the states are hit. Nevertheless, this approach is clearly not efficient, since in several time steps the estimator is forced to remain idle.*

Motivated by Remark 13, now we devise an efficient and straightforward sampling strategy, consisting in sampling $Y^{(s_i)}(j)$, *each* time the state s_i is hit, an equal number of times over all $i = 1, \dots, |S|$. Let us first show a useful classical result for Markov chains. Let η be the number of steps performed by the Markov chain. Let η_i be the number of visits to state s_i, i.e.

$$\eta_i = \sum_{t=0}^{\eta-1} \mathbb{I}(S_t = s_i).$$

Theorem 3.3.5 ([8])**.** *Let $\{S_t,\ t \geq 1\}$ be an ergodic Markov chain. Let $\hat{\boldsymbol{\pi}}_i^{(\eta)} \equiv \eta_i/\eta$. Then, for any distribution on the initial state and for all $i = 1, \dots, |S|$,*

$$\hat{\boldsymbol{\pi}}_i^{(\eta)} \overset{\eta\uparrow\infty}{\longrightarrow} \boldsymbol{\pi}_i \qquad \text{with probability } 1,$$

where $\boldsymbol{\pi}$ is the stationary distribution of the Markov chain.

It is evident from (3.41) that $\widetilde{\epsilon}(n, \delta) \in \Theta(n^{-1/2})$. Now we will show under which conditions the straightforward sampling strategy described above allows to achieve asymptotically for $n \uparrow \infty$ the best rate of convergence of $\widetilde{\epsilon}(n, \delta)$, for δ fixed. The reader can find the proof of the next Theorem in Appendix 3.3.10.

Theorem 3.3.6. *Suppose that Assumption 3 holds. Let the Markov chain of the simple Markovian game be ergodic. Fix the confidence probability $1 - \delta$. Under the average criterion, if each time the state s_i is hit then the estimator agent samples the r.v. $Y^{(s_i)}(j)$ a constant number of times not depending on i (e.g. 1), then with probability 1:*

$$\sqrt{n}\,\widetilde{\epsilon}(n, \delta) \overset{n\uparrow\infty}{\longrightarrow} \inf_{n \in \mathbb{N}} \min_{\substack{n_1, \dots, n_{|S|}: \\ \sum_i n_i = n}} \sqrt{n}\,\widetilde{\epsilon}(n, \delta) = \sqrt{\frac{\log(2/\delta)}{2}}.$$

Second dynamic approach

Since Hoeffding's inequality has a very general applicability and does not refer to any particular probability distribution of the random variables at issue, it is natural to look for confidence intervals especially suited to particular instances of games. In this section we will show a third confidence interval for the Shapley value of the Markovian game Γ which is tighter *i)* the higher the confidence probability $1-\delta$ is and *ii)* the tighter the confidence intervals $[l_i; r_i]$ are. As an example, in section 3.3.6 we will show a tight confidence interval for simple Markovian games.

We still assume that the estimator agent samples the r.v. $Y^{(s_i)}(j)$ n_i times, in each state s_i. Here we suppose to know beforehand that $\mathcal{S}h_j^{(s_i)}$ lies in the confidence interval $[l_i; r_i]$ with probability of at least $1-\delta_i$. In general, the extrema l_i and r_i may depend on n_i, $\sum_{t=1}^{n_i} Y_t^{(s_i)}(j)$, and δ_i.
As in the case of DCI1, the randomized approach proposed in this section also holds both under the static Assumption 2 and under the dynamic Assumption 3. It is based on the following Lemma, whose proof is in Appendix 3.3.10.

Lemma 3.3.7. *Let $A_1, \ldots, A_k$ be k random variables such that* $\Pr(A_i \in [l_i; r_i]) \geq 1-\delta_i$. *Let $c_i \geq 0$, for $i = 1, \ldots, k$. Then,*

$$\Pr\left(\sum_{i=1}^{k} c_i\, A_i \in \left[\sum_{i=1}^{k} c_i l_i \ ;\ \sum_{i=1}^{k} c_i r_i\right]\right) \geq \prod_{i=1}^{k} [1-\delta_i]$$

The reader should keep in mind that, the smaller the single confidence levels $\delta_1, \ldots, \delta_k$ are, the tighter the lower bound on the confidence probability $\prod_{i=1}^{k}(1-\delta_i)$ is. Now we are ready to present our second dynamic approach. Let the r.v. $Y^{(s_i)}(j)$ be defined as in (3.40).

Dynamic Confidence Interval 2 (DCI2). *Set $\delta_i \in (0;1)$, for all $i = 1, \ldots, |S|$. Let*

$$\left[l^{(s_i)}\left(n_i, \sum_{t=1}^{n} Y_t^{(s_i)}(j), \delta_i\right) \ ;\ r^{(s_i)}\left(n_i, \sum_{t=1}^{n} Y_t^{(s_i)}(j), \delta_i\right)\right] \tag{3.44}$$

be the confidence interval for $\mathcal{S}h^{(s_i)}$, with probability of confidence $1-\delta_i$, for all $i = 1, \ldots, |S|$. Let $1 \leq j \leq P$, $s \in S$. Then, with probability of confidence $\prod_{i=1}^{|S|}(1-\delta_i)$, $\mathrm{ShM}_j(\Gamma_s)$ *belongs to the confidence interval*

$$\left[\sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, l^{(s_i)}\left(n_i, \sum_{t=1}^{n_i} Y_t^{(s_i)}(j), \delta_i\right) \ ;\ \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\, r^{(s_i)}\left(n_i, \sum_{t=1}^{n_i} Y_t^{(s_i)}(j), \delta_i\right)\right].$$

We notice that the confidence interval DCI2 reveals the most natural connection between the issue of computing confidence intervals of Shapley value in static games, already addressed in [14], and in Markovian games under the dynamic Assumption 3.

We already saw in Section 3.3.6 that the accuracy of DCI1 can be maximized by adjusting the number of queries $n_1, \dots, n_{|S|}$ in each state. Here, in addition, we could optimize DCI2 also over the set of confidence levels $\delta_1, \dots, \delta_{|S|}$, under the nonlinear constraint:

$$\prod_{i=1}^{|S|} [1 - \delta_i] = \ 1 - \delta \, .$$

Simple Markovian games The aim of this section is twofold. Firstly, we suggest methods to compute a confidence interval for the Shapley-Shubik index in simple static games, as a complement of the study in [14]. Secondly, we stress that such methods can be utilized to compute efficiently the confidence interval DCI2 for SSM, as it is clear from the definition of DCI2 itself. In [14], the authors derived a confidence interval for the Shapley value of a single stage game, based on Hoeffding's inequality. Nevertheless, for simple static games, a tighter confidence interval can be obtained, by applying the following approach. Let $\chi \in X$ be a random permutation of $\{1, \dots, P\}$. Let us assume that $\{\chi_k \in X\}$, $k \geq 1$, are uniform and independent. Let us define the Bernoulli variable $Y^{(s)}(j)$ as in (3.40). As pointed out in [14], we can interpret the Shapley-Shubik index $\mathrm{SS}_j^{(s)}$ as

$$\mathrm{SS}_j^{(s)} = \ \Pr\left(Y^{(s)}(j) = 1\right) .$$

Let $Y_1^{(s)}(j), \dots, Y_n^{(s)}(j)$ be independent realization of $Y^{(s)}(j)$. It is evident that

$$\sum_{k=1}^{n} Y_k^{(s)}(j) \sim \mathcal{B}(n, \mathrm{SS}_j^{(s)}),$$

where $\mathcal{B}(a, b)$ is the binomial distribution with parameters a, b. Hence, computing a confidence interval for $\mathrm{SS}_j^{(s)}$ boils down to the computation of confidence intervals of the probability of success of the Bernoulli variable $Y^{(s)}(j)$ given the proportion of successes $\sum_{k=1}^{n} Y_k^{(s)}(j)/n$, which is a well know problem in literature. Of course, this might be accomplished by using the general Hoeffding's inequality as in [14], but over the last decades some more efficient methods have been proposed, like the Chernoff bound [26], the Wilson's score interval [101], the Wald interval [97], the adjusted Wald interval [1], and the "exact" Clopper-Pearson interval [27].

3.3.7 Comparison among the proposed approaches

In this section we focus on *simple Markovian games*, and we compare the accuracy of the proposed randomized approaches. We know that, under the static Assumption 2, we are allowed to use any of the three methods presented in this article, SCI, DCI1, and DCI2, to compute a confidence interval for the Shapley-Shubik index in simple Markovian games. In fact, DCI1 and DCI2 involve independent queries over the different states, and this can also be done under Assumptions 2. Therefore, it makes sense to compare the tightness of the two confidence intervals SCI and DCI1.

Lemma 3.3.8. *Consider simple Markovian games. Let $2\epsilon(n, \delta)$ be the accuracy of SCI (see eq. (3.39)). Let $2\widetilde{\epsilon}(n, \delta)$ be the accuracy of DCI1 (see eq. 3.42). Then, for any integer n and for any confidence probability $1 - \delta$,*

$$\epsilon(n, \delta) \le \widetilde{\epsilon}(n, \delta).$$

An interested reader can find the proof of Lemma 3.3.8 in Appendix 3.3.10.

Remark 14. *The reader should not be misled by the result in Lemma 3.3.8. In fact, n being equal in the two cases, the number of queries needed for confidence interval SCI is $|S|$ times bigger than for DCI1, since each sampling of the variable $Z(j)$, defined in (3.37), requires $|S|$ queries, one per each state. The comparison between the two confidence interval would be fair only if the estimator agent knew beforehand the coalition values of the long-run game $\{v(\Lambda, \Gamma_s)\}_{s,\Lambda}$.*

According to Remark 14, we should compare the length of the confidence interval for the static case, $2\,\epsilon(n, \delta)$, with the one for the dynamic case, $2\,\widetilde{\epsilon}(|S|n, \delta)$, calculated with $|S|$ times many queries. Intriguingly, the relation between the tightness of SCI and DCI is now, for a suitable query strategy, reversed, as we show next.

Theorem 3.3.9. *In the case of simple Markovian games, for any integer n,*

$$\min_{\substack{n'_1,\dots,n'_{|S|}:\\ \sum_i n'_i=|S|n}} \widetilde{\epsilon}(|S|n, \delta) \le \epsilon(n, \delta).$$

Proof. We can write

$$\min_{\substack{n'_1,\dots,n'_{|S|}:\\ \sum_i n'_i=|S|n}} \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i^2(s)}{n'_i} \le \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i^2(s)}{\sum_{k=1}^{|S|} n'_k/|S|} = \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i^2(s)}{n} \le \frac{\left[\sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\right]^2}{n}, \tag{3.45}$$

where the last inequality holds since $\boldsymbol{\sigma}_i(s) \ge 0$. Hence, by inspection over the expressions (3.39) and (3.42), the thesis is proved. □

Theorem 3.3.9 clarifies the relation between the confidence intervals SCI and DCI1, under the condition of simple Markovian games. We highlight its significance in the next two remarks.

