A solution concept with an exploration bias for repeated stochastic coalitional games

Josselin Guéneron, Grégory Bonnet

A solution concept with an exploration bias for repeated stochastic coalitional games

Josselin Guéneron and Grégory Bonnet

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE
{firstname.lastname}@unicaen.fr

Abstract. Classically, in coalition formation, agents know in advance the deterministic utilities they will obtain from coalitions. Relaxing these two assumptions (determinism and *a priori* knowledge) is important to deal with real-world applications. A way to do that is to consider the framework of repeated stochastic coalitional games. Here, agents decide at each time step which coalition to form on the basis of limited information. Then, their observations allow them to update their knowledge. We propose a solution concept that explicitly integrates an exploration bias to allow agents to sometimes form coalitions that have a low utility but that would be interesting to form to obtain more information. We compare this concept to a greedy approach and highlight its efficiency with respect to the structure of the real utilities, unknown to the agents.

Keywords: Coalition Formation · Cooperative Game Theory · Sequential Decision

1 Introduction

In MAS, individual agents are not always able to perform certain tasks alone. When the system is composed of selfish and rational agents, the agents may form groups, called coalitions, in order to jointly perform tasks that cannot be handled individually. However, the majority of work on coalition formation makes two strong assumptions. The first is that agents have perfect *a priori* knowledge of the payoff they obtain when forming a coalition. The second one is that this payoff is deterministic. Both assumptions do not seem adequate for real-world problems where the exact payoff obtained by a coalition is only known a posteriori. Moreover, if the same coalition is subsequently reformed, its payoff has no reason to be strictly the same, due to internal or external factors. For example, consider legal entities that have to repeatedly form consortia to work temporarily on projects. These consortia formations are repeated by the same agents. However, the quality of the results produced by each consortium may vary. For instance an internal factor may be the agents' individual skills, whose effects may be stochastic, coupled with their ability to interact better with some agents than with others. An external factor could be an environmental effect independent of the agents, such as a disaster in the office of one of them. Thus, it seems interesting to relax assumptions of determinism and *a priori* knowledge.

However, it raises new questions. If agents no longer have knowledge about coalitions, how can they obtain it? If the payoff from coalitions is stochastic, how can they estimate it? The literature then proposes to consider repeated games, which allow the outcome of the same game to be observed sequentially. Thus, agents can observe the state of the game at different times and are able to extract information from it. Nevertheless, the main objective of coalition formation is to partition the agents into coalitions. In this new context, where the utilities of coalitions are stochastic and unknown to the agents, how do we decide which coalitions to form? This choice can be broken down into several questions. How can agents favour the formation of one coalition over another based on what they know about them? At what point do they consider that they know enough about a coalition to properly assess its usefulness? How can agents collectively decide which interactions could be accepted by all?

In this article, we propose a new solution concept for repeated stochastic coalition formation, based on an exploration-exploitation trade-off, well-known in reinforcement learning. To do so, we redefine an existing solution concept, integrating a notion of interest in exploration in order to allow agents to form stable coalitions while accepting to form them to obtain more information. This consists in introducing a new notion of stability for repeated stochastic coalitional games, to allow a trade-off between exploiting a rather known payoff and exploring coalitions with unknown or very uncertain payoffs (taking into account mean and variance of the characteristic function known at a given time step). We show that our solution concept is very efficient on unstructured characteristic functions, and is better than an ϵ-greedy strategy except in the case of a highly structured characteristic function. These results extend previous results [11] which did not considered some informations (i.e. variance) and unstructured characteristic functions.

Section 2 presents coalitional games and their stochastic and repeated extensions. Section 3 describes our solution concept which explicitly integrates a notion of exploration, as well as an instantiation of such notion. Finally, section 4 presents experimental results, highligthing the interest of our solution concept.

2 State of the art

In cooperative game theory, agents cooperate by forming *coalitions* which produce some *utilities*.

Definition 1 (Coalitional game). *A* coalitional game *is a tuple* $\mathcal{G} = \langle N, v \rangle$ *where* $N = \{a_1, \ldots, a_n\}$ *is a set of agents and* $v : 2^N \to \mathbb{R}$ *is the characteristic function that associates a utility* $v(C)$ *to each coalition* $C \subseteq N$*.*

A partition of agents into coalitions is called a *coalition structure* and a *solution* to a coalitional game is defined as follows.

