Un concept de solutions avec un biais d'exploration pour les jeux de coalitions stochastiques répétés

J. Guéneron[a] josselin.gueneron@unicaen.fr
G. Bonnet[a] gregory.bonnet@unicaen.fr

[a]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000, France

Résumé

Classiquement, en formation de coalitions, les agents connaissent à l'avance les utilités déterministes qu'ils vont obtenir des coalitions. Relâcher ces deux hypothèses (déterminisme et connaissance a priori) nous place dans un cadre de jeux de coalitions stochastiques répétés. Les agents doivent décider à chaque pas de temps quelle coalition former sur la base d'informations limitées. Ils obtiennent alors des observations qui permettent de mettre à jour leurs connaissances. Nous proposons un concept de solutions qui intègre explicitement une notion d'exploration pour permettre aux agents de parfois former des coalitions ayant une faible utilité mais qui seraient intéressantes à former pour obtenir plus d'informations. Nous comparons ce concept à une approche gloutonne et mettons en lumière son efficacité en fonction de la structuration des utilités réelles, inconnues des agents.

Mots-clés : *Formation de coalitions, Théorie des jeux coopératifs, Décision séquentielle*

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) takes place in a framework of repeated stochastic coalition games. The agents must decide at each time step which coalition to form on the basis of limited information. They then obtain observations that allow them to update their knowledge. We propose a solution concept that explicitly integrates a notion of 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 effectiveness with respect to the structure of the real utilities, unknown to the agents.

Keywords: *Coalition Formation, Cooperative Game Theory, Sequential Decision*

1 Introduction

Dans un système multi-agents (SMA), les agents individuels ne sont pas toujours capables de réaliser certaines tâches seuls. Lorsque le système est composé d'agents égoïstes et rationnels, une des réponses à ce problème est la formation de coalitions. Ici, les agents forment des groupes, appelés coalitions, afin de réaliser conjointement les tâches qui ne peuvent pas être traitées individuellement. Cependant, la majorité des travaux sur la formation de coalitions font deux hypothèses fortes. La première est que les agents ont une connaissance parfaite a priori de l'intérêt à former une coalition, c'est-à-dire du gain obtenu en la formant. La seconde hypothèse est que ce gain est déterministe. Ces deux hypothèses ne semblent pas adaptées à des problèmes réels où le gain exact obtenu par une coalition n'est connu qu'a posteriori. De plus, si cette même coalition se reforme par la suite, ce gain n'a pas de raison d'être strictement le même, dû à des facteurs internes ou des externalités.

À titre d'exemple, considérons le cadre de l'*Industrie 4.0* où des robots de différentes natures doivent se coordonner. Les problématiques induites par ce problème sont nombreuses, comme celle des déplacements, de l'interaction planifiée et, ce qui nous intéresse ici, la constitution d'équipe en environnement incertain. En effet, certaines tâches ne pouvant être réalisées seul, les agents de cette industrie peuvent être amenés à former des groupes de manière répétée. Cependant, des facteurs internes et externes peuvent influer sur la qualité des résultats produits par ces groupes. Par exemple, un facteur interne peut être les compétences individuelles des agents dont les effets peuvent être stochastiques, couplées à leur capacité à mieux interagir avec certains agents plutôt qu'avec d'autres. Un facteur externe peut être un effet environnemental indépendant des agents, comme l'arrivée d'une sinistre dans les locaux de l'un d'entre eux. Il est donc approprié de relâcher ces deux hypothèses de déterminisme et connaissance a priori.

Cependant, cela soulève de nouvelles questions. Si les agents n'ont plus de connaissance a priori sur les coalitions, comment peuvent-ils en obtenir ? Si le gain produit par les coalitions est stochastique, comment peuvent-ils l'estimer ? Une façon de modéliser de tels jeux de coalitions aux hypothèses relâchées est d'utiliser des jeux répétés, qui permettent d'observer le résultat d'un même jeu séquentiellement. Ainsi, les agents peuvent observer l'état du jeu à différents moments et sont capables d'en extraire de l'information. Néanmoins, l'objectif principal de la formation de coalitions est de partitionner les agents en coalitions. Dans ce nouveau contexte où les utilités des coalitions sont stochastiques et inconnues des agents, comment décider des coalitions à former ? Ce choix peut être décomposé en plusieurs questions. Comment les agents peuvent-ils favoriser la formation d'une coalition plutôt qu'une autre en fonction de ce qu'ils savent d'elles ? À quel moment considèrent-ils qu'ils en savent suffisamment sur une coalition pour en évaluer correctement son utilité ? Comment les agents peuvent-ils décider collectivement des interactions qui pourraient être acceptées par tous ?

Cet article se place dans la continuité de nos travaux précédents publiés aux JFSMA, à savoir un état de l'art sur la formation de coalitions [13] et un protocole distribué de formation de coalitions classique (i.e. déterministe et non répété) [14]. Nous proposons alors dans cet article un nouveau concept de solutions pour la formation de coalitions stochastique répétée, fondé sur un principe exploration-exploitation, principe bien connu de l'apprentissage par renforcement. Pour cela, nous redéfinissons un concept de solutions existant, en y intégrant une notion d'intérêt à l'exploration afin de permettre aux agents de former des coalitions stables tout en obtenant plus d'information via un compromis entre l'exploitation d'un gain plutôt connu et l'exploration de coalitions dont le gain est inconnu ou très incertain. Nous montrons que notre concept de solutions est très efficace sur des fonctions caractéristiques non structurées, et est meilleur qu'une stratégie ϵ-glouton hormis dans le cas d'une fonction caractéristique très structurée.