Remark 15. *Theorem 3.3.9 claims that the approach DCI1 is more accurate than SCI for a suitable choice of $n'_1, \dots, n'_{|S|}$, when the number of queries is equal for the two methods. In essence, this occurs because the dynamic approach allows us to tune the number of queries in the coalition values according to the weight $\boldsymbol{\sigma}_i(s)$ of each state s_i in the long-run game. Moreover, the queries on coalition values are independent among the states, hence providing more diversity to the statistics.*

Remark 16. *As we already remarked, the dynamic Assumption 3 is more pragmatic and less restrictive than the static Assumption 2. Let us now give some insights on the accuracy that can be achieved by the approaches SCI and DCI1 under Assumptions 2 and 3. The approach DCI1 can be also utilized under static Assumption 2, and in finite time DCI1 is more accurate under Assumption 2 than under Assumption 3. Indeed, for a fixed n and under the static Assumption 2, the value of $n'_1, \dots, n'_{|S|}$ in (3.45) can always be set to the optimum value, since the algorithm DCI1 is run off-line. Instead, under the dynamic Assumption 3, the sequence of states over time $S_0, S_1, S_2, \dots$ is unknown a priori by the estimator agent, hence $n'_1, \dots, n'_{|S|}$ cannot be optimized for a finite n. Hence, in finite time, the static Assumption 2 has still an edge over the dynamic Assumption 3 for the implementation of DCI1. Nevertheless, we know from Theorem 3.3.6 that, for the average criterion in ergodic Markov chains, there exists a query strategy enabling to achieve an optimum rate of convergence for DCI1's accuracy. Therefore we can conclude with the following consideration. Under the average criterion, DCI1, when employed under the dynamic Assumption 3, can be* asymptotically *as accurate as DCI1 itself and more accurate than SCI, when both these approaches are employed under the stronger static Assumption 2.*

In addition to what has just been discussed, simulations showed that, when the number of queries n and the confidence level δ are equal for the two methods, then the *effective* confidence probability for SCI is generally higher than for DCI1, i.e. the lower bound $1 - \delta$ is loose. We explain this by reminding that the centers of the confidence intervals SCI and DC1, respectively

$$\frac{1}{m}\sum_{k=1}^{m} Z_k(j) \quad , \quad \sum_{i=1}^{|S|} \frac{\boldsymbol{\sigma}_i(s)}{n_i} \sum_{t=1}^{n_i} Y_t^{(s_i)}$$

are already two estimators for $\mathrm{SSM}(\Gamma_s)$, and the former possesses a smaller variance than the second one.

About the performance of confidence interval DCI2, the simulations confirmed our intuitions. We utilized the Clopper-Pearson interval to compute a confidence interval for the Shapley-Shubik index in simple static games, and we saw that the tightness of DCI2 increases when the confidence probability approaches 1. Let $a_{2 \succ 1}$ be the percentage of simple Markovian game

$1-\delta$	$a_{2\succ 1}$ (%)
.97	100
.95	99.9
.9	87.5
.8	57.7

Table 3.2: Percentage $a_{2\succ 1}$ of cases in which the confidence interval DCI2 is narrower than confidence interval DCI1, at different confidence probabilities. The Clopper-Pearson interval is considered for DCI2.

instances, generated randomly, in which the confidence interval DCI2 is narrower than confidence interval DCI1. In Table 3.2 we show, for each value of confidence probability $1-\delta$, the values of $a_{2\succ 1}$ obtained from simulations. We see that, for $1-\delta < 0.8$, the two confidence interval have a comparable length. For $1-\delta \geq 0.8$, the confidence interval DCI2 is apparently tighter than DCI1 under these settings.

3.3.8 Complexity of confidence intervals

In Section 3.3.4 we motivated the importance of devising an algorithm that approximates SSM with a polynomial accuracy in the number of players P without the need of an exponential number of queries. In this section we show that the proposed randomized approaches SCI and DCI1 fulfill this requirement, since they only require a polynomial number of queries to reach an accuracy which is polynomial in P. Interestingly, the number of queries required by SCI and DCI1 does not even depend on the number of players P.

Proposition 3.3.10. *Fix the confidence level δ and the length of confidence interval $2\,\epsilon$. Then n queries are required to compute the confidence interval SCI, where*

$$n = \frac{\left[\overline{y} - \underline{y}\right]^2 \log(2/\delta)}{2\,\epsilon^2}.$$

Proof. The proof follows straightforward from the expression of confidence interval SCI. □

Proposition 3.3.11. *Fix the confidence level δ and the length of confidence interval $2\,\widetilde{\epsilon}$. Then, there exist values of $n_1, \dots, n_{|S|}$, with $\sum_i n_i = n$, such that n queries are required to compute the confidence interval DCI1, where*

$$n \leq \frac{|S|\,\left[\overline{y} - \underline{y}\right]^2\, \log(2/\delta)}{2\,\widetilde{\epsilon}^2}.$$

Proof. The proof follows straightforward from Theorem 3.3.9. □

From Propositions 3.3.10 and 3.3.11 we derive the following fundamental result on the complexity of SCI and DCI1.

Theorem 3.3.12. *Let $p(P)$ be a polynomial in the variable P. The number of queries required to achieve an accuracy of $1/p(P)$ is $O(p^2(P))$, for both the confidence intervals SCI and DCI1.*

Since we did not provide an explicit expression for the confidence interval DCI2, then we can not provide a result analogous to Theorem 3.3.12 for DCI2 all the same. Anyway, we notice that the expression (3.44) of confidence interval DCI2 does not depend on the number of players P. Moreover, if the Hoeffding's inequality is used to compute the confidence interval for the Shapley value in the static games, then a result similar to Theorem 3.3.12 can be derived for DCI2.

Remark 17. *Corollary 3.3.4 and Theorem 3.3.12 explain in what sense the proposed randomized approaches SCI and DCI1 are better than any deterministic approach, according to our Definitions 19 and 20 of "accuracy". E.g., in order to achieve an accuracy in the order of P^{-1}, for a number of players P sufficiently high, the number of queries needed by SCI and DCI1 is always smaller than the number of queries employed by any deterministic algorithm.*

3.3.9 Conclusions

In Section 3.3.4 we proved that an exponential number of queries is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Hence, we directed our attention to randomized algorithms and we proposed three different methods to compute a confidence interval for SSM. The first one, described in Section 3.3.5 and called SCI, assumes that the coalition values in each state are available off-line to the estimator agent. SCI can be seen as a benchmark for the performance of the other two methods, DCI1 in Sections 3.3.6 and DCI2 in Section 3.3.6. The last two methods can be utilized also if we pragmatically assume that the estimator learns the coalition values in each static game while the Markov chain process unfolds. DCI2 reveals the most natural connection between confidence intervals of Shapley value in static games, presented in [14], and in Markovian games. As a by-product of the study of DCI2, we provided confidence intervals for the Shapley-Shubik index in static games, which are tighter than the one proposed in [14]. In Section 3.3.6 we proposed a straightforward way to optimize the tightness of DCI1. In Section 3.3.7 we compared the three proposed approaches in terms of tightness of the confidence interval. We proved that DCI1 is tighter than SCI, with an equal number of queries and for a suitable choice of the number of queries on coalition values in each state. This occurs essentially because DCI1 allows us to tune the number of samples according to the weight of the state. Hence we showed that, *asymptotically*, the dynamic Assumption 3 is not restrictive with respect to the much stronger static Assumption 2, under the average criterion and for what concerns SCI and DCI1. The simulations confirmed that DCI2 is more accurate than the SCI and DCI1 when both the confidence probability is close to 1 and a tight confidence interval for the Shapley-Shubik index of static games is available, like the Clopper-Pearson interval. Finally, in Section 3.3.8 we showed that a polynomial number of queries is sufficient to achieve a polynomial accuracy for the proposed algorithms. Hence, in order to compute SSM, the proposed randomized approaches are more accurate than any deterministic approach for a number of players sufficiently high. The three proposed randomized approaches can be utilized to compute confidence intervals for the Shapley value in *any* cooperative Markovian game, too. In Table 3.3 we summarize the features of the three proposed confidence intervals, SCI, DCI1, and DCI2.

3.3.10 Appendix

Proof of Theorem 3.3.3

Proof. We will prove that there exists a class $\mathcal{F}$ of game instances for which any deterministic algorithm computing $\mathrm{SS}_j^{(s)}$ with accuracy of at least $1/(2P)$ must utilize $\Omega(2^P/\sqrt{P})$ queries. Similarly to [14], let us construct $\mathcal{F}$ when P is odd. Let $\Lambda \subseteq \mathcal{P}\backslash\{j\}$. There exists a set D_o of $\binom{P-1}{[P-1]/2}/2$ coalitions of cardinality $[P-1]/2$ such that player j is critical only for D_o. In particular, for $|\Lambda| \leq [P-1]/2$, $v^{(s)}(\Lambda) = 0$; if $|\Lambda| = [P-1]/2$, then, if $\Lambda \in D_o$, $v^{(s)}(\Lambda \cup \{j\}) = 1$, otherwise $v^{(s)}(\Lambda \cup \{j\}) = 0$. The values of the remaining coalitions are 1 if and only if they contain a winning coalition among the ones constructed so far. The Shapley value for player j is thus:

$$\mathrm{SS}_j^{(s)} = \frac{([P-1]/2)!\,([P-1]/2)!}{2(P)!}\binom{P-1}{[P-1]/2} = \frac{1}{2P}$$

Hence, for any deterministic algorithm ALG_o employing a number of queries smaller than $\mu_o(P)$, where

$$\mu_o(P) = \frac{1}{2}\binom{P-1}{[P-1]/2},$$

there always exists an instance belonging to $\mathcal{F}$ for which ALG_o would answer $\mathrm{SS}_j^{(s)} = 0$. By Stirling's approximation, we can say that $\mu_o(P) \in \Omega(2^P/\sqrt{P})$. Let us now construct the class $\mathcal{F}$ of instances when P is even and $P > 2$. Let D_e be a set of $\binom{P-2}{[P-2]/2}$ coalitions of cardinality $[P-2]/2$, belonging to $\mathcal{C}\backslash\{j\}$, such that player j is critical only for D_e. Then,

$$\mathrm{SS}_j^{(s)} = \frac{(P/2-1)!\,(P/2)!}{(P)!}\binom{P-2}{[P-2]/2} = \frac{1}{2[P-1]} > \frac{1}{2P}.$$

Similarly to before, for any deterministic algorithm ALG_e using a number of queries smaller than

$$\mu_e(P) = \binom{P-1}{[P-2]/2} - \binom{P-2}{[P-2]/2} = \frac{P-2}{P}\binom{P-2}{[P-2]/2},$$

there always exists an instance belonging to $\mathcal{F}$ for which ALG_e would answer $\mathrm{SS}_j^{(s)} = 0$. By Stirling approximation, we can say that $\mu_e(P) \in \Omega(2^P/\sqrt{P})$. Hence, a number of samples $\mu \in \Omega(2^P/\sqrt{P})$ is needed to achieve an accuracy of at least $1/(2P)$. Hence, the thesis is proved. □

