International Journal of Game Theory manuscript No.
(will be inserted by the editor)

Cooperative Markov Decision Processes: Time Consistency, Greedy Players Satisfaction, and Cooperation Maintenance

Konstantin Avrachenkov - Laura Cottatellucci - Lorenzo Maggi (listed in alphabetical order).

Received: date / Accepted: date

Abstract We deal with multi-agent Markov Decision Processes (MDP's) in which cooperation among players is allowed. We find a cooperative payoff distribution procedure (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 MDP's.

Keywords cooperative Markov decision processes · stochastic games · payoff distribution procedure · time consistency · greedy players · cooperation maintenance

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). 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.)

Konstantin Avrachenkov
INRIA, BP95, 06902 Sophia Antipolis Cedex, France
E-mail: k.avrachenkov@sophia.inria.fr

Laura Cottatellucci, Lorenzo Maggi
Eurecom, Mobile Communications, BP193, F-06560 Sophia Antipolis Cedex, France
E-mail: laura.cottatellucci@eurecom.fr, lorenzo.maggi@eurecom.fr

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 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) and by Kranich, Perea, and Peters (2001) are the two independent pioneering works in this field.

While the theory of competitive Markov Decision Processes (MDP's), otherwise called non-cooperative stochastic games, has been thoroughly studied (Filar and Vrieze 1996 for an extensive survey), to the best of the authors' knowledge, there is very little work on cooperative MDP's in the literature. Unlike classic repeated games, in which the same game is played repeatedly over time, in cooperative MDP's 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 (2002), 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 MDP's, 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 (2002), 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), 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 2 is a short survey on non-cooperative and cooperative multi-agent MDP's. Following the lines of Petrosjan's work, in Section 3 we propose a stationary stage-wise CPDP for cooperative discounted MDP's (MDP-CPDP). In Section 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 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): 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 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 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), 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; Shapley 1967), 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 MDP's, dubbed as "Cooperation Maintaining solution". Finally, Section 8 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 val($\mathbf{A}$) stands for the value (e.g. Filar and Vrieze 1996) of the matrix $\mathbf{A}$. Let $\{C_i\}_i$ be a collection of sets; we define the sum set $\sum_i C_i$ as $\{\sum_i c_i : \ c_i \in C_i, \ \forall i\}$.

2 Discounted Cooperative Markov Decision Processes

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 $A_i(s)$, $i = 1, \ldots, P$, and $|A_i(s)| := m_i(s)$. To each $(P+1)$-tuple $(s, a_1, \ldots, a_P)$, with $a_i \in A_i(s)$, an immediate reward $r_i(s, a_1, \ldots, a_P)$ for player $i = 1, \ldots, P$ and a transition probability distribution $p(.|s, a_1, \ldots, a_P)$ 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 $\mathcal{P} := \{1, \ldots, P\}$ 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 $\boldsymbol{\Phi}_\Lambda^{(\beta)}(.) := [\Phi_\Lambda(s_1, .), \ldots, \Phi_\Lambda(s_N, .)]^T$ as

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

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.

$$\mathbf{f}_\mathcal{P}^{(\beta)*} = \operatorname*{argmax}_{\mathbf{f}_\mathcal{P}\in\mathbf{F}_\mathcal{P}} \boldsymbol{\Phi}_\mathcal{P}^{(\beta)}(\mathbf{f}_\mathcal{P}), \qquad \forall\, \beta \in [0; 1), \tag{2}$$

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

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

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), i.e.

$$v^{(\beta)}(\Lambda, \Gamma_s) := \max_{\mathbf{f}_{\Lambda} \in \mathbf{F}_{\Lambda}} \min_{\mathbf{f}_{\mathcal{P}\setminus\Lambda} \in \mathbf{F}_{\mathcal{P}\setminus\Lambda}} \Phi_{\Lambda}^{(\beta)}(s, \mathbf{f}_{\Lambda}, \mathbf{f}_{\mathcal{P}\setminus\Lambda}) \tag{3}$$

Nevertheless, we will consider the specific value formulation in (3) solely in Sections 6 and 7.3. Next we provide some useful definitions and preliminary results.