Definition 2 (Solution). *A* solution *to a coalitional game* $\mathcal{G}$ *is a tuple* $S_\mathcal{G} = \langle \mathcal{CS}, \vec{x} \rangle$ *where* $\mathcal{CS}$ *is a coalition structure of* N*,* $\vec{x} = \{x_1, \ldots, x_n\}$ *is a payoff vector for agents where* $x_i \geq 0$ *is the payoff of agent* a_i*.*

As agents are selfish, a solution must be accepted by all of them. This is why a solution must belong to a *solution concept*. A solution concept is the set of solutions that respect a certain notion of stability, e.g. the agents do not wish to form or join another coalition where they could earn more. We focus in this article on the concept of the core and its generalization, the *ϵ-core* [15, 17].

Definition 3 (ϵ-core). *A solution $\langle \mathcal{C}, \vec{x} \rangle$ belongs to the ϵ-core if and only if:*

$$\forall C \subseteq N, x(C) \geq v(C) - \epsilon \text{ with } x(C) = \sum_{a_i \in C} x_i$$

The ϵ-core allows to define the *least core*, which contains all ϵ-core solutions for the smallest value of ϵ for which the solution concept is non-empty [7]. On the first hand, the determinism of the characteristic function can be relaxed in the literature with stochastic coalitional games [8, 6, 12]. The nature of uncertainty in these models differs, from probabilitic distribution on deterministic characteristic functions [12] to agents having beliefs about capabilities of others [6]. The most general and abstract model we can consider was proposed by Charnes and Granot [9]. Here, the characteristic function $v : 2^N \to \mathcal{X}_{2^N}$ is simply defined by random variables, and the payoff vectors of the solutions are calculated on the expectation of those variables. In the sequel, we consider such kind of model.

On the other hand, relaxing the *a priori* knowledge of the characteristic function leads to repeated games [4]. Such games consists in repeating the following process at each time step: (1) the agents form coalitions based on their current knowledge; (2) the coalitions are formed and produce payoffs; (3) the agents update their knowledge based on the previous observed payoffs. Here again, models in the literature essentially differ on the nature of what the agents learn and how they estimate the coalitions. Generally, they learn a reliability value or skill expertise for each agent, which impacts in turn the characteristic function [5, 6, 13]. However from the most abstract point of view, they can simply learn the characteristic function [11]. In the sequel, we consider such an abstract model for shake of generality.

Definition 4 (Repeated stochastic coalitional game). *Let $\mathcal{G} = \langle N, \mathbb{T}, v, \hat{v} \rangle$ be a* repeated stochastic coalitional game *(RSCG) where: $N = \{a_1, \ldots, a_n\}$ is a set of agents, $\mathbb{T} \subset \mathbb{N}^+$ is a set of distinct time steps, $v : 2^N \to \mathcal{X}^{2^N}$ a stochastic characteristic function – unknown to the agents – that associates a random variable to each coalition, and $\hat{v} : 2^N \times \mathbb{T} \to \hat{\mathcal{X}}^{2^N}$ a characteristic function that associates an estimated utility to each coalition at each time step.*

At each time step, agents have to decide on a solution to the game, despite the fact that they do not know the characteristic function v *a priori*. A solution is, like in a deterministic context, a tuple made of a coalition structure and an *ex ante* payoff vector, i.e. an estimated payoff vector based on what the agents know about v.

Definition 5 (Solution to a RSCG). *A* solution S^t *at the time step $t \in \mathbb{T}$ to a RSCG $\mathcal{G}$ is a tuple $S^t = \langle \mathcal{CS}^t, \vec{x}^t \rangle$ where $\mathcal{CS}^t$ is a coalition structure (disjointed*

partition) of N, and $\vec{x}^t = \{x_1^t, \ldots, x_n^t\}$ is a payoff vector such that $x_i^t \geq 0$ is the payoff of the agent a_i based on $\hat{v}$ and the coalition a_i belongs to in $\mathcal{CS}^t$.

It has been shown that repeated coalition formation processes converge towards equilibria if agents sequentially form Pareto-efficient coalition structures [13]. However, forming Pareto-efficient coalition structures allows agents to form irrational solutions, i.e. at least one agent receive a payoff lesser than the utility he would receive while being alone. Interestingly, the processes still converge experimentally with greedy strategies [5, 6, 11] based on the expected values of coalitions. However, RSCG may allow to have more information than just the expected value of the characteristic function, for instance all standardized moments (mean, variance, skewness, kurtosis). As in other sequential decision-making problems it has been demonstrated that exploring can help on the long-term, i.e. making *a priori* sub-optimal decision in order to acquire knowledge [10, 14], we propose to extend RSCG with an explicit notion of information, and with a new solution concept which takes into account the collective interest to make such sub-optimal choices.