Proof of Corollary 3.3.4

Proof. Any deterministic algorithm employs a certain number of queries in each state s in order to compute $\mathrm{SSM}_j(\Gamma_s) = \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\mathrm{SS}_j^{(s_i)}$. Let I_0 be a game instance in which player j is a dummy player in all the single stage games $\{v^{(s)}\}_{s\in S}$, i.e. $\mathrm{SS}_j^{(s)} = 0$ for all $s \in S$. Let I_1 be a game instance such that $\mathrm{SS}_j^{(s)} = 0$ for all s except for s_k, for which $\sigma(s_k) \neq 0$, and such that the game $\Psi^{(s_k)}$ belongs to the class $\mathcal{F}$ of instances described in the proof of Theorem 3.3.3. Therefore,

$$\mathrm{SSM}_j(\Gamma_s) = \frac{\boldsymbol{\sigma}_k(s)}{2P}$$

in the case that P is odd and

$$\mathrm{SSM}_j(\Gamma_s) = \frac{\boldsymbol{\sigma}_k(s)}{2[P-1]}$$

if P is even. Hence, any deterministic algorithm needs $\Omega(2^P/\sqrt{P})$ queries in state s_k to achieve an accuracy better than $\boldsymbol{\sigma}_k(s)/(2P)$. Set $c = \boldsymbol{\sigma}_k(s)/2$. Hence, the thesis is proved. □

Proof of Lemma 3.3.7

Proof. We will provide the proof for continuous random variables; the proof for the discrete case is totally similar. By induction, it is sufficient to prove that, if $\Pr(A_1 \in [l_1; r_1]) \geq 1 - \delta_1$ and $\Pr(A_2 \in [l_2; r_2]) \geq 1 - \delta_2$, then

$$\Pr\left(A_1 + A_2 \in [l_1 + l_2 \ ; \ r_1; r_2]\right) \geq \ (1-\delta_1)(1-\delta_2).$$

Let f_A be the probability density function of the r.v. A. Let $\overline{f}_{A_i}(x) = f_{A_i}(x)\mathbb{I}(x \in [l_i; r_i])$, $i = 1, 2$. Then,

$$\begin{aligned}
\Pr\left(A_1 + A_2 \in [l_1 + l_2; r_1; r_2]\right) &= \int_{l_1+l_2}^{r_1+r_2} f_{A_1+A_2}(x)dx \\
&= \int_{l_1+l_2}^{r_1+r_2} \int_{\mathbb{R}} f_{A_1}(x-\tau)\, f_{A_2}(\tau)\, d\tau\, dx \\
&\geq \int_{l_1+l_2}^{r_1+r_2} \int_{\mathbb{R}} \overline{f}_{A_1}(x-\tau)\, \overline{f}_{A_2}(\tau)\, d\tau\, dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \overline{f}_{A_1}(x-\tau)\, \overline{f}_{A_2}(\tau)\, d\tau\, dx \\
&= \int_{\mathbb{R}} \overline{f}_{A_1}(x)\, dx \int_{\mathbb{R}} \overline{f}_{A_2}(x)\, dx \\
&= \Pr(A_1 \in [l_1; r_1]) \Pr(A_2 \in [l_2; r_2]) \\
&\geq (1-\delta_1)(1-\delta_2).
\end{aligned}$$

Hence, the thesis is proved. □

Proof of Theorem 3.3.6

Proof. Let us consider the following constrained minimization problem over the reals:

$$\begin{cases} \min\limits_{\omega_1,\ldots,\omega_{|S|}} \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i^2(s)/\omega_i \\ \sum_{i=1}^{|S|} \omega_i = n, \qquad \omega_i \in \mathbb{R}. \end{cases} \tag{3.46}$$

By using, e.g., the Lagrangian multiplier technique, it is easy to see that the optimum value for ω_i is

$$\omega_i^* = \frac{\boldsymbol{\sigma}_i(s)\, n}{\sum_{k=1}^{|S|} \boldsymbol{\sigma}_k(s)}$$

and that the minimum value of the objective function is

$$\xi^* = \frac{\left[\sum_{i=1}^{|S|} \boldsymbol{\sigma}_i(s)\right]^2}{n}. \tag{3.47}$$

The value ξ^* clearly represents a lower bound for the optimization problem over the integers in the case of simple games. Since we deal with the average criterion, let $\boldsymbol{\sigma}_i(s) \equiv \boldsymbol{\pi}_i$. Now we can find a lower bound for $\sqrt{n}\,\widetilde{\epsilon}(n,\delta)$ over

n that does not depend on the number of queries n:

$$\begin{aligned}
&\inf_{n\in\mathbb{N}} \min_{\substack{n_1,\dots,n_{|S|}:\\ \sum_i n_i = n}} \sqrt{n}\,\widetilde{\epsilon}(n,\delta) = \\
&= \inf_{n\in\mathbb{N}} \min_{\substack{n_1,\dots,n_{|S|}\in\mathbb{N}:\\ \sum_i n_i = n}} \sqrt{\frac{n\,\log(2/\delta)}{2}\sum_{i=1}^{|S|}\frac{\boldsymbol{\pi}_i^2}{n_i}} \\
&= \inf_{\substack{q_1,\dots,q_{|S|}\in\mathbb{Q}^+:\\ \sum_i q_i = 1}} \sqrt{\frac{\log(2/\delta)}{2}\sum_{i=1}^{|S|}\frac{\boldsymbol{\pi}_i^2}{q_i}} \\
&= \min_{\substack{x_1,\dots,x_{|S|}\in\mathbb{R}^+:\\ \sum_i x_i = 1}} \sqrt{\frac{\log(2/\delta)}{2}\sum_{i=1}^{|S|}\frac{\boldsymbol{\pi}_i^2}{x_i}} \qquad (3.48)\\
&= \sqrt{\frac{\log(2/\delta)}{2}}
\end{aligned}$$

and the optimum value of x_i in (3.48) is

$$x_i^* = \frac{\boldsymbol{\pi}_i}{\sum_{k=1}^{|S|}\boldsymbol{\pi}_k} = \boldsymbol{\pi}_i\,.$$

For Theorem 3.3.5,

$$n_i/n \overset{n\uparrow\infty}{\longrightarrow} \boldsymbol{\pi}_i \qquad \text{with probability 1.}$$

Hence, n_i/n converges with probability 1 to the optimum value x_i^* and, by continuity, the thesis is proved. □

Proof of Lemma 3.3.8

Proof. In the case of simple Markovian games, the optimization problem (3.43) turns into

$$\begin{cases} \min\limits_{n_1,\dots,n_{|S|}} \sum_{i=1}^{|S|} \boldsymbol{\sigma}_i^2(s)/n_i \\ \sum_{i=1}^{|S|} n_i = n, \qquad n_i \in \mathbb{N}. \end{cases} \qquad (3.49)$$

Let us consider the constrained minimization problem over the reals in (3.46). Since evidently ξ^*, defined in (3.47), is not greater than the minimum value of the objective function in (3.49), then by straightforward inspection over the expressions (3.39) and (3.42) the thesis is proved. □

SCI	Confidence interval based on Hoeffding's inequality. Valid under static Assumption 2. Its general formulation in (3.38) holds for Shapley value in any Markovian game, as well. Polynomial number of queries required to obtain a polynomial accuracy (Theorem 3.3.12).
DCI1	Confidence interval based on Hoeffding's inequality. Valid under both static Assumption 2 and dynamic Assumption 3. Its general formulation in (3.41) holds for Shapley value in any Markovian game, as well. Theorem 3.3.6 provides a sampling strategy maximizing its accuracy, applicable under dynamic Assumption 3. Polynomial number of queries required to obtain a polynomial accuracy (Theorem 3.3.12).
DCI2	Confidence interval valid under both static Assumption 2 and dynamic Assumption 3. Its formulation holds for Shapley value in any Markovian game, as well.
SCI vs. **DCI1**	Under the static Assumption 2, there exists a sampling strategy for which DCI1 is at least as tight as SCI, for any number of sampling n (see Theorem 3.3.9). Under dynamic Assumption 3 and average criterion, DCI1 can be made at least as tight as SCI asymptotically, for $n \uparrow \infty$ (see Theorem 3.3.6).
DCI1 vs. **DCI2**	By utilizing Clopper-Pearson intervals, DCI2 is tighter than DCI1 for all $1 - \delta > 0.8$, simulations suggest (see Table 3.2).

Table 3.3: Summary of results for the three proposed approaches to compute confidence intervals for Shapley-Shubik power index in Markovian games, i.e. SCI, DCI1, and DCI2.

Chapter 4

Restless Bandits for Dynamic Channel Section: An MDP formulation

In this chapter we utilize an MDP formulation with uncountable state space to model a Restless Multi-Armed Bandit problem. We deal with a multi-access wireless network in which transmitters dynamically select a frequency band to communicate on. The slow fading channel attenuations follow an autoregressive model. In the single user case, we formulate this selection problem as a restless multi-armed bandit problem and we propose two strategies to dynamically select a band at each time slot. Our objective is to maximize the SNR in the long run. Each of these strategies is close to the optimal strategy in different regimes. In the general case with several users, we formulate the problem as a Competitive MDP with uncountable state space, where the objective is the SINR. Then we propose two strategies to approximate the best response policy for one user when the other users' strategy is fixed.
We remark that our approach can be applied to any Restless Multi-Armed Bandit with autoregressive arms of order 1.

4.0.11 Introduction

Next generation of wireless networks is expected to be characterized by a high decentralization/distribution of control functions among nodes to support self-organizing and self-healing capabilities. Network devices shall be able to monitor and sense the surroundings, learn from their monitoring and smartly and dynamically allocate resources. This perspective scenario is attracting a considerable amount of research efforts to develop learning techniques able to optimize the trade-off between exploration and exploitation of environment and resources. A relevant class of learning algorithms is the Multi-Armed Bandit (MAB) one. In the classic MAB problem there exist several "arms" that offer a reward when pulled (in analogy with gambling on bandits in casinos). Each arm is associated with a Markov process, and the reward of an arm is a function of its state. Gittins provided a dynamic allocation procedure (see Gittins et al., 1989 [37]), then dubbed Gittins index, which is optimal if the arms that are not pulled do not evolve over time. The more general case when the arms that are not pulled keep evolving in time is known as Restless MAB. It was proven by Papadimitriou and Tsitsiklis (1999, [67]) that restless MAB are PSPACE-hard in general. Whittle (1989, [100]) proposed to adopt a heuristic Lagrangian relaxation to extend the Gittins index to the restless case, which is asymptotically optimal under certain limiting regime (Weber and Weiss, 1990 [98]).