Dans la suite, nous présentons en section 2 les jeux de coalitions et les travaux relatifs à la formation de coalitions stochastique répétée. Nous décrivons en section 3 un nouveau concept de solutions intégrant explicitement une notion d'exploration, ainsi qu'un exemple de définition de cette dernière. Enfin, la section 4 est consacrée aux résultats expérimentaux.

2 État de l'art

Dans cette section, nous présentons succinctement les fondements de la formation de coalitions classique ainsi que des approches stochastiques et répétées de ce problème. Pour une vue d'ensemble plus complète, nous renvoyons le lecteur vers l'article de synthèse [13].

2.1 Formation de coalitions classique

Quand les agents coopèrent, ils forment une *coalition*. Celle-ci produit une certaine *utilité*.

Definition 1 (Jeu de coalitions) *Un* jeu de coalitions *est un tuple* $\mathcal{G} = \langle N, v \rangle$ *où* $N = \{a_1, \ldots, a_n\}$ *est un ensemble d'agents, et* $v : 2^N \to \mathbb{R}$ *est la fonction caractéristique qui indique l'utilité* $v(C)$ *de chaque coalition* $C \subseteq N$.

Une partition des agents en coalitions est appelée une *structure de coalitions* et une *solution* à un jeu de coalitions est définie comme suit.

Definition 2 (Solution) *Une* solution *à un jeu de coalitions* $\mathcal{G}$ *est un tuple* $S_\mathcal{G} = \langle \mathcal{CS}, \vec{x} \rangle$ *où* $\mathcal{CS}$ *est une structure de coalitions de* N, $\vec{x} = \{x_1, \ldots, x_n\}$ *est un vecteur de gains pour les agents où* $x_i \geq 0$ *est le gain de l'agent* a_i.

Les agents étant supposés égoïstes, une solution doit être *acceptable* par tous les agents, c'est-à-dire qu'ils ne doivent pas vouloir former ou rejoindre une autre coalition (où ils gagneraient plus par exemple). Une telle solution est dite *stable*. Différents critères de stabilité existent et l'ensemble des solutions respectant un de ces critères est appelé un *concept de solutions*.

Intéressons-nous au concept du *cœur* et sa généralisation, l'ϵ-*cœur* [18, 20]. Le cœur est l'ensemble des solutions $\langle \mathcal{CS}, \vec{x} \rangle$ pour lesquels il n'existe aucune autre coalition qui pourrait être formée et qui produirait une utilité supérieure à la somme des gains de ses agents dans $\vec{x}$. Si le cœur peut être vide, il existe une variante non vide, appelée ϵ-cœur.

Definition 3 (ϵ-cœur) *Une solution* $(\mathcal{CS}, \vec{x})$ *appartient à l'*ϵ*-cœur si et seulement si :*

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

Cette variante autorise à réduire le gain des agents d'ϵ afin de rendre la solution stable. Ici, ϵ est la pire perte de gain parmi tous les agents qui assure la stabilité. Dans la suite, nous considérons uniquement l'ϵ-cœur, car toute solution à un jeu appartient à un ϵ-cœur pour un ϵ, et plus précisément, nous nous intéressons à l'ϵ-cœur ayant le plus petit ϵ pour lequel une solution existe, appelé le *dernier cœur* [9].

2.2 Fonctions caractéristiques stochastiques

Comme mentionné précédemment, le fait que l'utilité de la coalition soit déterministe n'est pas toujours réaliste. Dans la littérature, certains travaux proposent de relâcher cette contrainte et d'intégrer la stochasticité dans les fonctions caractéristiques [11, 8, 15]. Cependant, la nature de l'incertitude dans ces modèles diffère.

Par exemple, Ieong et Shoham ont proposé une distribution de probabilité sur des mondes représentant des jeux de coalitions, chacun d'entre eux ayant une fonction caractéristique déterministe [15]. Les vecteurs de gains sont alors des distributions de l'utilité des coalitions dans chaque monde possible, et les agents expriment des préférences sur ces distributions.

Chalkiadakis et Boutilier ont considéré une fonction caractéristique déterministe modélisée dans un environnement stochastique avec des processus de décision markoviens partiellement observables [8]. Les agents ont des croyances sur les capacités des autres et la même structure de coalitions peut conduire à différents états du monde. Ici, les auteurs considèrent un concept de stabilité appelé le noyau bayésien [7, 10].

Charnes et Granot considèrent simplement que la valeur d'une coalition est une variable aléatoire : [12]. La fonction caractéristique est alors réécrite comme $v : 2^N \rightarrow \mathcal{X}_{2^N}$. Ainsi, lorsqu'une coalition est formée, l'utilité produite est déterminée par la variable aléatoire, qui suit une loi normale. Dans ce modèle, ils calculent leurs vecteurs de gains en associant à chaque agent d'une coalition une part égale de l'espérance de la variable aléatoire associée à la coalition.

Dans cet article, l'objectif est que notre modèle puisse représenter des situations de stochasticité hétérogène. Par exemple, où l'incertitude pourrait provenir de la fiabilité des agents qui la composent pour une coalition, et être due à l'environnement pour une autre. Les deux premiers modèles présentés ci-dessus ne correspondent pas à notre exigence, puisque la stochasticité provient d'éléments particuliers spécifiques. Le modèle de Ieong et Shoham nécessite la définition de mondes dans lesquels les fonctions caractéristiques sont déterministes, et la nature de la stochasticité ne peut provenir que de l'incertitude sur le monde réel. Si cela nous permet de modéliser des fonctions caractéristiques stochastiques, cela ne correspond pas à une nature hétérogène de l'incertitude. En ce qui concerne le modèle de Chalkiadakis et Boutilier, il repose sur l'incertitude de l'environnement et les croyances sur les compétences des agents. Ce modèle est plus permissif que le précédent, mais pas encore assez général de notre point de vue car les compétences des agents sont toujours déterministes en elles-mêmes. C'est pourquoi nous nous positionnons dans la continuité des travaux de Charnes et Granot, ce qui nous permet de modéliser l'hétérogénéité de la stochasticité par l'utilisation d'une variable aléatoire.