Definition 1 (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 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:

$$\begin{cases} \sum_{i\in I} b_i \mathbf{c}'_j(i) = \widetilde{\mathbf{c}}_j & \forall\, j : \{j\} \in \Lambda \\ \sum_{j:\{j\}\in\Lambda} \mathbf{c}'_j(i) \leq v(\Lambda, \Delta_i) & \forall\, i \in I \end{cases} \tag{4}$$

Let us prove that (4) admits a solution. Let $\epsilon_i \geq 0$, for all $i \in I$, be such that

$$\sum_{i\in I} \epsilon_i = \sum_{i\in I} b_i v(\Lambda, \Delta_i) - \sum_{j:\{j\}\in\Lambda} \widetilde{\mathbf{c}}_j \geq 0 \tag{5}$$

We write the following linear system

$$\begin{cases} \sum_{i\in I} b_i \mathbf{c}'_j(i) = \widetilde{\mathbf{c}}_j & \forall\, j : \{j\} \in \Lambda \\ b_i \sum_{j:\{j\}\in\Lambda} \mathbf{c}'_j(i) = b_i v(\Lambda, \Delta_i) - \epsilon_i & \forall\, i \in I \end{cases} \tag{6}$$

Evidently, any solution to (6) is also a solution to (4). Thanks to (5), the sum of the first $|\Lambda|$ equations of (6) equals the sum of the remaining $|I|$ equations. By discarding the last equation of (6) we obtain a linear system with $|\Lambda| + |I| - 1$ linearly independent equations in $|\Lambda||I| > |\Lambda| + |I| - 1$ unknowns. Hence, a solution to (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 2 (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.*

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

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

$$\begin{cases} \sum_{i\in\mathcal{P}} x_i = v^{(\beta)}(\mathcal{P}, \Gamma_s) \\ \sum_{i\in\Lambda} x_i \geq v^{(\beta)}(\Lambda, \Gamma_s), \ \forall\, \Lambda \subset \mathcal{P}. \end{cases} \tag{7}$$

A game with nonempty Core is said to be balanced. The strict Core $\mathbf{sCo}^{(\beta)}(\Gamma_s)$ is defined as in (7), but with the strict inequality signs.
The terminal cooperative solution $\mathbf{T} \equiv \mathbf{Sh}^{(\beta)}(\Gamma_s)$ stands for the Shapley value of the β-discounted game Γ_s, i.e. for all $i = 1, \ldots, P$,

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

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

Proposition 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 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 1 *For all $\beta \in [0; 1)$, $\sum_{i=1}^N b_i \mathbf{Sh}^{(\beta)}(\Gamma_{s_i}) = \mathbf{Sh}^{(\beta)}(\sum_{i=1}^N b_i \Gamma_{s_i})$, where $b_i \geq 0, \ \forall\, i$.*

Proof The proof follows straightforward from Proposition 1 and from the linearity property of the Shapley value.

3 Cooperative Payoff Distribution Procedure

In cooperative MDP's, 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 1 All the results presented in the current section, as well as the ones in Sections 4, 5, 7, 8 can be easily extended to undiscounted transient MDP's, i.e. games for which $\beta = 1$ and

$$\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} p_t(s'|s, \mathbf{f}_{\mathcal{P}}) < \infty, \quad \forall\, s \in \mathcal{S},\ \mathbf{f}_{\mathcal{P}} \in \mathbf{F}_{\mathcal{P}}. \tag{8}$$

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 (8).

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

Definition 3 (CPDP) *The cooperative payoff distribution procedure (CPDP)* $g^{(\beta)} := [g_1^{(\beta)}, \ldots, g_P^{(\beta)}]$ *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* $\mathbf{h}_t = [S_0, g^{(\beta)}(\mathbf{h}_0), S_1, \ldots, g^{(\beta)}(\mathbf{h}_{t-1}), S_t]$ *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 4 (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 (2002) introduced a CPDP for games on finite trees. Following his lines, we now propose a stationary CPDP for cooperative MDP's (MDP-CPDP) with β-discounted criterion, with $\beta \in [0;1)$ fixed *a priori*.

Definition 5 (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*

$$\gamma^{(\beta)}(s, \overline{\mathbf{T}}) := \sum_{s' \in \mathcal{S}} \left[\delta_{s,s'} - \beta\, p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\right] \overline{\mathbf{T}}^{(\beta)}(\Gamma_{s'}), \qquad \forall\, s \in \mathcal{S}. \tag{9}$$

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 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 (9).