In this work, we consider a wireless network where transmitters can select a frequency band from a shared pool to communicate on. The evolution of the slow fading channel attenuation associated to each frequency band and each transmitter is a random process that can be well approximated by an autoregressive process (Aguero et al., [2]). We assume that all such random processes are independent of each other. The goal of each transmitter is to maximize its average Signal to Interference and Noise Ratio (SINR) in the long run.

To get insight into this problem, first we focus on a single transmitter system to investigate the exploration-exploitation trade-off for the randomness introduced in the system by the autoregressive channel attenuations. Then, we consider the multi-transmitter case where the problem is further complicated by the randomness introduced by the autonomous band selections of multiple transmitters. For the single terminal case, the problem of dynamic frequency allocation for SNR maximization can be modeled as a restless Multi Armed Bandit (MAB) since the transmitter only knows the instantaneous attenuations on the bands utilized in the past and they evolve also when not utilized. To the best of the authors' knowledge, there are no available results on the MAB problem for autoregressive processes. We propose two heuristic frequency allocation strategies, one called "myopic" and the other "randomized". When the AR processes possess similar autocorrelation functions, we suggest to use the myopic strategy. Instead, when there is one AR process having a much higher autocorrelation, we suggest to use the randomized strategy. In the scenario with multiple transmitters the problem is formulated as a Competitive MDP with uncountable state space. We focus on a two-user system and we assume that user 1 is oblivious of the presence of user 2 and follows a plain single-user myopic approach. Then we propose two strategies for user 2 to approximate its best response against user 1's

strategy. Again, one strategy is myopic and the other is randomized, with respect to the SINR objective function.
A lexical remark. We say that we "sample" a frequency band when we utilize it for the communication in a certain time slot.

4.0.12 Model

In Section 4.0.13 we consider one transmitters and one receiver, while in Section 4.0.14 we deal with a model with two transmitters. Time is divided into slots and, at the beginning of a time slot, each transmitter (or user) selects a frequency band, out of a pool of M different ones, to transmit. At the receiver, a single-user decoder per transmitter is deployed. In the two-transmitter case, when a both users access the same frequency band i at time slot t, they interfere with each other, and the SINR (Signal to Interference plus Noise Ratio) for user $j = 1, 2$ at time t is

$$\text{SINR}_{i,j}[t] = \frac{P_j |h_{i,j}[t]|^2}{N_0 + \sum_{q \neq j} P_q |h_{i,q}[t]|^2}$$

where P_j is the transmit power of user j, $h_{i,j}[t] \in \mathbb{C}$ is the i-th channel coefficient of user j at time t and N_0 is the variance of the additive white Gaussian noise at the receiver. When only one user is present, the SINR definition boils down to the classic SNR. For simplicity of notation, henceforth we will denote the channel attenuation coefficient $|h_{i,j}[t]|^2$ as $g_{i,j}[t]$.
Let us describe now our channel model. In (Aguero et al., 2007 [2]) it is shown that, under slow fading conditions, the SNR (Signal to Noise Ratio) of indoor wireless channels can be well approximated by an autoregressive (AR) model. This means that, under such conditions, we can model the channel attenuations as

$$g_{i,j}[t] = \sum_{k=1}^{p_{i,j}} a_{i,j}^{(k)} \, g_{i,j}[t-k] + c_{i,j} + \epsilon_{i,j}[t]$$

where $a_{i,j} \in \mathbb{R}$, $\{\epsilon_{i,j}[t]\}_t$ is an *i.i.d.* Gaussian process with zero mean and variance $\sigma_{i,j}^2$, $c_{i,j} > 0$, and $p_{i,j}$ is the order of the model. Moreover, all the channels considered are independent of each other, i.e. $\epsilon_{i_1,j_1}[t]$ is independent of $\epsilon_{i_2,j_2}[t]$ when either $i_1 \neq i_2$ or $j_1 \neq j_2$.
We assume the AR process to be wide sense stationary (WSS), i.e. the roots of the polynomial $z^p - \sum_{k=1}^{p} a_{i,j}^{(k)} z^{p-k}$ must lie inside the unit circle.

4.0.13 Single user: MDP formulation

In this section we consider the single user case. In order to simplify the notation, we drop the user index. In our study we consider an AR(1) channel attenuation model, i.e.

$$g_i[t] = \; a_i \, g_i[t-1] + c_i + \epsilon_i[t]$$

For $|a_i| < 1$, the process is WSS, and the (unconditioned) expected value of channel attenuation $g_i[t]$ at any time instant t can be expressed as

$$m_i = \mathbb{E}(g_i[t]) = \frac{c_i}{1 - a_i} \qquad \forall\, t.$$

Therefore we can say that $\mathbb{E}(\mathrm{SNR}_i[t]) = Pm_i/N_0$, for all t. Straightforward computations show that the autocovariance function of the channel attenuation can be written as

$$\mathbb{E}\big((g_i[t] - m_i)(g_i[t-n] - m_i)\big) = a_i^{|n|}\, \frac{\sigma_i^2}{1 - a_i^2}. \tag{4.1}$$

We now illustrate the two fundamental assumptions of this paper. First, the coefficients a_i and σ_i are known by the transmitter, which might have estimated them during a training phase. Second, the transmitter, at time t, *only knows the instantaneous attenuations of the frequency bands utilized* up to time $t-1$. Indeed, we assume that the receiver estimates $g_{i,j}$ and broadcast this information on the channel along with an identifier for the transmitter and the frequency band. The *goal* of the user is to dynamically switch among the channels at each time slot in order to maximize the expected average SNR over an infinite horizon. Equivalently, it wants to maximize the expected average over time of channel attenuations, denoted by $\Phi^{(av)}(\mathbf{f})$:

$$\max_{\mathbf{f}} \left\{ \Phi^{(av)}(\mathbf{f}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\mathbf{f}}(g_{\mathbf{f}(t)}[t]) \right\} \tag{4.2}$$

where $\mathbf{f}$ is a dynamic sampling strategies over the channels $1, \ldots, M$. The reader should notice that a channel sampling strategy $\mathbf{f}$ at time t may depend on the whole history of the observed channels and of the sampling decisions up to time t. This class also includes static strategies, that choose one channel once for all. Intuitively, when there exists a channel i with much lower *unconditioned* expected attenuation, i.e. $m_i \gg m_k$ for all $k \neq i$, a static selection of the channel i is the nearly optimal strategy, since with high probability $g_i[t] > g_k[t]$ for all $k \neq i$ for almost all t.

In this section, we want to study how to dynamically select the band on which to transmit when, *a priori*, all of them are nearly equivalent, i.e. there exists $m \approx m_k$, for all k. At each time slot, there is always one channel better than the others, hence we wish to track dynamically the evolution over time of the best channel.

Intuitively, the sampling choice at each instant has to be a trade-off between exploration and exploitation. To give a hint, the most natural policy, that we will call *myopic*, at each time step t aims at maximizing the expected value of SNR$[t]$, given all the previous channel observations. On the other hand, the statical information about channels that are not used becomes more and more obsolete, therefore in some cases it might be better to explore different channels with a *randomized* strategy.

We can formulate the optimization problem (4.2) as a restless Multi Armed Bandit problem (MAB for short), in which a user at each time instant t selects an arm (here, frequency band) which gives a reward (here, the SNR) and all the arms, including the ones that have not been selected, evolve according to a certain stochastic process (here, an autoregressive process).

More specifically, we can describe the decision problem at hand as a Markov Decision Process (MDP) with an uncountable set of states $\mathcal{S}$ or, equivalently, as a Partially Observable MDP. Let us describe it in detail. At time t, we call $n_i(t)$ the number of steps ago in which channel i has been last used. The attenuation of channel i at time t conditioned on its last observation is a Gaussian r.v., and we denote its mean and variance as $\mu_i(t)$ and $\nu_i(t)$, respectively:

$$\begin{aligned}\mu_i(t) &= \mathbb{E}\big(g_i[t]\,\big|\,g_i[t-n_i(t)]\big)\\ &= a_i^{n_i(t)} g_i[t-n_i(t)] + c_i\frac{1-a_i^{n_i(t)}}{1-a_i} \end{aligned}\tag{4.3}$$

$$\nu_i(t) = \ \mathrm{Var}\left(g_i[t]\,\big|\,g_i[t-n_i(t)]\right) = \ \sigma_i^2\,\frac{1-a_i^{2n_i(t)}}{1-a_i^2}\tag{4.4}$$

where $g_i[t-n_i(t)]$ is the attenuation of channel i during its last utilization. At time step t, thanks to the Markov property of the AR(1) process, the whole statistical information about channel i is hence contained in $(\mu_i(t), \nu_i(t))$. We observe that $\mu_i \in \mathbb{R}$, while ν_i is bounded between $[\sigma_i^2; \sigma_i^2/(1-a_i^2)]$. The decision on which channel to utilize at time t hinges on the set S_t:

$$S_t = \{\mu_1(t), \nu_1(t),\ \mu_2(t), \nu_2(t), \dots\ ,\ \mu_M(t), \nu_M(t)\}.\tag{4.5}$$

By utilizing the MDP jargon, we call by S_t the state of the decision problem at time t. The state space $\mathcal{S}$ is the uncountable collection of all the possible states. In each state $S \in \mathcal{S}$, a set of actions $\mathcal{A} = \{1, 2, \dots, M\}$ is available to the transmitter, which represents the collection of channels that can be selected at time slot t. If channel i is selected, then we map the "reward" for the user in state S_t to the expected channel attenuation at time t conditioned on the last observation of channel i itself, i.e. $\mu_i(t)$. The state of the system at time $t+1$ evolves stochastically, according to the following Markovian rule. If channel i is selected at time t, then at time $t+1$,

$$\begin{aligned}&\mu_i(t+1) = a_i Y + c_i, \quad \text{where } Y \sim \mathcal{N}\big(\mu_i(t), \nu_i(t)\big)\\ &\nu_i(t+1) = \sigma_i^2.\end{aligned}$$

Instead if channel i is not selected at time t,

$$\begin{aligned}\mu_i(t+1) &= a_i\mu_i(t) + c_i\\ \nu_i(t+1) &= a_i^2\nu_i(t) + \sigma_i^2.\end{aligned}$$

Heuristic algorithms

The theory of MDP allows us to claim that there exists an optimal stationary strategy $\mathbf{f}^O$ for the problem (4.2). Unfortunately, the computation of $\mathbf{f}^O$ turns out to be a difficult task. Indeed the solution to a Markov Decision Problem with uncountable state can only be approximated by means of discretization algorithms (Bäuerle and Rieder, 2011 [17], see also Section 1.3.2), and even in this case the curse of dimensionality entails that the size of the discretized state space increases exponentially with the number of arms. A different approach would be to compute the Whittle index (Whittle, 1989 [100]) of each channel, but this approach is not guaranteed to be optimal. Hence, it becomes crucial to devise a simple policy whose performance is reasonably close to the optimal $\Phi^{(av)}(\mathbf{f}^O)$.