2.3 Jeux de coalitions répétés

Si nous relâchons l'hypothèse de connaissance parfaite de la fonction caractéristique, qu'elle soit stochastique ou non, les agents ne savent rien a priori des utilités produites par les coalitions. Il devient alors intéressant d'utiliser un jeu répété [4]. En effet, dans un contexte de formation de coalitions, répéter le jeu permet d'observer les utilités des coalitions formées, d'estimer la fonction caractéristique et de trouver une solution optimale stable dans le temps.

Par exemple, Blankenburg *et al.*, Louati *et al.* et Bettinelli *et al.* ont proposé des modèle de jeux de coalitions répétés ainsi que des protocoles fondés sur des notions de confiance [6, 17] ou de proximité sociale [5] pour déterminer avec quels agents former des coalitions. Ces protocoles s'appuient sur des étapes similaires : (1) communications à propos des croyances sur les compétences, la similarité ou la confiance des agents ; (2) formation (parfois dynamique) des coalitions ; (3) paiement des agents si nécessaire avant l'achèvement de la tâche pour encourager le bon comportement ; (4) exécution et évaluation des tâches ou des services. De cette façon, les agents apprennent un degré de fiabilité ou de pertinence envers les autres en observant les utilités des coalitions auxquelles ils participent. Toutefois, il s'agit d'approches *gloutonnes* qui ne cherchent pas à former des coalitions pour obtenir plus d'information.

Nous pouvons également mentionner à nouveau Chalkiadakis et Boutilier [8] dont le modèle est

basé sur un processus d'apprentissage par renforcement bayésien, qui peut être vu comme un jeu répété. À chaque pas de temps, les agents doivent choisir de former une coalition. Une fois les coalitions formées, ils prennent une action qui provoque une transition stochastique à un état du monde. Les agents observent ce qu'il s'est passé et mettent à jour leurs croyances. Cela leur permet d'apprendre les compétences des autres agents mais aussi le modèle de transition stochastique entre les états.

Konishi et Debraj [16] ont proposé un processus répété de formation de coalitions dans le but d'étudier les équilibres dans ces processus. À chaque pas de temps, les coalitions doivent choisir une action commune à réaliser. Elles peuvent ne rien faire, ou se déplacer dans un état autorisé (qui correspond à la description d'une structure de coalitions, y compris le vecteur de gains). Les états autorisés d'une coalition, appelés *mouvements coalitionnels*, sont limités à des mouvements Pareto-efficaces afin de permettre la convergence.

Les deux premiers modèles sont intéressants mais sont principalement axés sur l'apprentissage des compétences ou des fiabilités individuelles influant sur une fonction caractéristique déterministe. Le troisième modèle est également intéressant mais la contrainte de mouvements Pareto-efficaces autorise les agents à faire des mouvements irrationnels. Par conséquent, nous allons plutôt nous éloigner de ces modèles pour apprendre une fonction caractéristique stochastique telle que définie par Charnes et Granot. Dans les sections suivantes, nous présentons notre modèle ainsi qu'un concept de solutions intégrant une notion d'intérêt à l'exploration, puis un protocole expérimental pour le mettre en pratique, et enfin quelques résultats.

3 Un ϵ-cœur avec exploration

Afin de répondre à notre problématique, nous définissons en premier lieu un modèle de jeux de coalitions stochastiques répétés. Nous présentons un concept de solutions intégrant un équilibre exploration-exploitation, principe souvent utilisé dans les bandits manchots. En effet, dans les deux cadres, les objets de la décision (respectivement les coalitions [1] et les bras des bandits) génèrent un gain (ou une utilité) lorsque choisis, ce gain dépendant d'une distribution de probabilité inconnue des agents. Une différence néanmoins est que dans les bandits manchots, un seul bras est tiré à chaque pas de temps, tandis que dans la formation de coalitions, cela concerne (potentiellement) plusieurs coalitions, et toutes ne peuvent pas être formées en même temps. Cet équilibre est défini de manière générique afin de prendre en compte différentes caractérisations d'intérêt, par exemple un biais d'exploration qui décrit un intérêt à l'information, ou bien un biais de réputation, qui décrit un intérêt lié à la fiabilité des agents.

1. Nous pourrions envisager de considérer un bras non pas par coalition mais par structure de coalitions. Toutefois, cela conduirait à devoir envisager un nombre de bras très supérieur aux $2^{|N|}$ coalitions.

3.1 Jeux de coalitions stochastiques répétés

Le modèle de jeux de coalitions stochastiques répétés que nous définissons est inspiré des travaux de Charnes et Granot.