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 (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 1 *The MDP-CPDP $\gamma^{(\beta)}(s, \overline{\mathbf{T}}) \in \mathbb{R}^P$, defined in (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 $i \in \mathcal{P}$,

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

If we substitute (9) in the equation above, we find that $\gamma_i^{(\beta)}$ defined in (9) satisfies the relation:

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

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 2 *The MDP-CPDP $\gamma^{(\beta)}$, defined in (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 (2) that it solves can be formulated as a Markov Decision Process (Puterman 1994). Hence, $r_i(s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})$ is also deterministic, for all $i \in \mathcal{P}$. Let us sum (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 (9) also fulfills a *terminal efficiency* property, i.e.

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

5 Time Consistency

Time consistency is a well known concept in dynamic cooperative theory (Filar and Petrosjan 2000, Zaccour 2008, 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 $\mathbf{h}_t\in\mathcal{H}_t$,*

$$\mathbb{E}\left[\sum_{k=t}^{\infty}\beta^k g^{(\beta)}(S_k,\mathbf{h}_{k-1})\Big|\mathbf{h}_t\right]\in\beta^t\mathbf{T}^{(\beta)}(\Gamma_{\bar{s}}),\tag{10}$$

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 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 *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 (10) as

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k \gamma^{(\beta)}(S_{t+k}) \Big| S_t = \bar{s}\right] \in \mathbf{T}^{(\beta)}(\Gamma_{\bar{s}}). \tag{11}$$

Thanks to Theorem 1, (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 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 1 we claim that, if $\mathbf{T} \equiv \mathbf{Sh}$, then

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

Thus (11) is verified for $\mathbf{T} \equiv \mathbf{Co}$ and $\mathbf{T} \equiv \mathbf{Sh}$, and the thesis is proven.

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 MDP's 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. $\mathbf{T} \equiv \mathbf{Sh}$.
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). 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), for all $\Lambda \subseteq \mathcal{P}$, $s \in \mathcal{S}$. The analogous expression holds for $v(\Lambda, \Omega_s)$.*

Lemma 1 *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

$$y^{(\beta_{n_k})}(s, \mathbf{f}_{\mathcal{P}}) := \Phi^{(\beta_{n_k})}_{\mathcal{P}}(s, \underline{\mathbf{f}}^*_{\mathcal{P}}) - \Phi^{(\beta_{n_k})}_{\mathcal{P}}(s, \mathbf{f}_{\mathcal{P}}) \geq 0, \qquad \forall\, k \in \mathbb{N}. \tag{12}$$

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 (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 $\Phi_{\mathcal{P}}^{(\beta)}(s, \underline{\mathbf{f}}_{\mathcal{P}}^*)$ 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:

$$\Theta_s : \left\{ \mathbf{x} \in \mathbb{R}^P : \sum_{i \in \mathcal{P}} x_i = \sum_{i \in \mathcal{P}} r_i(s, \underline{\mathbf{f}}_{\mathcal{P}}^*) \right\}, \tag{13}$$

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

$$\exists\, \beta^* > 0 \,:\, \underline{\mathbf{f}}_{\mathcal{P}}^* = \operatorname*{argmax}_{\mathbf{f}_{\mathcal{P}} \in \mathbf{F}_{\mathcal{P}}} \boldsymbol{\Phi}_{\mathcal{P}}^{(\beta)}(\mathbf{f}_{\mathcal{P}}) \quad \forall\, \beta \in [0; \beta^*). \tag{14}$$

Corollary 2 *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 2 and from Lemma 1.

Next we present a useful result.

Lemma 2 *Let $\mathbf{T} \equiv \mathbf{Sh}$. Under Condition 1, $\lim_{\beta \downarrow 0} \gamma^{(\beta)}(s) = \mathbf{Sh}^{(0)}(\Gamma_s) \equiv \mathbf{Sh}(\Omega_s)$.*

Proof Let us rewrite (9) as

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

It is sufficient to prove that $\lim_{\beta \downarrow 0} \mathbf{Sh}^{(\beta)}(\Gamma_s) = \mathbf{Sh}^{(0)}(\Gamma_s)$, $\forall\, s \in \mathcal{S}$. Since each component of the vector $\mathbf{Sh}^{(\beta)}(\Gamma_s)$ 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 1996)

$$|\operatorname{val}(\mathbf{B}) - \operatorname{val}(\mathbf{C})| \leq \max_{i,j} |\mathbf{B}_{i,j} - \mathbf{C}_{i,j}| \tag{15}$$

where $\mathbf{B}, \mathbf{C}$ are matrices with the same size. We know from Filar and Vrieze (1996) that