3 An exploration-based ϵ-core

We propose to adapt the RSCG framework and the associated ϵ-core solution concept by considering that the value of a coalition, i.e. its interest to be formed at a given time step, depends on two elements: an estimation of its utility from which the payoffs are directly derived, and an interest that the agents have in forming it in order to obtain more information on its real utility. Notice that unlike what was done in [11], the interest is intrinsic to the game definition.

Definition 6 (Interest-biased repeated stochastic coalitional game). *Let $\mathcal{G} = \langle \mathcal{G}, i \rangle$ be an* interest-biased repeated stochastic coalitional game *(IRSCG) where $\mathcal{G}$ is a RSCG and $i : 2^N \times \mathbb{T} \to \mathbb{R}$ an interest function that associates a quantitative interest to each coalition at each time step. We denote $i(C, t)$ the interest of the coalition C at time step t.*

As mentioned in Section 2, the agents' payoff for a given solution is an estimate. Once the solution is found and coalitions are formed, the actual utilities they produced are the result of stochastic processes parameterised by the characteristic function. We assume that these utilities are observed by all agents. We denote X_t^C the observation of the utility produced by C at time step t.

Definition 7 (Observations). *Let $\mathcal{O}_t = \{(C, t', X_{t'}^C) : C \subseteq 2^N, t' \in \mathbb{T}, t' < t\}$ be a set of observations at time step t corresponding to the set of the coalitions formed at each time step before t and their ex-post payoffs.*

Thereafter, let us note $\mathcal{O}_t(C)$ the set of observations at time step t associated with the coalition $C \subseteq 2^N$. This set of observations allows to update the knowledge of agents about the characteristic function. In the following, we assume that agents estimate the utility of coalitions as normal distributions. Thus, for a

given coalition $C \subseteq 2^N$, $\hat{v}(C,t)$ is characterised by the expectation and variance of a normal distribution over all observations.

Definition 8 (Utility estimation). *At time step $t \in \mathbb{T}$ and for the coalition $C \subseteq 2^N$, the estimated value of C, $\hat{v}(C,t)$, is given by $\hat{\mu}^2(C,t)$ its expectation and $\hat{\sigma}^2(C,t)$ its variance, which are computed from the observations of $\mathcal{O}_t(C)$.*

Thus, the learning method we used is tabular and is similar to what is used for multi-armed bandits. Since there is uncertainty about the utility produced by the coalitions once formed, a solution must take this uncertainty into account to be stable.

3.1 Interest of coalitions

The exact nature of the interest that agents have in a coalition may depend on the problem. However, the purpose of this interest is to make it possible to explore other solutions that are potentially interesting for the agents but that might be considered unstable in the sense of a classical solution concept. However, it is important to note that in coalition formation we need to compare coalition structures, which therefore involves comparing different coalitions. For example, in the case of the core solution concept, checking the stability of a solution involves comparing the utility of a coalition to the sum of the individual gains of the agents in that same coalition wherever they are in the solution. We need to consider a form of interest that allows us to make such comparisons, i.e. to calculate from the interest of a coalition the individual interest of the agents that compose it.

Definition 9 (Individual interest). *The* individual interest $i_j(C_j,t)$ *of a agent a_j for a coalition C_j to which he belongs at a time step t is:*

$$i_j(C_j,t) = \frac{i(C_j,t)}{|C_j|}$$

This egalitarian distribution is one of many ways of distributing interest, and it represents the fact that each agent in a coalition has the same interest in that coalition, regardless of the other coalitions to which they may belong. However, it should be noted that the more agents a coalition contains, the lower their individual interest will be. This distribution therefore tends to favour coalitions of low cardinality, as several distinct observations can yield more information than a single one. This individual interest allows us to define the interest of a coalition with respect to a given coalition structure, regardless of whether or not the coalition agents are together in the structure.