Definition 4 (RSCG) *Soit* $\mathcal{G} = \langle N, \mathbb{T}, v, \hat{v}, i \rangle$ *un* jeu de coalitions stochastique répété *où :*
- $N = \{a_1 \dots a_n\}$ *est un ensemble d'agents,*
- $\mathbb{T} \subset \mathbb{N}^+$ *est un ensemble de pas de temps distincts,*
- $v : 2^N \rightarrow \mathcal{X}^{2^N}$ *est une fonction caractéristique qui à chaque coalition associe une variable aléatoire. Pour une coalition* $C \subseteq 2^N$ *donnée, nous notons* $v(C) = \mathcal{X}^C$*. Cette fonction caractéristique est inconnue des agents.*
- $\hat{v} : 2^N \times \mathbb{T} \rightarrow \hat{\mathcal{X}}^{2^N}$ *est une fonction caractéristique qui associe à chaque coalition au pas de temps* $t \in \mathbb{T}$ *une estimation de l'utilité.*
- $i : 2^N \times \mathbb{T} \rightarrow \mathbb{R}$ *une fonction d'intérêt qui à chaque coalition associe un intérêt quantitatif à un pas de temps donné. Nous notons* $i(C, t)$ *l'intérêt de la coalition* C *à un pas de temps* t*.*

À chaque pas de temps, les agents de N doivent décider d'une solution au jeu, malgré le fait qu'ils ne connaissent pas a priori la fonction caractéristique v. Une solution est, comme dans un contexte déterministe, un tuple constitué d'une structure de coalitions et d'un vecteur de gains. Cependant, le gain des agents est un gain estimé *ex ante* basé sur ce qu'ils savent de v.

Definition 5 (Solution à un RSCG) *Une* solution S^t *à un pas de temps* $t \in \mathbb{T}$ *à un jeu* $\mathcal{G}$ *est un tuple* $S^t = \langle \mathcal{CS}^t, \vec{x}^t \rangle$ *tel que :*
- $\mathcal{CS}^t$ *est une structure de coalitions de* N*,*
- $\vec{x}^t = \{x_1^t, \dots, x_n^t\}$ *est un vecteur de gains tel que* $x_i^t \geq 0$ *est le gain de l'agent* a_i *calculé selon l'utilité estimée de la coalition à laquelle il appartient dans la structure* $\mathcal{CS}^t$*.*

Nous proposons d'adapter le concept de solutions de l'ϵ-cœur en considérant que la valeur d'une coalition, c'est-à-dire son intérêt à être formée à un pas de temps donné, dépend de deux éléments : une estimation de son utilité dont les gains des agents sont directement dérivés, et d'un intérêt que les agents ont de la former afin d'obtenir plus d'informations sur son utilité réelle.

Comme mentionné ci-dessus, le gain des agents pour une solution donnée est une estimation. Une fois la solution trouvée et les coalitions formées, les utilités réelles produites par les coalitions sont le résultat de processus stochastiques paramétrés par la fonction caractéristique. Nous supposons que ces utilités sont observées par tous les agents. Notons X_t^C l'observation de l'utilité produite par la coalition C au pas de temps t.

Definition 6 (Observations) *Soit $\mathcal{O}_t$ un ensemble d'observations au pas de temps t correspondant à l'ensemble des coalitions formées à chaque pas de temps avant t et leurs utilités réelles produites :*

$$\mathcal{O}_t = \{(C, t', X_{t'}^C) : C \subseteq 2^N, t' \in \mathbb{T}, t' < t\}$$

Par la suite, notons $\mathcal{O}_t(C)$ l'ensemble des observations au pas de temps t associé à la coalition $C \subseteq 2^N$. Cet ensemble d'observations permet de mettre à jour les connaissances des agents sur la fonction caractéristique.

Dans la suite, nous faisons l'hypothèse que les agents estiment l'utilité des coalitions comme des lois normales. Ainsi, pour une coalition donnée $C \subseteq 2^N$, $\hat{v}(C, t)$ est caractérisé par la espérance et la variance d'une loi normale sur l'ensemble des observations.

Definition 7 (Estimation de l'utilité) *Au pas de temps $t \in \mathbb{T}$ et pour la coalition $C \subseteq 2^N$, la valeur estimée de C, $\hat{v}(C, t)$, est donnée par $\hat{\mu}(C, t)$ son espérance et $\hat{\sigma}^2(C, t)$ sa variance, calculées à partir des observations $\mathcal{O}_t(C)$.*

L'incertitude portant sur l'utilité produite par les coalitions une fois formées, une solution doit tenir compte de cette incertitude pour être stable.

3.2 Intérêt des coalitions

La nature exacte de l'intérêt que les agents ont pour une coalition peut dépendre du problème. Toutefois, cet intérêt a pour objectif de rendre possible l'exploration d'autres solutions potentiellement intéressantes pour les agents mais qui pourraient être jugées non stables au sens d'un concept de solutions classique.

Remarquons que dans la formation de coalitions, nous devons comparer des structures de coalitions, ce qui implique donc de comparer des coalitions différentes. Par exemple dans le cœur, vérifier la stabilité d'une solution consiste à comparer l'utilité d'une coalition à la somme des gains individuels des agents de cette même coalition où qu'ils soient dans la solution. Nous devons considérer une forme d'intérêt nous permettant de telles comparaisons, c'est-à-dire calculer à partir de l'intérêt d'une coalition l'intérêt individuel des agents qui la composent.

Definition 8 (Intérêt individuel) *L'*intérêt individuel $i_j(C_j, t)$ *d'un agent $a_j \in C_j$ à un pas de temps t est :*

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

Cette répartition égalitaire est une des nombreuses façons de distribuer l'intérêt et, contrairement à la valeur de Shapley, représente le fait que chaque agent d'une coalition possède le même intérêt pour cette coalition, indépendamment des autres coalitions auxquelles ils peuvent appartenir. De plus, plus une coalition contient d'agents, plus leur intérêt individuel sera faible. Cette répartition va donc avantager les coalitions de faible cardinalité, car plusieurs observations distinctes peuvent rapporter plus d'information qu'une seule. Cet intérêt individuel permet de définir l'intérêt d'une coalition au regard d'une structure de coalitions donnée, indépendamment du fait que les agents de cette coalition soient ensemble ou non dans la structure.