In the following we propose the most natural stationary strategy one can think of, i.e. the myopic policy $\mathbf{f}^M$ that aims at maximizing the instantaneous expected SNR in each state. Such a policy does not take into account that the statistics of the channel that have not been selected for a long period might become too stale. First, we need to initialize the algorithm, and we choose to sample the coefficient of each channel once.

Algorithm 4.0.13. Myopic policy $\mathbf{f}^M$.
For $0 \leq t \leq M-1$ select channel t, i.e. $\mathbf{f}^M(S_t) = t+1$. For $t \geq M$,

$$\mathbf{f}^M(S_t) = \underset{i\in\{1,\dots,M\}}{\operatorname{argmax}} \mu_i(t).$$

We intend to compare the performance of the myopic policy with a more sophisticated one, that we call randomized strategy and is inspired by the Thompson sampling strategy for Bayesian Multi Armed Bandit problems (Thompson, 1933 [92]). We suggests to draw, in each state S_t, one realization of the random variable $\xi_i = g_i[t]\big|g_i[t-n_i(t)]$, for each channel $i = 1, \dots, M$. Then, the arm corresponding to the highest realization of ξ is chosen. This procedure does not always follow the myopic rule, but with a certain probability explore the arms that, though possessing a lower μ, might be optimal since their last observation is too stale.

Algorithm 4.0.14. Randomized policy $\mathbf{f}^R$.
For $0 \leq t \leq M-1$ select channel t, i.e. $\mathbf{f}^R(S_t) = t+1$. For $t \geq M$, draw a realization of the Gaussian variable $\xi_i \sim \mathcal{N}(\mu_i(t), \nu_i(i))$ for all $i = 1, \dots, M$. Select

$$\mathbf{f}^R(S_t) = \underset{i=1,\dots,M}{\operatorname{argmax}} \xi_i.$$

Simulations

In this section we show the results of some simulations, giving a hint about the performance of the myopic and the randomized policies, described respectively in Algorithm 4.0.13 and 4.0.14. Given a stationary policy $\mathbf{f}$, we want to assess its average reward $\Phi^{(av)}(\mathbf{f})$. We compare the myopic and randomized policies with *i)* the optimal policy $\mathbf{f}^O$, approximated by means of a state discretization technique (see Section 1.3.2), with *ii)* the upper bound for the performance of any strategy, computed by selecting the channel with the highest coefficient g at each time step:

$$\mathbf{f}^U(t) = \operatorname*{argmax}_{i=1,\dots,M} g_i[t], \quad \forall\, t \geq 0 \tag{4.6}$$

and with *iii)* the static policy $\mathbf{f}^S$, that selects off-line the arm with the highest expected value, and no longer switches to other channels, i.e.

$$\mathbf{f}^S_t = \operatorname*{argmax}_{i=1,\dots,M} m_i, \quad \forall\, t \geq 0.$$

Of course, the strategy $\mathbf{f}^U$ is not applicable, since it is not causal. In theory, its performance is achievable only when the channels are deterministic hence perfectly predictable, i.e. $\sigma_i = 0$ for all $i = 1, \dots, M$. We now show the performance of the five policies under scrutiny, the myopic $\mathbf{f}^M$, the randomized $\mathbf{f}^R$, the static $\mathbf{f}^S$, the optimal $\mathbf{f}^O$, and the upper bound policy $\mathbf{f}^U$, under different channel conditions.

First, we consider 3 arms, where arms 2,3 are statistically equivalent, and $a_2 = a_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, and $m_2 = m_3 = 8$. Arm 1 has the same coefficients $a_1 = 0.3$, $\sigma_1^2 = 1$ as arms 2,3. In Figure 4.1 we show the performance of the five policies when m_1 varies within $[7; 9]$. We see that, under these conditions, the myopic policy outperforms the randomized one since the latter wastes too much time in exploring arms that are not optimal. As intuition confirms, the static policy $\mathbf{f}^S$ performs as well as the myopic $\mathbf{f}^M$ when arm 1 has the highest expected value $m_1 > m_2 = m = 3$. Instead, for $m_1 < m_2 = m_3$, dynamically switching between the arms 2,3 is beneficial with respect to statically selecting one of the two.

As we see in Figure 4.1, when all the arms are characterized by the same unconditioned expectation, i.e. $m_i = 8$, for $i = 1, 2, 3$, the static policy $\mathbf{f}^S$ is outperformed by both the myopic and the randomized strategies. It is indeed better to switch among the channels to attempt to track the best instantaneous channel at *each* time instant, based on the previous observations. Remarkably, the performance of the myopic policy $\mathbf{f}^M$ is close to the optimal $\mathbf{f}^O$.

Hence, we evaluate our algorithms in a different scenario, in which the value m's are the same for all the channels, but there exists one channel (say, 1) whose autocovariance function (4.1) decays considerably more slowly than the others. It is clear from Figure 4.2 that there are lapses in which channel 1 is by far the best, and some others in which its channel coefficient g_1 plummets below the others. From Figure 4.2 we observe that the myopic strategy often fails to track channel 1 when it is the best. The reason is quite intuitive: during the lapse in which channel 1 is the worst one, the

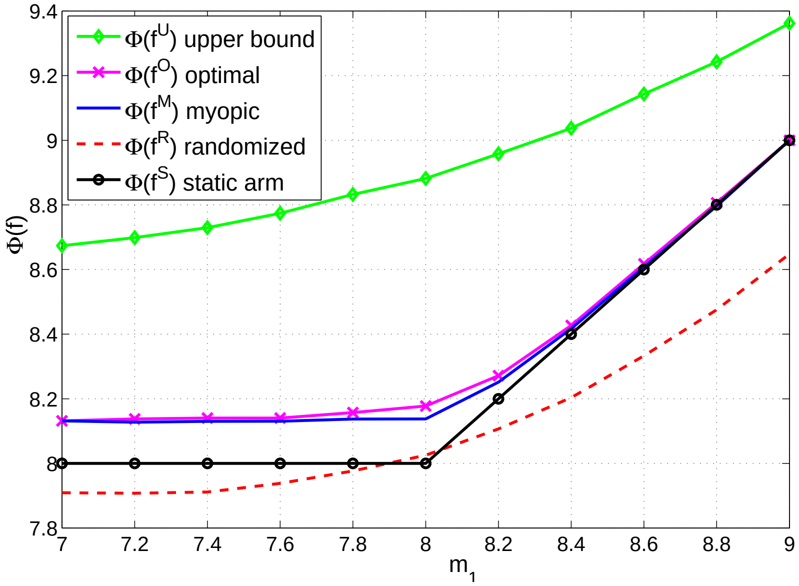

Figure 4.1: Performance of myopic and randomized algorithm with 3 arms (channels). Arms 2 and 3 are statistically equivalents, with $a_2 = a_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, and $m_2 = m_3 = 8$. Arm 1 has the same $a_1 = 0.3$, $\sigma_1^2 = 1$ as arms 2,3, while the performance of the proposed algorithms are assessed when m_1 varies within $[7; 9]$.

myopic strategy does not choose it, then its last observation become obsolete, and consequently the prediction $\mu_1(t)$ tends to $m_1 = 10$. Thus, it is highly probable (and this probability increases with M) that one of the other, suboptimal, channels, having a fresher observation, offers a higher prediction. It easily follows that, for its inherent features, the randomized policy is more suitable to such kind of situations, as results in Figure 4.3 confirm. We considered 3 arms (frequency bands). Arms 2 and 3 are statistically equivalents, with $a_2 = a_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, $c_2 = c_3 = 10$. Arm 1 has the same coefficients $c_1 = 10$, $\sigma_1^2 = 1$ as arms 2,3, while the performance of the proposed algorithms are assessed when the coefficient a_1 varies within $[0.3; 0.98]$. As we intuitively explained before, when the coefficient a_1 is sufficiently high, i.e. $a > 0.85$, the randomized strategy outperforms the myopic one. Notably, the myopic policy is quasi-optimal for $a_1 < 0.6$, while the the randomized one is nearly optimal for $a_1 > 0.9$.

4.0.14 Multi user: Competitive MDP formulation

In this section we discuss the more general scenario described in Section 4.0.12, in which two transmitters dynamically select one among M channels at each time slot. If some users choose the same channel in one time

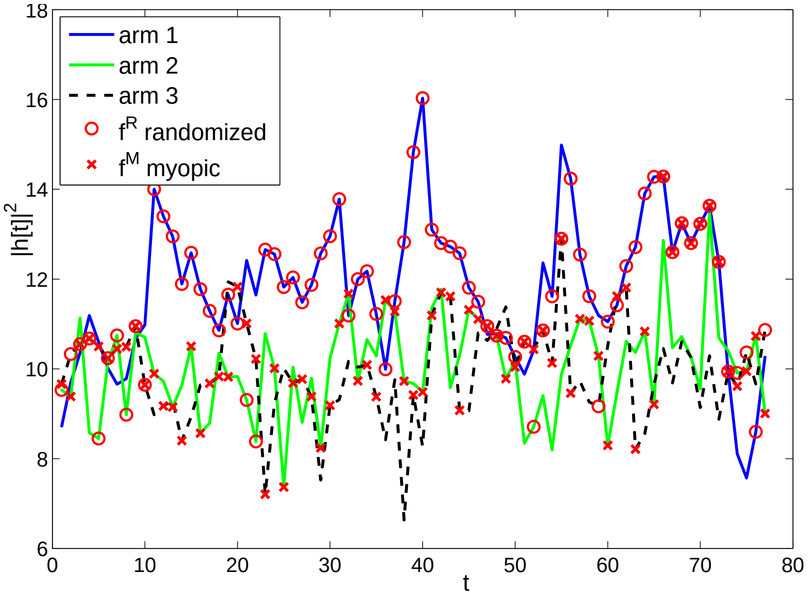

Figure 4.2: Channel (or arm) selection when $c_i = 10$ for all i, $a_1 = 0.9, a_2 = a_3 = 0.3$, $\sigma_1^2 = 1.5, \sigma_2^2 = \sigma_3^2 = 0.5$. The randomized strategy succeeds in tracking the first channel with higher autocorrelation, when it is the best one.

slot, they interfere with each other. Therefore, the objective function for each user is its SINR, and no longer its SNR. Since in the single user case the decision process can be described as an MDP, then the scenario with two users can be formalized as a Competitive MDP, also called competitive MDP (Filar and Vrieze, 1997 [32]), with uncountable state space.

In our case, the set of channels $h_{1,j}, \dots, h_{M,j}$ for player j evolve independently from the ones available to any other player $k \neq j$, and the action space for each player is still $\mathcal{A} = \{1, \dots, M\}$, i.e. the channel indices to be selected at each slot. Therefore, we are allowed to formulate the game as a Competitive MDP in which each user j controls its own Markov chain on the state space $\mathcal{S}_j$. As in the single user case $\mathcal{S}_j$ is the set of all the possible states (4.5). Formally, the state space of the Competitive MDP at hand is the Cartesian product $\mathcal{S}^* = \mathcal{S}_1 \times \mathcal{S}_2$.