Definition 10 (Collective interest). *The* collective interest $i^{\mathcal{CS}}(C,t)$ *of the agents of the coalition C w.r.t. of a coalition structure $\mathcal{CS}$ at a time step t is:*

$$i^{\mathcal{CS}}(C,t) = \sum_{a_j \in C} i_j(C_j^{\mathcal{CS}},t)$$

where $C_j^{\mathcal{CS}}$ is the coalition of the agent a_j in the structure $\mathcal{CS}$.

3.2 λ-core

In order to integrate this interest in coalitions into the solution concept, we must aggregate it with utility. In order to remain generic at first, we consider in an abstract way an aggregation operator noted $\oplus$. Depending on the exact nature of the interest, this operator can take different forms, for example an *addition*, a *multiplication* or even a *maximum*. The various elements describing the interest of the agents being defined, we can now build our solution concept, the λ-core, based on an exploration-exploitation trade-off. To do this, we adapt the concept of the ϵ-core by integrating the aggregation operator and the interest function. We therefore add, on one side of the inequation, the interest of a coalition to the expected utility of the coalition, and on the other side the collective interest to the sum of the agents' gains with respect to the solution considered.

Definition 11 (λ-core). *A solution $\langle \mathcal{CS}^t, \vec{x}^t \rangle$ belongs to the λ-core if and only if $\forall\ C \subseteq N$:*

$$\vec{x}^t(C) \oplus i^{\mathcal{CS}^t}(C,t) \geq \hat{\mu}(C,t) \oplus i(C,t) - \lambda \text{ with } \vec{x}^t(C) = \sum_{a_i \in C} \vec{x}_i^t$$

In a similar way to the ϵ-core, the least core for this concept of the λ-core is defined as the one with the smallest λ for which a solution exists. We can now propose an example of instantiation of this solution concept by defining the interest as an exploration bias, and the aggregation operator as an *addition*.

3.3 Example of interest: exploration bias

A relevant notion of interest is that of exploration, which we find in the multi-armed bandit problem. For this problem, many strategies have been proposed, and in particular strategies based on a *Upper Confidence Bound* called *UCB* [1]. Among the strategies based on this principle, there is *UCB-V*, which was proposed for the multi-armed bandit problem by Audibert *et al.* [2]. This describes an exploration bias taking into account the variance of the underlying probability distributions of the multi-armed bandit's arms and has been shown to be more efficient than the strategy *UCB-1* [3]. We therefore adapt *UCB-V* to apply it to interest-biased repeated stochastic coalitional games.

Definition 12 (UCB-V exploration bias). *The* UCB-V *exploration bias for a given coalition C at a time step t is defined as follows:*

$$i(C,t) = \sqrt{\frac{2\hat{\sigma}^2(C,t)\eta}{|O_t(C)|+1}} + c\frac{3b\eta}{|O_t(C)|+1} \text{ with } \eta = \zeta.log(|O_t|+1)$$

Some constants must be defined. The constant b defines the upper bound of the problem's payoffs, so this is dependent on the latter. However, we can assume that the utilities are normalized over the interval $[0,1]$ as in multi-armed bandit problem, and thus define $b = 1$. The constants ζ and c are control parameters of the exploration (in particular ζ). We take here the values of the original article, in which Audibert *et al.* show the efficiency of these constants when they are defined as $\zeta = 1.2$ and $c = 1$.

4 Experiments

In order to evaluate the performance of our solution concept, we proceed empirically. We generate random sets with different characteristic functions, constructed in structured and random ways. We then apply our solution concept to these games, as well as to mixtures of these games, in order to test our concept on different degrees of structuring, from fully structured to fully random games. The performance is measured with the instant regret of the stable solutions found at each time step, in order to calculate the cumulative regret over all time steps.

4.1 Experimental protocol

In a first step, we construct 200 different pairs of games with unique characteristic functions, for 6 agents. Each pair of games is constructed with two different characteristic function structures. The first characteristic function is drawn according to the NDCS (*Normally Distributed Coalition Structures*) [16] model. This model makes it possible to construct structured characteristic functions, but without strongly constraining the model as with monotonic or superadditive structures [7]. Thus, the utility expectation μ_C of each coalition $C \subseteq N$ is drawn according to a normal distribution $\mathcal{N}(|C|, \sqrt{|C|})$. The second characteristic function is unstructured, as it is drawn randomly and uniformly for each coalition. Thus, the utility expectation μ_C of each coalition $C \subseteq N$ is drawn according to a uniform distribution $\mathcal{U}(0, 1)$. In both structuring models, the variances σ_C^2 of each coalition C are drawn according to a uniform distribution $\mathcal{U}(0, \frac{\mu_C}{2})$. Each characteristic function is then normalized on the interval $[0, 1]$.