Definition 9 (Intérêt collectif) *L'*intérêt collectif $i^{\mathcal{CS}}(C, t)$ *des agents d'une coalition C au regard d'une structure de coalitions $\mathcal{CS}$ à un pas de temps t est :*

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

où $C_j^{\mathcal{CS}}$ est la coalition de a_j dans $\mathcal{CS}$ et C n'appartient pas nécessairement à $\mathcal{CS}$.

3.3 λ-cœur

Afin d'intégrer cet intérêt des coalitions au concept de solutions, nous devons l'agréger à

l'utilité. Pour rester générique dans un premier temps, nous considérons de manière abstraite un opérateur d'agrégation noté $\oplus$. Selon la nature exacte de l'intérêt, cet opérateur peut prendre différentes formes, par exemple une *addition*, une *multiplication* ou encore un *maximum*.

Les différents éléments décrivant l'intérêt des agents étant définis, nous pouvons désormais construire notre concept de solutions, le λ-cœur, fondé sur un principe exploration-exploitation. Pour cela, nous adaptons le concept de l'ϵ-cœur en intégrant l'opérateur de composition ainsi que la fonction d'intérêt. Nous ajoutons donc d'un côté de l'inéquation, l'intérêt d'une coalition à l'espérance d'utilité de la coalition, et de l'autre côté l'intérêt collectif à la somme des gains des agents au regard de la solution considérée.

Definition 10 *Une solution $\langle \mathcal{CS}^t, \vec{x}^t \rangle$ appartient au λ-cœur si et seulement si $\forall\, C \subseteq N$:*

$$x^t(C) \oplus i^{\mathcal{CS}^t}(C,t) \geq \hat{\mu}(C,t) \oplus i(C,t) - \lambda$$
$$\textit{avec } x^t(C) = \sum_{a_i \in C} x_i^t$$

De manière similaire à l'ϵ-cœur, le dernier cœur pour ce concept du λ-cœur est défini comme étant celui ayant le plus petit λ pour lequel une solution existe. Nous pouvons désormais proposer un exemple d'instanciation de ce concept de solutions en définissant l'intérêt comme un biais d'exploration, et l'opérateur d'agrégation comme étant une *addition*.

3.4 Exemple d'intérêt : biais d'exploration

Une notion d'intérêt pertinente est celle de l'exploration, que nous retrouvons dans le problème des bandits manchots. Pour ce problème, de nombreuses stratégies ont été proposées, et notamment les stratégies fondées sur une *borne supérieure de confiance* appelée *UCB* (pour *Upper Confidence Bound*) [1]. Parmi les stratégies fondées sur ce principe, il existe *UCB-V*, qui a été proposée pour le problème des bandits manchots par Audibert *et al.* [2]. Celle-ci décrit un biais d'exploration prenant en compte la variance des distributions de probabilité sous-jacentes des bras du bandit manchot a été prouvée plus efficace que la stratégie *UCB-1* [3]. Nous adaptons donc *UCB-V* pour l'appliquer aux jeux de coalitions stochastiques répétés.

Definition 11 (Biais d'exploration UCB-V) *Le biais d'exploration* UCB-V *pour une coalition C à un pas de temps t est défini comme suit :*

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

Certaines constantes doivent être définies. La constante b définit la borne supérieure des gains du problème, cela est donc dépendant de ce dernier. Cependant, nous pouvons faire l'hypothèse que les utilités sont normalisées sur l'intervalle $[0,1]$ comme dans les bandits manchots, et ainsi définir $b = 1$. Les constantes ζ et c sont des paramètres de contrôle de l'exploration (en particulier ζ). Nous reprenons ici les valeurs de l'article originel, dans lequel Audibert *et al.* montrent l'efficacité de ces constantes lorsqu'elles sont donc définies comme $\zeta = 1,2$ et $c = 1$.

4 Expérimentations

Nous évaluons de manière empirique les performances de notre concept de solutions.

4.1 Protocole expérimental

Dans un premier temps, nous construisons 200 couples de jeux différents avec des fonctions caractéristiques uniques, pour 6 agents. Chaque couple de jeux est construit avec deux structures de fonctions caractéristiques différentes. La première fonction caractéristique est tirée selon le modèle NDCS (*Normally Distributed Coalition Structures)* [19]. Ce modèle permet de construire des fonctions caractéristiques structurées, mais sans contraindre fortement le modèle comme avec des structures monotones ou superadditives [9]. Ainsi, l'espérance d'utilité μ_C de chaque coalition $C \subseteq N$ est tirée selon une loi normale $\mathcal{N}(|C|, \sqrt{|C|})$. La fonction caractéristique est ensuite normalisée sur l'intervalle $[0,1]$. La deuxième fonction caractéristique n'est pas structurée, car tirée aléatoirement de façon uniforme pour chaque coalition. Ainsi, l'espérance d'utilité μ_C de chaque coalition $C \subseteq N$ est tirée selon une loi uniforme $\mathcal{U}(0,1)$. Dans les deux modèles de structuration, les variances σ_C^2 de chaque coalition C sont tirées selon la loi uniforme $\mathcal{U}(0, \frac{\mu_C}{2})$. Chaque fonction caractéristique est ensuite normalisée sur l'intervalle $[0,1]$.

Dans un second temps, afin de créer une série de jeux des plus aux moins structurés, pour

chaque couple de jeux, nous créons des jeux intermédiaires à l'aide d'une transformation linéaire en appliquant un facteur de transformation $w \in [0, 1]$. Ainsi, un facteur de transformation de 0 correspond au jeu structuré NDCS, tandis que le facteur 1 correspond au jeu structuré aléatoirement. Un jeu est créé par pas de 0.05 pour w entre les deux jeux du couple, ce qui correspond donc à 19 jeux intermédiaires supplémentaires. Notre concept de solutions est donc évalué sur 4200 jeux et sur 100 pas de temps chacun.