Let us denote by $\mathbf{f}_j$ a sampling strategy for user j and by $\mathbf{f}_{-j}$ the one for the other users. Possibly, $\mathbf{f}_j, \mathbf{f}_{-j}$ are randomized policies. We define the instantaneous reward for user j in state $S_t^* \in \mathcal{S}^*$ as the expected reward

$$\mathbb{E}\big(\mathrm{SINR}_j[t] \big| S_t^*, \mathbf{f}_i, \mathbf{f}_{-i}\big).$$

Thus, the interaction on the players occurs only on the instantaneous rewards gained in each state, through the SINR expression. Thus we can say that our model is a reward-coupled Competitive MDP. This model is very similar with the one dealt with in (Altman at al., 2008 [4]), except that here the state space is uncountable and there are no constraints on the rewards.

Heuristic Best Response

We now propose a heuristic best response policy for user 2. Suppose that user 1 is oblivious of the presence of user 2 and performs a myopic policy $\mathbf{f}_1^M$ to maximize the expected average of channel attenuations over time, as in the single user case. On the other hand, user 2 knows the parameters of the channels, the current state, and the strategy of user 1. Thus, user 2 still faces an MDP with uncountable states, which is equivalent to the Competitive MDP described before, when user 1 fixes its own stationary strategy. Let us give an insight on a possible strategy for user 2. Assume that, for user 2, channel i_1 presents at time t the highest coefficient $g_{i_1,2}[t]$, but the expected SINR guaranteed by channel i_2 with suboptimal attenuation is higher, since the interference is much weaker. Then, it is in general not clear what user 2 should do. A myopic solution would suggest to switch to the free channel i_2, but on the other hand, in such a way the information about channel i_1 becomes stale, and moreover channel i_1 itself might become free in a near future. Then, in analogy with the single player case, we propose two strategies, one myopic and one randomized, to approximate the best response for user 2 against a myopic policy $\mathbf{f}_1^M$ that user 1 implements regardless of user 2's behaviour. We suppose the algorithms are initialized by sampling each channel once.

Algorithm 4.0.15. SINR myopic policy $\mathbf{f}_2^{MS}$ for user 2, against myopic policy $\mathbf{f}_1^M$ for user 1.

$$\mathbf{f}_2^{MS}(S_t^*, \mathbf{f}_1^M) = \underset{i \in \{1,\dots,M\}}{\operatorname{argmax}} \mathbb{E}(\mathrm{SINR}_{i,2}[t] \big| S_t^*, \mathbf{f}_1^M, i).$$

Algorithm 4.0.16. Randomized policy $\mathbf{f}_2^{RS}$ for user 2, against myopic policy $\mathbf{f}_1^M$ for user 1.
Draw a realization of the random variable $\xi_i = \mathrm{SINR}_{i,2}[t] \big| (S_t^*, \mathbf{f}_1^M, i)$*, for all* $i = 1, \dots, M$*. Select*

$$\mathbf{f}_2^{RS}(S_t^*, \mathbf{f}_1^M) = \underset{i=1,\dots,M}{\operatorname{argmax}} \xi_i.$$

About the performance of policies $\mathbf{f}^{MS}, \mathbf{f}^{RS}$, we can do similar considerations to the one made for the myopic and randomized algorithms in the single user case. Let us explain the results illustrated in Figure 4.4. We considered 2 users and 2 channels. The noise variance is $N_0 = 1$ and $P_1 = P_2 = 1$. The channels for user 1 are almost deterministic, i.e. $\sigma_{1,1}^2 = \sigma_{2,1}^2 = 0.1$ and $a_{1,1} = a_{2,1} = 0.3$, $m_{1,1} = 2$, $m_{2,1} = 0.5$. Thus user 1, that is unaware of the presence of user 2 and adopts a myopic policy $\mathbf{f}_1^M$, selects channel 1 almost always. For user 2, $\sigma_{1,2}^2 = 0.8, \sigma_{2,2}^2 = 0.4$, $m_{1,2} = 8, m_{2,2} = 3$, $a_{2,2} = 0.3$.

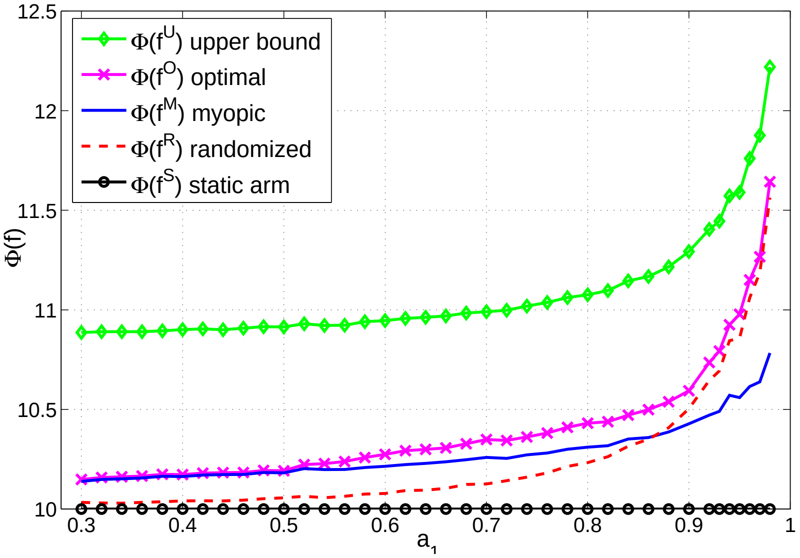

Figure 4.3: Performance of myopic and randomized algorithm with 3 arms (frequency bands). Arms 2 and 3 are statistically equivalents, with $a_2 = a_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, $c_2 = c_3 = 10$. Arm 1 has the same coefficients $c_1 = 10$, $\sigma_1^2 = 1$ as arms 2,3. a_1 varies within $[0.3; 0.98]$.

Hence, a static strategy for user 2 would suggest not to collide and to select channel 2. Anyway, sometimes it is beneficial for user 2 to select channel 1 when this is good enough. Indeed, for values of $a_{1,2}$ approaching 1, the autocorrelation of channel 1 for user 2 increases, and the randomized policy $\mathbf{f}^{RS}$ succeeds in tracking channel 1 in the time slots in which its coefficient g is large enough to overwhelm the interference caused by user 1.

4.0.15 Conclusions

We proposed two strategies to dynamically select one out of a pool of M slow fading channels, modelled as autoregressive processes of order 1. The decision process is modelled as a restless bandit, or equivalently as a Markov Decision Process. The myopic channel selection strategy is nearly optimal when the channels are similarly correlated. Instead we suggest to adopt a randomized strategy when one channel shows higher autocorrelation. When two users are present, they interfere with each other, and we model the competitive learning process as a Competitive MDP. We finally propose two ways to approximate a best response selection strategy for the transmitters.

We remark that, of course, our approach can be applied to any Restless Multi-Armed Bandit process with independent AR(1) arms. This is seemingly a promising research field. We believe indeed that AR(1) processes are

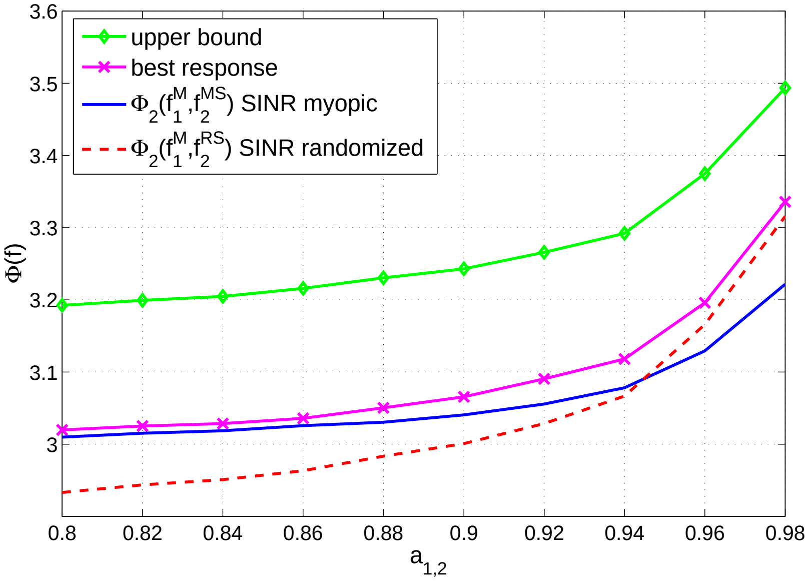

Figure 4.4: Best response strategy of user 2 against a myopic policy for user 1. For user 1, $\sigma^2_{1,1} = \sigma^2_{2,1} = 0.1$ and $a_{1,1} = a_{2,1} = 0.3$, $m_{1,1} = 2$, $m_{2,1} = 0.5$. For user 2, $\sigma^2_{1,2} = 0.8, \sigma^2_{2,2} = 0.4$, $m_{1,2} = 8, m_{2,2} = 3$, $a_{2,2} = 0.3$. $a_{1,2}$ varies within $[0.8; 0.98]$. $\Phi_2^{(av)}(\mathbf{f}_1^M, \mathbf{f}_2)$ is the expected long run average SINR for user 2 when user 1 adopts strategy $\mathbf{f}_1^M$.

easy enough to provide nice analytical formulations, and at the same time they are a widely used paradigm for correlated time-series.

List of publications

We specify the correspondence among the parts of this manuscript and the publications during the 3 years of Ph.D..

Section 2.1 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Algorithms for uniform optimal strategies in two-player zero-sum stochastic games with perfect information,” *Operations Research Letters*, vol. 40, pp. 56–60, 2011.

Section 2.4 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Stochastic games for cooperative network routing and epidemic spread”, *Communications Workshops (ICC), 2011 IEEE International Conference on*, 2011, pp. 1–5.

Section 3.1 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Cooperative Markov Decision Processes: Time Consistency, Greedy Players Satisfaction, and Cooperation Maintenance”, *International Journal of Game Theory*, Volume 42, Issue 1, pages 239-262, 2013.

Section 3.2 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Dynamic Rate Allocation in Markovian Quasi-Static Multiple Access Channels: a Game Theoretic Approach”, *11th Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2013.

Section 3.3 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Confidence intervals for Shapley value in Markovian dynamic games,” EURECOM, Tech. Rep. RR-12-264, 2012, to appear in *Dynamic Games and Applications*, DOI: 10.1007/s13235-012-0060-9.

Chapter 4 K. Avrachenkov, L. Cottatellucci, and L. Maggi, “Slow fading channel selection: A Restless Multi-Armed Bandit formulation”, *Ninth International Symposium on Wireless Communication Systems, ISWCS*, 2012.

Two publications have been left out of this manuscript, since they do not fall within the main scope of the thesis:

- L. Maggi and L. Cottatellucci, “Retrospective interference alignment for interference channels with delayed feedback,” *WCNC 2012, IEEE Wireless Communications and Networking Conference*, 2012.