In a second step, in order to create a series of games from the most to the least structured, for each pair of games, we create intermediate games using a linear transformation by applying a transformation factor $w \in [0, 1]$. Thus, a transformation factor of 0 corresponds to the NDCS structured game, while the factor 1 corresponds to the randomly structured game. A game is created in 0.05 steps for w between the two games of the pair, which corresponds to 19 additional intermediate games. Our solution concept is thus evaluated over 4200 games and 100 time steps each.

Example 1. Let C and C' be two coalitions, and v_1 and v_2 be two characteristic functions, respectively randomly and NDCS structured:

$$v_1 = \{C = \mathcal{N}(0.6, 0.2), C' = \mathcal{N}(0.1, 0.4)\}$$
$$v_2 = \{C = \mathcal{N}(0.2, 0.4), C' = \mathcal{N}(0.5, 0.1)\}$$

For a transformation factor of 0.4, the utilities of C and C' are such that:

$$v_{(1,2)}^{0.4} = \{C = \mathcal{N}(0.36, 0.32), C' = \mathcal{N}(0.34, 0.22)\}$$

For a transformation factor of 1, the resulting characteristic function is v_1:

$$v_{(1,2)}^{1} = \{C = \mathcal{N}(0.6, 0.2), C' = \mathcal{N}(0.1, 0.4)\}$$

For a transformation factor of 0, the resulting characteristic function is v_2:

$$v^0_{(1,2)} = \{C = \mathcal{N}(0.2, 0.4), C' = \mathcal{N}(0.5, 0.1)\}$$

These games are also played with the ϵ-greedy strategy, which will be our reference strategy. It also describes an exploration-exploitation trade-off, exploring randomly with probability ϵ, and exploiting with probability $1 - \epsilon$. In our implementation, the exploitation consists in using the concept of the least core with an ϵ (for ϵ-greedy) value of 0.05.

4.2 Performance measures

The first measure is the *instant regret*, which is the difference between the maximum social welfare – i.e. the maximum sum of payoffs produced by a coalition structure – of the game and the sum of the actual expected utilities of the coalitions of the structure formed at time step t. Formally:

Definition 13 (Instant regret). *Given the optimal solution $S^* = \langle \mathcal{CS}^*, \vec{x}^* \rangle$ in the sense of social welfare, the instant regret at time step t, noted R^t, is defined such that:*

$$R^t = \sum_{C^* \in \mathcal{CS}^*} \mu_{C^*} - \sum_{C \in \mathcal{CS}^t} \mu_C$$

Due to stochasticity, instant regret can oscillate (sometimes with a large amplitude), which is why the second measure is the *cumulative regret*. This measures the evolution of instant regret over time and highlights the convergence of regret, i.e. the time step at which the strategies have reached their exploration-exploitation equilibrium and therefore produce constant instant regret. At a time step t, the cumulative regret is the sum of the instant regrets of each time step $t' \leq t$. Formally:

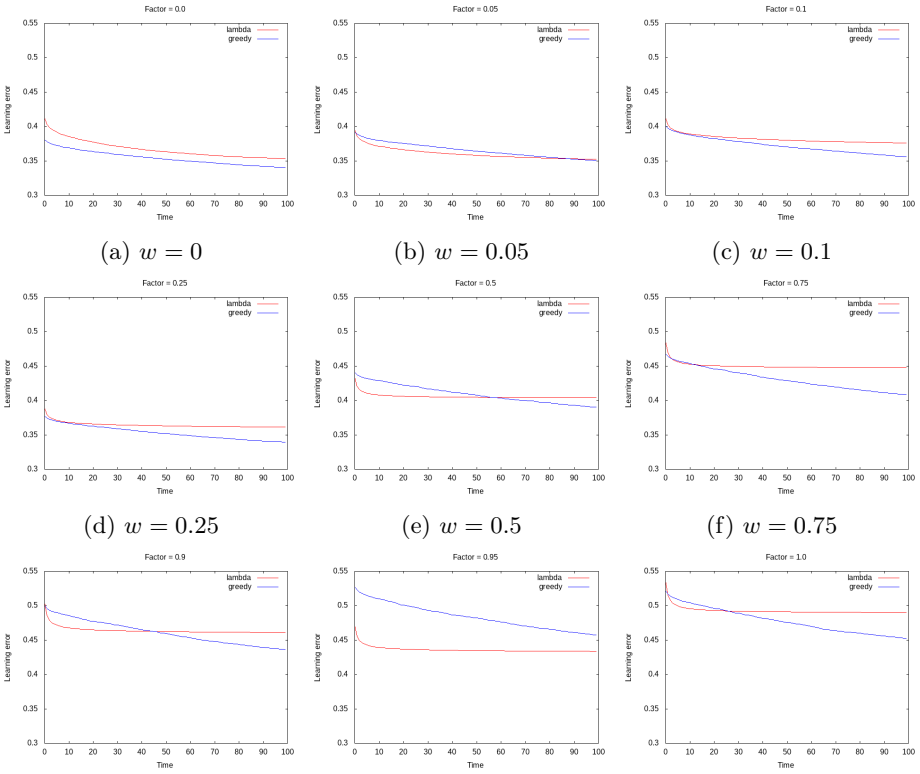
Definition 14 (Cumulative regret). *Given the optimal solution $S^* = (\mathcal{CS}^*, \vec{x}^*)$ in the sense of social welfare, the cumulative regret at time step t, noted R^t_c, is defined such that:*