Exemple 1 *Soient C et C' deux coalitions, et v_1 et v_2 deux fonctions caractéristiques respectivement structurées aléatoirement et NDCS :*

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

Pour un facteur de transformation de 0.4, les utilités de C et C' sont telles que :

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

Pour un facteur de transformation de 1, la fonction caractéristique résultante est v_1 :

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

Enfin, pour un facteur de transformation de 0, la fonction caractéristique résultante est donc v_2 :

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

Ces jeux sont également joués avec la stratégie ϵ-glouton, qui est une stratégie de référence dans le domaine des bandits manchonts [21]. Elle décrit également un équilibre exploration-exploitation, en explorant aléatoirement avec la probabilité ϵ, et en exploitant avec la probabilité $1 - \epsilon$. Dans notre implémentation, l'exploitation consiste à utiliser le concept du dernier cœur avec une valeur ϵ utilisée de 0.05.

4.2 Mesures de performance

La première mesure est le *regret cumulé*. Celui-ci mesure l'évolution du regret instantané (c'est-à-dire la différence entre le bien-être social maximal [2] du jeu et la somme des utilités réelles espérées des coalitions de la structure formée au pas de temps t) au cours du temps. Cela permet de mettre en lumière la convergence du regret, c'est-à-dire le pas de temps à partir duquel les stratégies ont atteint leur équilibre exploration-exploitation et produisent donc un regret instantané constant. À un pas de temps t, le regret cumulé est la somme des regrets instantanés de chaque pas de temps $t' \leq t$. L'objectif de cette mesure est de pouvoir observer les différences entre les stratégies et non la valeur absolue. Formellement :

Definition 12 (Regret cumulé) *Soit la solution optimale $S^* = (\mathcal{CS}^*, \vec{x}^*)$ au sens du bien-être social, le regret cumulé à un pas de temps t, noté R^t_c, est défini tel que :*

$$R^t_c = \sum_{t'=0}^{t} \left(\sum_{C^* \in \mathcal{CS}^*} \mu_{C^*} - \sum_{C \in \mathcal{CS}^t} \mu_C \right)$$

Enfin, afin d'évaluer l'apprentissage que les agents font de la fonction caractéristique réelle au cours du temps, nous utilisons l'*erreur moyenne absolue* (MAE) sur les utilités estimées et réelles des coalitions. Plus la MAE est proche de 0, plus la fonction caractéristique estimée est précise. La MAE est définie telle que :

Definition 13 (Erreur moyenne absolue) *Soient v et $\hat{v}$ deux fonctions caractéristiques, l'*erreur moyenne absolue *(MAE) D^t_{MAE} entre v et $\hat{v}$ au pas de temps t est définie telle que :*

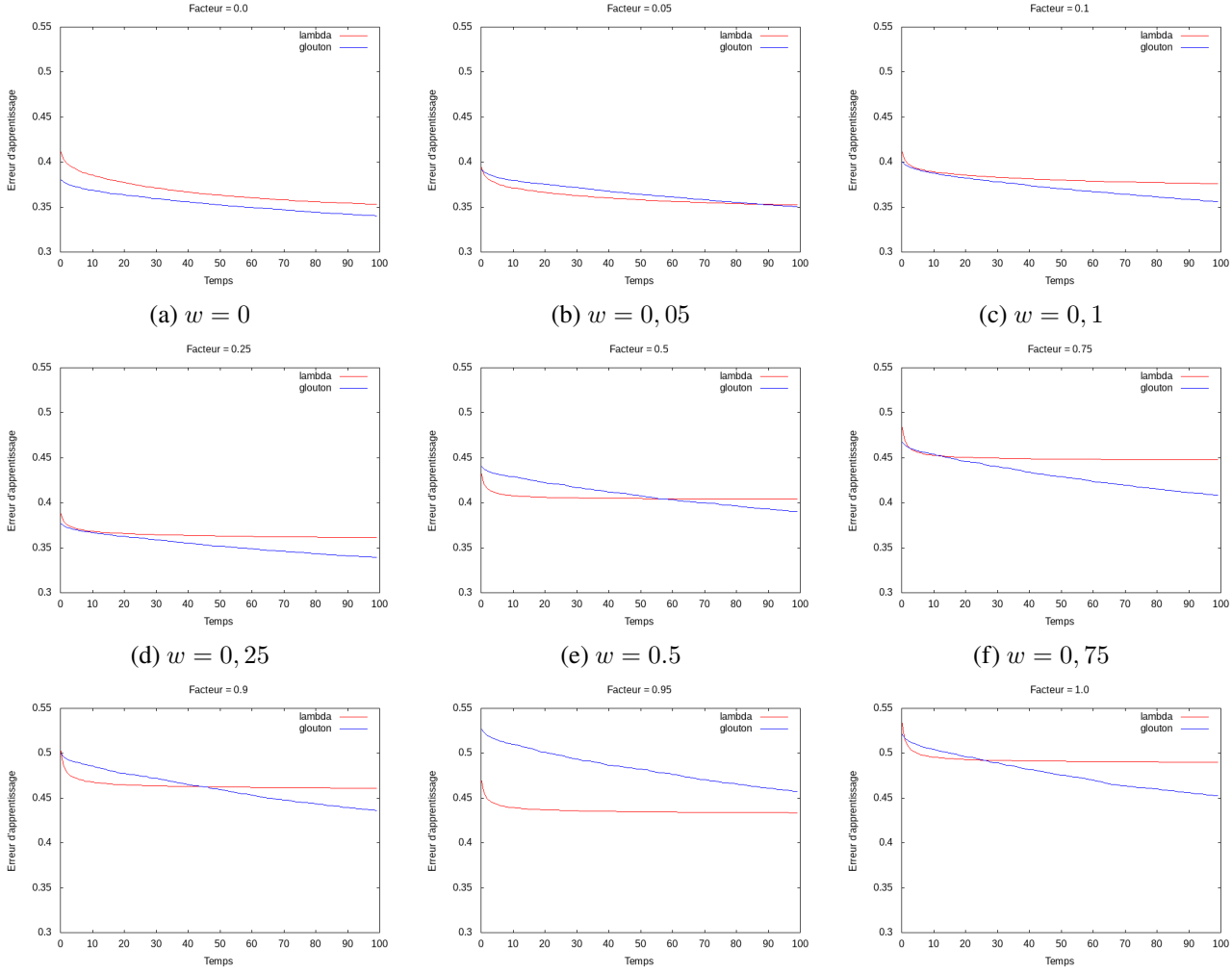
$$D^t_{MAE} = \frac{\sum_{C \in 2^N} |\hat{\mu}(C, t) - \mu_C|}{|2^N|}$$