$$\begin{aligned} v^{(\beta)}(\Lambda, \Gamma_s) = \operatorname{val} \Bigg(\Big[\sum_{i \in \Lambda} r_i(s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda}) + \ldots \\ + \beta \sum_{s' \in \mathcal{S}} p(s'|s, a_\Lambda, a_{\mathcal{P}\setminus\Lambda})\, v^{(\beta)}(\Lambda, \Gamma_{s'}) \Big]_{a_\Lambda = 1, a_{\mathcal{P}\setminus\Lambda} = 1}^{m_\Lambda(s), m_{\mathcal{P}\setminus\Lambda}(s)} \Bigg), \end{aligned} \tag{16}$$

where $a_\Lambda \in A_\Lambda(s)$ and $a_{\mathcal{P}\setminus\Lambda} \in A_{\mathcal{P}\setminus\Lambda}(s)$. Thus, from (15,16) we can say that, for all $\Lambda \subseteq \mathcal{P}$,

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

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) 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 (13). Note that, in general, the affine hull of $\mathbf{Co}(\Omega_s)$ could be a proper subset of Θ_s.

Corollary 3 *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). The proof of *ii)* is straightforward.

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

Theorem 4 *Under Conditions 1 and 2, the greedy players satisfaction property is verified by $\gamma^{(\beta)}(\mathbf{Sh}^{(\beta)})$ for all discount factors β sufficiently close to 0.*

Proof Fix $s \in \mathcal{S}$. We know from Corollary 3 that $\mathbf{Sh}(\Omega_s)$ 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 2 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 (14). Hence, for Lemma 2 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 $s \in \mathcal{S}$.

7 Cooperation Maintenance

The (single step) cooperation maintenance property was first introduced by Mazalov and Rettieva (2010), 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 $\Lambda \subset \mathcal{P}$,*

$$\sum_{i\in\Lambda} g_i^{(\beta)}(s) + \beta v^{(\beta)}\left(\Lambda, \sum_{s'\in\mathcal{S}} p(s'|s, \mathbf{f}_{\mathcal{P}}^{(\beta)*})\, \Gamma_{s'}\right) \geq v^{(\beta)}(\Lambda, \Gamma_s). \qquad (17)$$

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 (17) are invariant with respect to a change of strategy by Λ, which can only withdraw at the following time step.

7.1 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 5 *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*

$$\left[\mathbf{I} - \left[\beta\mathbf{P}^{*(\beta)}\right]^n\right] \mathbf{v}^{(\beta)}(\Lambda,\Gamma) := \widetilde{\mathbf{v}}^{(\beta,n)}(\Lambda,\Gamma), \qquad \Lambda \subseteq \mathcal{P} \tag{18}$$

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 (9) as:

$$\boldsymbol{\gamma}_i^{(\beta)}(\overline{\mathbf{T}}) = \left[\mathbf{I} - \beta\mathbf{P}^{*(\beta)}\right]\overline{\mathbf{T}}_i^{(\beta)}, \quad \forall\, i \in \mathcal{P} \tag{19}$$

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 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

$$\begin{cases} \left[\mathbf{I} - \left[\beta\mathbf{P}^{*(\beta)}\right]^n\right]\sum_{i\in\Lambda}\overline{\mathbf{T}}_i^{(\beta)} \geq \left[\mathbf{I} - \left[\beta\mathbf{P}^{*(\beta)}\right]^n\right]\mathbf{v}^{(\beta)}(\Lambda,\Gamma), \quad \forall\,\Lambda\subset\mathcal{P} \\ \sum_{i\in\mathcal{P}}\overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases} \tag{20}$$

Since the matrix $\mathbf{I} - [\beta \mathbf{P}^{*(\beta)}]^n$ is invertible for any $n \in \mathbb{N}$, then we can equivalently rewrite (20) as

$$\begin{cases} \sum_{i \in \Lambda} \widetilde{\overline{\mathbf{T}}}_i^{(\beta,n)} \geq \widetilde{\mathbf{v}}^{(\beta,n)}(\Lambda, \Gamma), \ \forall \Lambda \subset \mathcal{P} \\ \sum_{i \in \mathcal{P}} \widetilde{\overline{\mathbf{T}}}_i^{(\beta,n)} = \widetilde{\mathbf{v}}^{(\beta,n)}(\mathcal{P}, \Gamma) \end{cases} \tag{21}$$

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 (21) are component-wise, for the Bondareva-Shapley Theorem (Bondareva 1963; Shapley 1967) the thesis is proven.