$$R^t_c = \sum_{t'=0}^{t} R^{t'}$$

Finally, in order to evaluate the learning that the agents do of the real characteristic function over time, we use the *mean absolute error* (MAE) on the estimated and real utilities of the coalitions. The closer the MAE is to 0, the more accurate the estimated characteristic function is. The MAE is defined as:

Definition 15 (Mean absolute error). *Let v and $\hat{v}$ be two characteristic functions, the* mean absolute error *D^t_{MAE} between v and $\hat{v}$ at time step t is defined as:*

$$D^t_{MAE} = \frac{\sum_{C \in 2^N} |\hat{\mu}(C, t) - \mu_C|}{|2^N|}$$



(a) $w = 0$ (b) $w = 0.05$ (c) $w = 0.1$

(d) $w = 0.25$ (e) $w = 0.5$ (f) $w = 0.75$

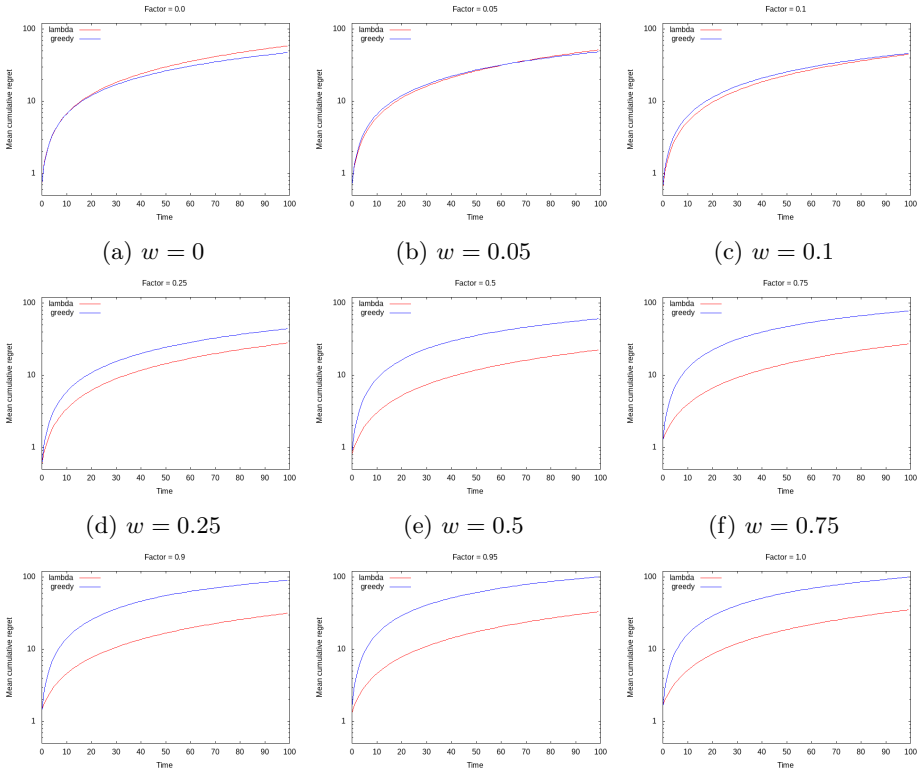
(g) $w = 0.9$ (h) $w = 0.95$ (i) $w = 1$

Fig. 1: Learning error for 6 agents

4.3 Results

Figures 1 and 2 show respectively the evolution of the means of the learning error and the cumulative regret of the set of games for a given configuration (i.e. a linear transformation factor w) over the 100 time steps. Figure 3 summarizes the results with the relative percentage of efficiency of the λ-core versus ϵ-greedy for the different transformation factors.