4.3 Résultats

Les figures 1 et 2 montrent respectivement l'évolution des moyennes de l'erreur d'apprentissage et du regret cumulé de l'ensemble des jeux pour une configuration donnée (c'est-à-dire un facteur de transformation linéaire w) au cours des 100 pas de temps. La figure 3 synthétise les résultats avec le pourcentage relatif d'efficacité du λ-cœur contre ϵ-glouton pour les différents facteurs.

Concernant l'erreur d'apprentissage en figure 1, un premier point à souligner est que plus la fonction caractéristique est structurée (donc plus le facteur de transformation w est proche de 0), moins l'erreur d'apprentissage est grande. De manière générale, la stratégie ϵ-gloutonne est

2. Donné par la structure de coalitions qui maximise la somme des utilités de ses coalitions.



(a) $w = 0$ (b) $w = 0,05$ (c) $w = 0,1$

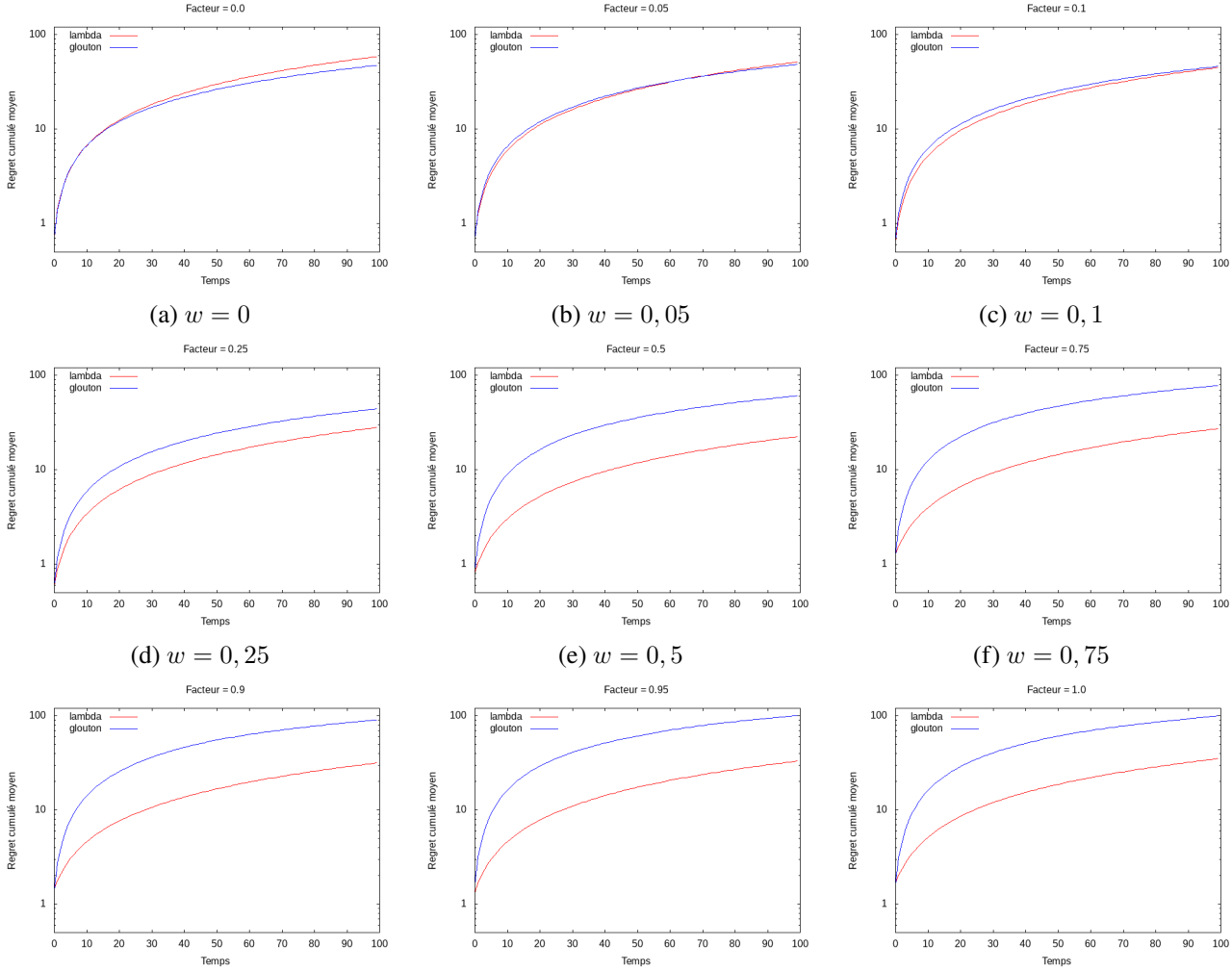
(d) $w = 0,25$ (e) $w = 0.5$ (f) $w = 0,75$