The reader should notice that, in the limit for $n \to \infty$, the result of Theorem 5 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 6 *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

$$\begin{cases} \beta \mathbf{P}^{*(\beta)} \left[\sum_{i \in \Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda, \Gamma)\right] \geq \sum_{i \in \Lambda} \overline{\mathbf{T}}_i^{(\beta)} - \mathbf{v}^{(\beta)}(\Lambda, \Gamma), \quad \forall \Lambda \subset \mathcal{P} \\ \sum_{i \in \mathcal{P}} \overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P}, \Gamma). \end{cases} \tag{22}$$

By iteratively left multiplying by the nonnegative matrix $\beta \mathbf{P}^{*(\beta)}$ both sides of the first expression in (22), then we obtain for each coalition $\Lambda \subset \mathcal{P}$:

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

Hence, the thesis is proven.

7.2 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 4 *Set $\beta \in [0;1)$. If an MDP-CPDP $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$ satisfies the single step cooperation maintenance property, then*

$$\mathbb{E}\left[\sum_{t\geq 0}\beta^t\gamma^{(\beta)}(S_t)\big|S_0=s\right] \in \mathbf{Co}^{(\beta)}(\Gamma_s), \quad \forall\, s\in\mathcal{S}. \tag{23}$$

Proof Since $\gamma^{(\beta)}$ satisfies Property 6, then (20) is verified, with $n=1$. By left multiplying each set of inequalities in (20) by the nonnegative matrix $(\mathbf{I}-\beta\mathbf{P}^{*(\beta)})^{-1}$, we obtain the following expressions:

$$\begin{cases} \sum_{i\in\Lambda}\overline{\mathbf{T}}_i^{(\beta)} \leq \mathbf{v}^{(\beta)}(\Lambda,\Gamma), & \forall\,\Lambda\subset\mathcal{P},\\ \sum_{i\in\mathcal{P}}\overline{\mathbf{T}}_i^{(\beta)} = \mathbf{v}^{(\beta)}(\mathcal{P},\Gamma). \end{cases} \tag{24}$$

Thanks to Theorem 1, we can say that the relations in (24) are equivalent to (23), hence the thesis is proven.

Interestingly, Corollary 4 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 7.2.1 we will show that it is actually a proper refinement. Therefore, it is worth coining a new terminal cooperative solution for cooperative MDP's, that we dub *Cooperation Maintaining solution*, grounded on the cooperation maintenance property.

Definition 6 (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 5 *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 (18), are balanced, for* $k = 1,\ldots,N$*.*

Corollary 6 *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 7 *For all* $\beta \in [0,1)$*,* $\mathbf{Cm}^{(\beta)}(\Gamma) \subseteq \cup_{s\in\mathcal{S}}\mathbf{Co}^{(\beta)}(\Gamma_s)$*.*

7.2.1 The Cooperation Maintaining solution is a proper refinement of the Core

It is natural to ask whether the converse of Corollary 7 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 MDP's.

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 7.2.1.

	(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 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

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

We find that:

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

Therefore, the converse of Corollary 7 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 $s \in \mathcal{S}$.

7.3 Strictly convex static games

In the same spirit of Section 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 6, we assume that Condition 1 holds, i.e. the coalition values are computed *'a la* von Neumann and Morgenstern.

Theorem 7 *Suppose that Conditions 1,2 hold. Then,* $\gamma^{(\beta)}(., \mathbf{Sh})$ *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 1) we can reformulate Property 6 as

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

From (9), considering $\mathbf{T}\equiv\mathbf{Sh}$,

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

By hypothesis, for all $s\in\mathcal{S}$ the Shapley value $\mathbf{Sh}(\Omega_s)=\mathbf{Sh}^{(0)}(\Gamma_s)$ 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 2), we conclude that, for all β sufficiently close to 0,

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

where $\underline{\mathbf{f}}_{\mathcal{P}}^{*}$ is the optimal strategy for grand coalition for all β sufficiently small, as in (14). Hence, the thesis is proven.