Concerning the learning error in figure 1, a first point to underline is that the more the characteristic function is structured (thus the closer the transformation factor w is to 0), the less the learning error is. In general, the ϵ-greedy strategy is the one that learns best, with a few exceptions such as for $w = 0.95$ where the λ-core allows better learning, or $w = 0.05$ where the results of the two methods are very close. However, we can see graphically a difference in behaviour between them according to the structuring of the characteristic functions. Indeed, the more structured the characteristic functions are, the more the ϵ-greedy strategy learns between the beginning and the end of the experiments. For example, its learning error decreases respectively by 10.59%, 11.45% and 13.20% for $w = 0$,



(a) $w = 0$ (b) $w = 0.05$ (c) $w = 0.1$

(d) $w = 0.25$ (e) $w = 0.5$ (f) $w = 0.75$

(g) $w = 0.9$ (h) $w = 0.95$ (i) $w = 1$

Fig. 2: Mean cumulative regret for 6 agents

$w = 0.5$ and $w = 1$. Let us note that this decrease is quasi-linear with the variation of the factor w. Concerning the λ-core, we can see that the learning converges quickly, due to the *UCB-V* exploration bias, and this more and more quickly as the characteristic functions are destructured. For example, for $w = 0$, the learning error decreases throughout the experiment, while for $w = 1$, the error almost stops decreasing after the time step $t = 20$. From a more general point of view for both methods, the more unstructured the characteristic functions are, the larger the learning error is initially.

Let us then look at the mean cumulative regret in figure 2. For a transformation factor $w = 0$, i.e. with a pure NDCS structure, the mean cumulative regret is in favour of the ϵ-greedy strategy, just as for a $w = 0.05$. However, from $w = 0.1$ onwards, the λ-core performs better in terms of regret, and the gap is larger for larger values of w. From these results, we can deduce that the ϵ-greedy strategy performs well on highly structured characteristic functions but that the less structuring there is, the less well it performs. However, it should be noted that when the ϵ-greedy strategy is outperformed by the λ-core, it is mainly the

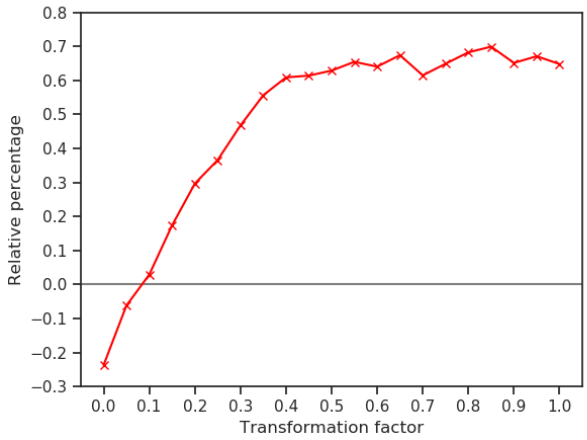


Fig. 3: Relative percentage of efficiency of λ-core against the ϵ-greedy strategy

latter that gains in performance more than the ϵ-greedy strategy loses. Indeed, the latter obtains a mean cumulative regret of 45.91 for $w = 0$ and 47.13 for $w = 0.1$, i.e. a difference of 1.22. For its part, the λ-core obtains a mean cumulative regret of 58.28 for $w = 0$ and 44.58 for $w = 0.1$, i.e. a difference of 13.70. This difference is 16.07 for $w = 0.25$, 37.91 for $w = 0.5$, 35.11 for $w = 0.75$, up to a difference of 64.39 for $w = 1$. In the latter case, the mean cumulative regret for the λ-core is 34.98 while it is 99.37 for the ϵ-greedy strategy. The relative efficiency of λ-core against the ϵ-greedy strategy is highlighted in figure 3. On the latter, we can see that the gap in favour of the λ-core only increases until $w = 0.4$ and then stabilises. For $w = 0$, λ-core is $23,66\%$ less efficient than ϵ-greedy. It becomes $2,9\%$ more efficient from $w = 0.1$, until $60,87\%$ for $w = 0.4$. Then, for $w \geq 0.4$, the relative efficiency in favour of λ-core stabilises around 65%, with a maximum of 69.90% for $w = 0.85$. Thus, the λ-core solution concept performs very well on unstructured characteristic functions, and remains more efficient than the ϵ-greedy strategy as long as the structuring is not very important. It is however necessary to note that the λ-core is more efficient on slightly structured characteristic functions. For example, it obtains a mean cumulative regret of 22.39 with $w = 0.5$, while for $w = 1$ it is 34.98 (with a minimum for $w = 0.45$ with 21.94 of mean cumulative regret).