- K. Avrachenkov, L. Cottatellucci, L. Maggi, and Y. Mao “Maximum entropy mixing time of circulant Markov chains,” *Statistics & Probability Letters*, Volume 83, Issue 3, March 2013, Pages 768773

Conclusions and future perspectives

In this dissertation we mainly studied Competitive and Cooperative Game Theory on Markov Decision Processes (MDPs). MDPs are controlled Markov process. The actions taken by a set of players influences both the rewards for each player in each state and the transition probabilities among the states.

In Chapter 2, Section 2.1, we investigated zero-sum Competitive MDPs with two players and perfect information, i.e. in each state at most one player has more than one action available. We considered the discounted criterion and we provided two algorithms to compute the optimal strategies at the Nash equilibrium, for both players, for all discount factors sufficiently close to 1. We adopted a linear programming technique in the field of rational functions with real coefficients. We proved the convergence in finite time for one algorithm. Our algorithms also produce the range of discount factors in which the strategies are optimal.
In Section 2.4 we utilized the techniques developed in Section 2.1 to analyze a routing game in which several service providers (SPs) share the same network and provide connection toward a unique server (destination) to their customers. In each node of the network, one SP controls the routing of the incoming packets. Each link between adjacent nodes has a different cost for each SP. SPs must cooperate to carry out the transmission of the packets successfully. We utilized a long-run cooperative game approach to distribute the costs of the transmission among the SPs. We proposed to adopt the algorithms developed in Section 2.4 to compute the value of each coalition of SPs, by following a max-min methodology.

In Chapter 3 we dealt with Dynamic Cooperative Games on MDPs. The main difference with respect to Section 2.4 lies in the fact that coalitions are allowed to form throughout the game, and it is necessary to allocate a payoff to each player in each state. In Section 3.1 a Transferable Utility (TU) MDP cooperative game is considered. we devised a Cooperative Payoff Distribution Procedure on MDPs (MDP-CPDP) which satisfies a time consistency property. We then studied under which conditions the MDP-CPDP contents the greedy players, having a myopic perspective of the game. Most importantly, we investigated a Cooperation Maintenance property, and we found that it is a refinement of the concept of Core on MDPs. Such property strengthen the cohesiveness of the grand coalition throughout the game. Indeed, at each time step, each coalition is always enticed to postpone the decision of withdrawing from the grand coalition, which is cohesive by induction.

In Section 3.2 we applied some concepts developed in Section 3.1 to a multiple access channel with Markovian quasi-static channel coefficient. We allocated the rate to each user in each channel state. In the corresponding MDP model, the transition probabilities among the (channel) states do not depend on the players (users). Hence, under this perspective, the model is simpler than the one in Section 3.1. On the other hand, the rewards (rates) cannot be distributed in any manner in each state, but only within a feasibility (capacity) region. We first studied how to obtain a maximum sum-rate allocation both in each state and in the long-run process. Then, we found a sufficient condition ensuring the existence of an allocation which is fair throughout the game, from each intermediate step onwards. Finally, we found that the set of global optimum rates coincides with both the time consistent Core set and the Cooperation Maintaining set.

In Section 3.3 we dealt with Cooperative MDPs under the TU assumption and by considering an endogenous Markov chain on the state space. We call them Markovian games. We propose three procedures to compute a confidence interval for the Shapley value in the long-run game. One of them relies on the assumption that all coalition values in all states are known off-line, while the other two are also valid under the more realistic assumption that the coalition values are learned during the game. Afterwards we provided some results on the accuracy of the proposed methods in the specific case of simple games, i.e. the coalition values in each state take on binary values, i.e. 0 and 1. These games are appropriate to assess the power of members within a committee. We proposed these three randomized approaches to overcome complexity issues. Indeed, in order to achieve a polynomial accuracy in the number of players, an exponential number of queries on the coalition values is needed even to approximate the Shapley value with polynomial accuracy. Instead, our methods only require a polynomial number of queries to achieve a polynomial accuracy.

Finally, in Chapter 4 we utilize an MDP formulation with uncountable state space to tackle a problem of optimal dynamic selection of slow fading channels. Our approach can be generalized to any Restless Multi-Armed Bandit (MAB) problem with independent autoregressive AR(1) arms. Only one frequency can be utilized at a time, and the instantaneous value of the associated attenuation coefficient is revealed. We propose two strategies to maximize the average SNR in the long-run. The first strategy is myopic, i.e. it prescribes to choose the channel with the lowest expected attenuation, conditioned on the previous observations. The second one is randomized and suggests to explore with a certain probability the seemingly worse channels, still based on the previous observations. Almost counter-intuitively, when one channel shows much higher correlation than the others, the randomized strategy outperforms the myopic one and it is quasi-optimal. Indeed, under the myopic policy, the highly correlated channel is rarely utilized and its statistics tend to become stale. Hence, in the periods in which it is better than the other channels, it is not detected.

Future perspectives I feel that the Cooperative MDP model, described in Section 3.1, can be extended in several directions.

First of all, in the Cooperative MDP model the probability of transition from a state s to a state s' only depends on the strategy adopted by each player, or agent, in state s, hence disregarding the fashion in which the play-off, earned by the grand coalition as a "reward" of their joint strategies, is shared among the players in state s. More formally, the stochastic kernel T of the process, governing the transition law between state at state t and $t+1$, S_t and S_{t+1} respectively, is described by the following relation:

$$S_{t+1} = T(S_t, \mathbf{f}_{\mathcal{P}}(S_t)),$$

where $\mathbf{f}_{\mathcal{P}}(S_t)$ is the strategy of the grand coalition $\mathcal{P}$. On the other hand, each state of the Markov process can be considered as a "snapshot" of the relative power of each player inside a coalition, which reasonably also depends on the previous allocations. Hence, it would be interesting to introduce the dependency of the transition probabilities among the states on the payoff allocation in each state. In order to make the model even more versatile, one could also assume that each player will consume a certain portion δ of the assigned pay-off, and invest the remainder in order to boost its power inside a coalition in the next step. More formally, one could define the stochastic kernel T of the process as

$$S_{t+1} = T(S_t, \mathbf{f}_{\mathcal{P}}(S_t), \boldsymbol{\gamma}(S_t), \boldsymbol{\delta}(S_t)),$$

where $\boldsymbol{\gamma}(S_t)$ is the CPDP, and $\boldsymbol{\delta}(S_t)$ is the investment policy for all players, in state S_t. The objective is devising a pay-off allocation in each state which enforces an agreement among all the players, which is both stable over time and social optimum. I expect that this task is considerably more challenging than the one dealt with in Section 3.1, due to the further complications brought in the model.

In my opinion, another promising research direction consists in examining the case of imperfect monitoring of players' strategies in a MDP game, that has been studied only in the case of repeated games (see Mailath and

Samuelson, 2006 [55], for a survey). The grand coalition is seen as a meta-agent, i.e. it does not simply coincide with the union of all the single agents, but it is rather an external entity whose duty is to impose the strategy that each agent should implement - but not necessarily actually does - in order to attain a social optimum solution. As a naïve example, one can identify the grand coalition with a Union of countries, and the players with the single countries belonging to it. The Union imposes restrictions on the policy of each country, which sometimes are not respected. In classic Cooperative Game Theory, the only way such a coalition has to stand out against the grand coalition is breaching the agreement and not cooperate, or even not signing the cooperation contract in the first place. Here one could study an intermediate situation, in which each unsatisfied subset of agents can "take the law into their own hands" and deviate from the strategy recommended by the grand coalition. By assuming the imperfect monitoring condition, one may further assume that the grand coalition has not a direct monitoring of the implemented strategies, but rather indirect, through the observation of the sequence of visited states $S_1, S_2, \dots$. Indeed, the sequence of states visited by the stochastic process now depends on the implemented strategy, and the grand coalition can acquire some statistical knowledge by inferring on its observation. Nevertheless, this is a "noisy" observation, since the transition rule is not deterministic.
The optimal CPDP for the grand coalition should take into account some form of punishment for those agent which deviated from the recommended strategy with a sufficiently high probability. This procedure should act as a deterrent for future deviations.

As to the work on uniform optimal strategies, developed in Section 2.1, I would like to extend the linear programming approach on the field $F(\mathbb{R})$ to the value iteration algorithm, that works for fixed discount factors.

Finally, I believe that the work on Multi-Armed Bandits with autoregressive arms, dealt with in Chapter 4, is a very insightful and interesting model. In order to be able to provide some analytical results, I am considering to discretize the state space and assume a Markov chain on it. Note that the original AR model can be approximated when the discretization is fine enough and the transition probability are chosen *ad hoc*. I am confident about the possibility to well approximate the dynamic allocation problem through an MDP formulation on a reduced state space, whose dimension is polynomial (and no longer exponential, as in the exact case) in the number of players.

Bibliography

[1] A. Agresti and B. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, pages 119–126, 1998.

[2] R. Aguero, M. Garcia, and L. Mufioz. BEAR: A bursty error auto-regressive model for indoor wireless environments. In *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, pages 1–5. IEEE, 2007.

[3] E. Altman. Applications of markov decision processes in communication networks. *International series in Operations Research and Management Science*, pages 489–536, 2002.

[4] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. Sadoc Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

[5] E. Altman, K. Avrachenkov, L. Cottatellucci, M. Debbah, G. He, and A. Suarez. Operating point selection in multiple access rate regions. In *Teletraffic Congress. ITC 21st International*, pages 1–8. IEEE, 2009.

[6] E. Altman, K. Avrachenkov, and J. A. Filar. Asymptotic linear programming and policy improvement for singularly perturbed markov decision processes. *Mathematical methods of operations research*, 49(1):97–109, 1999.

[7] E. Altman, E. A. Feinberg, and A. Shwartz. Weighted discounted stochastic games with perfect information. *Annals of the International Society of Dynamic Games*, 5:303–324, 2000.

[8] K. B. Athreya and C. D. Fuh. Bootstrapping Markov chains: Countable case. *Journal of Statistical Planning and Inference*, 33:311–331, 1992.

[9] R. J. Aumann. Economic Applications of the Shapley Value. In S. S. J.-F. Mertens, editor, *Game-Theoretic Methods in General Equilibrium Analysis*, pages 121–133. Kluwer Academic Publisher, 1994.

[10] R. J. Aumann and S. Hart. *Handbook of Game Theory with Economic Applications*, volume 2. Elsevier, 1994.

[11] K. Avrachenkov, L. Cottatellucci, and L. Maggi. Cooperative Markov decision processes: Time consistency, greedy players satisfaction, and cooperation maintenance. *International Journal of Game Theory*, 2012.

[12] K. Avrachenkov, L. Cottatellucci, and L. Maggi. Dynamic rate allocation in Markovian quasi-static multiple access channels. Technical Report RR-12-269, EURECOM, 2012.