8 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.* $p(s'|s,a_1,\ldots,a_P)=p(s'|s)$*, for all* $a_i\in A_i(s)$ *and for each* $s,s'\in\mathcal{S}$*.*

As in Sections 6 and 7.3, 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), 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 8 *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

$$\begin{cases} \sum_{i\in\Lambda}\boldsymbol{\gamma}_i^{(\beta)} \geq \ \mathbf{v}(\Lambda,\Omega) & \forall\,\Lambda\subset\mathcal{P} \\ \sum_{i\in\mathcal{P}}\boldsymbol{\gamma}_i^{(\beta)} = \ \mathbf{v}(\mathcal{P},\Omega). \end{cases} \tag{25}$$

From the linearity property of coalition values (see Proposition 1) we claim that

$$\mathbf{v}^{(\beta)}(\Lambda,\Gamma) = \ \left[\mathbf{I}-\beta\mathbf{P}\right]^{-1}\mathbf{v}(\Lambda,\Omega) \quad \forall\,\Lambda\subseteq\mathcal{P}, \tag{26}$$

where $\mathbf{v}(\Lambda,\Omega) := \ [v(\Lambda,\Omega_{s_1}),\dots,v(\Lambda,\Omega_{s_N})]^T$. Thus, by left multiplying the expressions in (25) by the nonnegative matrix $(\mathbf{I}-\beta\mathbf{P})^{-1}$ we obtain

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

Hence, *i)* and *ii)* are proven by the construction of $\gamma^{(\beta)}(.,\overline{\mathbf{T}})$. By plugging (26) in (25), we can write

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

which coincides with the definition of the single step cooperation maintenance property. For Theorem 6, *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 8 for β sufficiently high, as pointed out in Section 6.
Moreover, Theorem 8 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 8 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 1 and equation (26).

Corollary 8 *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, $\gamma^{(\beta)}(s, \overline{\mathbf{T}}) = \mathbf{Sh}(\Omega_s)$, for all $s \in \mathcal{S}$, if and only if $\overline{\mathbf{T}}^{(\beta)}(\Gamma_s) = \mathbf{Sh}^{(\beta)}(\Gamma_s)$, for all $s \in \mathcal{S}$.*

9 Conclusions

This paper deals with Cooperative Markov Decision Processes, in which subcoalition 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 MDP's. 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.

References

1. O.N. Bondareva, Some applications of linear programming methods to the theory of cooperative games, Problemy Kybernetiki, Vol. 10, pp. 119-139 (1963).
2. J. Filar, L.A. Petrosjan, Dynamic Cooperative Games, International Game Theory Review, Vol. 2, No. 1, pp. 47-65 (2000).
3. J. Filar, K. Vrieze, Competitive Markov Decision Processes, Springer (1996).
4. L. Kranich, A. Perea, H. Peters, Dynamic Cooperative Games, Unpublished Mimeo (2001).
5. A. Hordijk, R. Dekker, L.C.M. Kallenberg, Sensitivity Analysis in Discounted Markov Decision Processes, OR Spektrum, Vol. 7, No. 3, pp. 143-151 (1985).

6. V.V. Mazalov, A.N. Rettieva, Fish wars and cooperation maintenance, Ecological Modelling, Vol. 221, Issue 12, pp. 1545-1553 (2010).
7. R.B. Myerson, Game Theory - Analysis of conflict, Harvard University Press (1991).
8. J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton: Princeton University Press (1944).
9. J. Oviedo, The Core of a Repeated n-Person Cooperative Game, European Journal of Operational Research, Vol. 127, Issue 3, pp. 519-524 (2000).
10. B. Peleg, P. Sudhölter, Introduction to the Theory of Cooperative Games, Springer, 2nd edition (2007).
11. L.A. Petrosjan, Cooperative Stochastic Games, Proceedings of the 10th International Symposium on Dynamic Games and Applications, Vol. 2 (2002).
12. A. Predtetchinski, The strong sequential core for stationary cooperative games, Games and Economic Behavior, Vol. 61, pp. 50-66 (2007).
13. M.L. Puterman, Markov Decision Processes, Wiley (1994).
14. L.S. Shapley, On balanced sets and cores, Naval Research Logistics Quarterly, Vol. 14, pp. 453-460 (1967).
15. L.S. Shapley, Cores of Convex Games, International Journal of Game Theory, Vol. 1, No. 1, pp. 11-26 (1971).
16. G. Zaccour, Time consistency in cooperative differential games: A tutorial, INFOR: Information Systems and Operational Research, Vol. 46, No. 1, pp. 81-92 (2008).

Acknowledgements This research was supported by "Agence Nationale de la Recherche" with reference ANR-09-VERS-001, and by the European research project SAPHYRE, which is partly funded by the European Union under its FP7 ICT Objective 1.1 - The Network of the Future. The authors would like to thank both Professor Eitan Altman for fruitful discussions and the reviewer for useful remarks.