5 Conclusion

In this paper, we proposed the interest-biased repeated stochastic coalitional games. This model allows a new solution concept, the λ-core, based on an exploration-exploitation trade-off by integrating a notion of interest for the agents. By setting this interest to an exploration bias and defining the aggregation as an addition, we have shown that this solution concept is efficient on

repeated stochastic coalitional games, especially when the characteristic functions are not very strongly structured. However, the computation of the λ-core is time consuming due to exploration bias. Indeed, this bias leads the least core to have a high value of λ, and thus to traverse more the space of the solutions because a naive approach of this calculation consists in seeking λ-core by iteratively incrementing the value of λ. Thus, it would be relevant to work on a distributed or decentralised approach of the calculation.

References

1. Agrawal, R.: Sample mean based index policies by $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem. Advances in Applied Probability **27**(4), 1054–1078 (1995)
2. Audibert, J., Munos, R., Szepesvári, C.: Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theor. Comput. Sci. **410**(19), 1876–1902 (2009)
3. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine learning **47**(2), 235–256 (2002)
4. Benoit, J.P., Krishna, V.: Finitely repeated games. Foundations in Microeconomic Theory pp. 195–212 (1984)
5. Blankenburg, B., Dash, R.K., Ramchurn, S.D., Klusch, M., Jennings, N.R.: Trusted kernel-based coalition formation. In: 4th AAMAS. pp. 989–996 (2005)
6. Chalkiadakis, G., Boutilier, C.: Sequential decision making in repeated coalition formation under uncertainty. In: 7th AAMAS. pp. 347–354 (2008)
7. Chalkiadakis, G., Elkind, E., Wooldridge, M.: Computational aspects of cooperative game theory. Synthesis Lectures on Artificial Intelligence and Machine Learning **5**(6), 1–168 (2011)
8. Charnes, A., Granot, D.: Prior solutions: Extensions of convex nucleus solutions to chance-constrained games. Tech. rep., Texas Univ. (1973)
9. Charnes, A., Granot, D.: Coalitional and chance-constrained solutions to n-person games. i: The prior satisficing nucleolus. SIAM J. Appl. Math. **31**(2), 358–367 (1976)
10. Gittins, J.C.: Bandit processes and dynamic allocation indices. J. R. Stat. Soc. Series B Stat. Methodol. **41**(2), 148–164 (1979)
11. Guéneron, J., Bonnet, G.: Are exploration-based strategies of interest for repeated stochastic coalitional games? In: 19th PAAMS. pp. 89–100 (2021)
12. Ieong, S., Shoham, Y.: Bayesian coalitional games. In: 23rd AAAI Conference on Artificial Intelligence. pp. 95–100 (2008)
13. Konishi, H., Ray, D.: Coalition formation as a dynamic process. Journal of Economic theory **110**(1), 1–41 (2003)
14. Mahajan, A., Teneketzis, D.: Multi-armed bandit problems. Foundations and applications of sensor management pp. 121–151 (2008)
15. Mochaourab, R., Jorswieck, E.A.: Coalitional games in MISO interference channels: Epsilon-core and coalition structure stable set. IEEE Transactions on Signal Processing **62**(24), 6507–6520 (2014)
16. Rahwan, T., Ramchurn, S.D., Jennings, N.R., Giovannucci, A.: An anytime algorithm for optimal coalition structure generation. Journal of Artificial Intelligence Research **34**, 521–567 (2009)
17. Shapley, L.S., Shubik, M.: Quasi-cores in a monetary economy with nonconvex preferences. Econometrica: Journal of the Econometric Society pp. 805–827 (1966)