[13] K. Avrachenkov, J. Elias, F. Martignon, G. Neglia, and L. Petrosyan. A Nash bargaining solution for Cooperative Network Formation Games. *NETWORKING 2011*, pages 307–318, 2011.

[14] Y. Bachrach, E. Markakis, E. Resnick, A. Procaccia, J. Rosenschein, and A. Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, 2010.

[15] Y. Bachrach, R. Meir, M. Feldman, and M. Tennenholtz. Solving cooperative reliability games. *Arxiv preprint arXiv:1202.3700*, 2012.

[16] J. F. Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317–343, 1964.

[17] N. Bäuerle and U. Rieder. *Markov Decision Processes with applications to finance*. Springer Verlag, 2011.

[18] R. Bellman. A markovian decision process. *J. Math. Mech.*, 6:679–684, 1957.

[19] T. Bewley and E. Kohlberg. The asymptotic theory of stochastic games. *Mathematics of Operations Research*, pages 197–208, 1976.

[20] J. Bilbao, J. Fernandez, A. J. Losada, and J. Lopez. Generating functions for computing power indices efficiently. *Top*, 8(2):191–213, 2000.

[21] D. Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2):719–726, 1962.

[22] O. N. Bondareva. Some applications of linear programming methods to the theory of cooperative games. *Problemy kibernetiki*, 10:119–139, 1963.

[23] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[24] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[25] G. Chalkiadakis, E. Markakis, and C. Boutilier. Coalition formation under uncertainty: Bargaining equilibria and the Bayesian core stability concept. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 64. ACM, 2007.

[26] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[27] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934.

[28] G. B. Dantzig. *Linear programming and extensions*. Princeton Univ Pr, 1998.

[29] B. C. Eaves and U. G. Rothblum. Formulation of linear problems and solution by a universal machine. *Mathematical Programming*, 65(1):263–309, 1994.

[30] J. Edmonds. Submodular functions, matroids, and certain polyhedra. *Combinatorial Optimization - Eureka, You Shrink!*, pages 11–26, 2003.

[31] J. Filar and L. A. Petrosjan. Dynamic cooperative games. *International Game Theory Review*, 2(1):47–65, 2000.

[32] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer Verlag, 1996.

[33] J. A. Filar, E. Altman, and K. Avrachenkov. An asymptotic simplex method for singularly perturbed linear programs. *Operations Research Letters*, 30(5):295–307, 2002.

[34] F. Forges, E. Minelli, and R. Vohra. Incentives and the core of an exchange economy: a survey. *Journal of Mathematical Economics*, 38(1):1–41, 2002.

[35] D. Gale. The core of a monetary economy without trust. *Journal of Economic Theory*, 19(2):456–491, 1978.

[36] D. Gillette. Stochastic games with zero stop probabilities. *Contributions to the Theory of Games*, 3:179–187, 1957.

[37] J. C. Gittins, R. Weber, and K. D. Glazebrook. *Multi-armed bandit allocation indices*, volume 25. Wiley Online Library, 1989.

[38] P. J. J. Herings, A. Predtetchinski, and A. Perea. The weak sequential core for two-period economies. *International journal of game theory*, 34(1):55–65, 2006.

[39] J. Herzog and T. Hibi. *Monomial Ideals*. Springer, 2010.

[40] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[41] A. Hordijk, R. Dekker, and L. Kallenberg. Sensitivity-analysis in discounted markovian decision problems. *OR Spectrum*, 7(3):143–151, 1985.

[42] R. A. Howard. Dynamic programming and markov processes. *The M.I.T. Press*, 1960.

[43] F. Javadi, M. R. Kibria, and A. Jamalipour. Bilateral Shapley Value Based Cooperative Gateway Selection in Congested Wireless Mesh Networks. In *IEEE GLOBECOM 2008*, pages 1–5. IEEE, 2008.

[44] M. Katz and S. Shamai. Transmitting to colocated users in wireless ad hoc and sensor networks. *Information Theory, IEEE Transactions on*, 51(10):3540–3563, 2005.

[45] B. Klinz and G. J. Woeginger. Faster algorithms for computing power indices in weighted voting games. *Mathematical Social Sciences*, 49(1):111–116, 2005.

[46] K. Knopp. *Theory and application of infinite series*. Dover Publications, 1990.

[47] L. Kranich, A. Perea, and H. Peters. Dynamic cooperative games. *Discussion Papers*, 2000.

[48] L. Kranich, A. Perea, and H. Peters. Core concepts for dynamic TU games. *International Game Theory Review*, 7(1):43–61, 2005.

[49] F. E. Kydland and E. C. Prescott. Rules rather than discretion: The inconsistency of optimal plans. *The Journal of Political Economy*, pages 473–491, 1977.

[50] R. J. La and V. Anantharam. A game-theoretic look at the Gaussian multiaccess channel. In *Advances in network information theory: DIMACS Workshop Network Information Theory*, volume 66, pages 87–105. American Mathematical Society, 2004.

[51] D. G. Luenberger and Y. Ye. *Linear and nonlinear programming*, volume 116. Springer Verlag, 2008.

[52] R. T. B. Ma, D. Chiu, J. Lui, V. Misra, and D. Rubenstein. Internet Economics: The use of Shapley value for ISP settlement. In *Proceedings of CoNEXT 2007*, Columbia University, New York, December 2007.

[53] M. A. Maddah-Ali, A. Mobasher, and A. K. Khandani. Fairness in multiuser systems with polymatroid capacity region. *Information Theory, IEEE Transactions on*, 55(5):2128–2138, 2009.

[54] M. Madiman. Cores of cooperative games in information theory. *EURASIP Journal on Wireless Communications and Networking*, (318704), 2008.

[55] G. J. Mailath and L. Samuelson. *Repeated games and reputations: long-run relationships.* Oxford University Press, USA, 2006.

[56] I. Mann and L. S. Shapley. Values of large games, IV. *RAND Corporation, Memoradum RM-2651*, 1960.

[57] T. Matsui and Y. Matsui. A survey of algorithms for calculating power indices of weighted majority games. *JOURNAL-OPERATIONS RESEARCH SOCIETY OF JAPAN*, 43:71–86, 2000.

[58] Y. Matsui and T. Matsui. Np-completeness for calculating power indices of weighted majority games. *Theoretical Computer Science*, 263(1):305–310, 2001.

[59] V. V. Mazalov and A. N. Rettieva. Fish wars and cooperation maintenance. *Ecological Modelling*, 221(12):1545–1553, 2010.

[60] R. Meir, M. Tennenholtz, Y. Bachrach, and P. Key. Congestion games with agent failures. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[61] J. F. Mertens and A. Neyman. Stochastic games. *International Journal of Game Theory*, 10(2):53–66, 1981.

[62] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, 2000.

[63] R. B. Myerson. *Game theory: analysis of conflict.* Harvard University Press, 1997.

[64] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the U.S.A.*, 36(1):48–49, 1950.

[65] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*, chapter 18, pages 461–486. Cambridge University Press, 2007.

[66] J. Oviedo. The core of a repeated n-person cooperative game. *European Journal of Operational Research*, 127(3):519–524, 2000.

[67] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queueing network control. *Mathematics of Operations Research*, 24, 1999.

[68] T. Parthasarathy and T. E. S. Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.

[69] B. Peleg and P. Sudhölter. *Introduction to the Theory of Cooperative Games*. Springer Verlag, 2007.

[70] M. Penn, M. Polukarov, and M. Tennenholtz. Congestion games with failures. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 259–268. ACM, 2005.

[71] M. Penn, M. Polukarov, and M. Tennenholtz. Congestion games with load-dependent failures: identical resources. *Games and Economic Behavior*, 67(1):156–173, 2009.

[72] L. A. Petrosjan. Stable solutions of differential games with many participants. *Viestnik of Leningrad University*, 19:46–52, 1977.

[73] L. A. Petrosjan. Cooperative stochastic games. *Advances in Dynamic Games*, pages 139–145, 2006.

[74] K. Prasad and J. S. Kelly. Np-completeness of some problems concerning voting games. *International Journal of Game Theory*, 19(1):1–9, 1990.

[75] A. Predtetchinski. The strong sequential core for stationary cooperative games. *Games and economic behavior*, 61(1):50–66, 2007.

[76] A. Predtetchinski, P. Herings, and H. Peters. The strong sequential core for two-period economies. *Journal of mathematical economics*, 38(4):465–482, 2002.

[77] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[78] T. E. S. Raghavan and Z. Syed. A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. *Mathematical programming*, 95(3):513–532, 2003.

[79] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

[80] S. Shamai and A. D. Wyner. Information-theoretic considerations for symmetric, cellular, multiple-access fading channels. I. *Information Theory, IEEE Transactions on*, 43(6):1877–1894, 1997.

[81] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the U.S.A.*, 39(10):1095, 1953.

[82] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:31–40, 1953.

[83] L. S. Shapley. On balanced sets and cores. *Naval research logistics quarterly*, 14(4):453–460, 1967.

[84] L. S. Shapley. Cores of convex games. *International journal of game theory*, 1(1):11–26, 1971.

[85] L. S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954.

[86] K. W. Shum and C. W. Sung. Fair rate allocation in some Gaussian multiaccess channels. In *Information Theory, 2006 IEEE International Symposium on*, pages 163–167. IEEE, 2006.

[87] K. W. Shum and C. W. Sung. On the fairness of rate allocation in Gaussian multiple access channel and broadcast channel. *Arxiv preprint cs/0611015*, 2006.

[88] E. Solan and N. Vieille. Computing uniformly optimal strategies in two-player stochastic games. *Economic Theory*, 42(1):237–253, 2010.

[89] R. Stanojevic, N. Laoutaris, and P. Rodriguez. On economic heavy hitters: Shapley value analysis of 95th-percentile pricing. In *Proceedings of the 10th annual conference on Internet measurement*, pages 75–80. ACM, 2010.

[90] P. Tannenbaum. Power in weighted voting systems. *Mathematica Journal*, 7(1):58–63, 1997.

[91] A. Taylor and A. Pacelli. *Mathematics and Politics: Strategy, Voting, Power, and Proof.* Springer Verlag, 2008.

[92] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[93] F. Thuijsman and T. E. S. Raghavan. Perfect information stochastic games and related classes. *International Journal of Game Theory*, 26(3):403–408, 1997.

[94] D. Tse and P. Viswanath. *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[95] J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

[96] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

[97] A. Wald and J. Wolfowitz. Confidence limits for continuous distribution functions. *Annals of Mathematical Statistics*, 10:105–118, 1939.

[98] R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, pages 637–648, 1990.

[99] D. J. White. A survey of applications of markov decision processes. *Journal of the Operational Research Society*, pages 1073–1096, 1993.

[100] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298, 1988.

[101] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *JASA*, pages 209–212, 1927.

[102] G. Zaccour. Time consistency in cooperative differential games: A tutorial. *INFOR: Information Systems and Operational Research*, 46(1):81–92, 2008.