(g) $w = 0,9$ (h) $w = 0,95$ (i) $w = 1$

FIGURE 1 – Erreur d'apprentissage moyenne pour 6 agents sur 4200 jeux

celle qui apprend le mieux, avec quelques exceptions comme pour $w = 0,95$ où le λ-cœur permet d'apprendre mieux, ou bien $w = 0,05$ où les résultats des deux méthodes sont très proches. Toutefois, nous pouvons voir graphiquement une différence de comportement entre elles selon la structuration des fonctions caractéristiques. En effet, plus les fonctions caractéristiques sont structurées, plus la stratégie ϵ-gloutonne apprend entre le début et la fin des expérimentations. Par exemple, son erreur d'apprentissage diminue respectivement de $10,59\%$, $11,45\%$ et de $13,20\%$ pour $w = 0$, $w = 0,5$ et $w = 1$. Remarquons que cette diminution est quasi-linéaire avec la variation du facteur w. Concernant le λ-cœur, nous pouvons voir que l'apprentissage converge rapidement, dû au terme d'exploration *UCB-V*, et cela de plus en plus rapidement à mesure que les fonctions caractéristiques sont déstructurées. Par exemple, pour $w = 0$, l'erreur d'apprentissage diminue tout au long de l'expérimentation, tandis que pour $w = 1$, l'erreur cesse quasiment de diminuer après le pas de temps $t = 20$. D'un point de vue plus général pour les deux méthodes, plus les fonctions caractéristiques sont déstructurées, plus l'erreur d'apprentissage est initialement importante.

Intéressons-nous ensuite au regret cumulé moyen en figure 2. Pour un facteur de transformation $w = 0$, c'est-à-dire avec une structure purement NDCS, le regret cumulé moyen est en faveur de la stratégie ϵ-gloutonne, tout comme pour un $w = 0,05$. Toutefois, à partir de $w = 0,1$, le λ-cœur est plus performant en termes de regret, et l'écart est plus important pour des valeurs de w plus grande. À partir de ces résultats, nous pouvons déduire que la stratégie ϵ-gloutonne est performante sur des fonctions



(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) $w = 1$

FIGURE 2 – Regret cumulé moyen pour 6 agents sur 4200 jeux

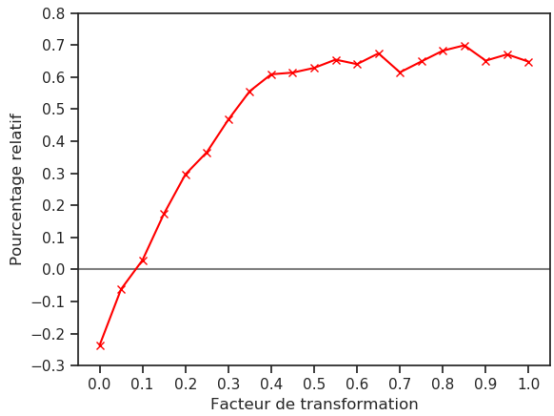


FIGURE 3 – Pourcentage relatif moyen d'efficacité de λ-cœur contre la stratégie ϵ-gloutonne

caractéristiques structurées mais que moins il y a de structuration, moins elle l'est. Il faut toutefois souligner que lorsque la stratégie ϵ-gloutonne est surpassée par le λ-cœur, c'est principalement ce dernier qui gagne en performance plus que la stratégie ϵ-gloutonne n'en perd. En effet, cette dernière obtient un regret cumulé moyen de $45,91$ pour $w = 0$ et de $47,13$ pour $w = 0,1$, soit une différence de $1,22$. De son côté, le λ-cœur obtient un regret cumulé moyen de $58,28$ pour $w = 0$ et de $44,58$ pour $w = 0,1$, soit une différence de $13,70$. Cet écart est de $16,07$ pour $w = 0,25$, de $37,91$ pour $w = 0,5$, de $35,11$ pour $w = 0,75$, jusqu'à une différence de $64,39$ pour $w = 1$. Dans ce dernier cas, le regret cumulé moyen pour le λ-cœur est de $34,98$ alors qu'il est de $99,37$ pour la stratégie ϵ-gloutonne. L'efficacité relative de λ-cœur contre la stratégie ϵ-gloutonne est mise en lumière sur la figure 3. Sur cette dernière, nous pouvons voir que l'écart

en faveur du λ-cœur ne fait qu'augmenter jusqu'à $w = 0,4$ puis se stabilise. Pour $w = 0$, λ-cœur est $23,66\%$ moins efficace que ϵ-glouton. Il devient $2,9\%$ plus efficace à partir de $w = 0,1$, jusqu'à $60,87\%$ pour $w = 0,4$. Ensuite, pour $w \geq 0,4$, l'efficacité relative en faveur de λ-cœur se stabilise autour de 65%, avec un maximum de $69,90\%$ pour $w = 0,85$. Ainsi, le concept de solutions λ-cœur se montre très performant sur des fonctions caractéristiques déstructurées, et reste plus performant que la stratégie ϵ-gloutonne tant que la structuration n'est pas importante. Il est toutefois nécessaire de noter que le λ-cœur est plus efficace sur des fonctions caractéristiques légèrement structurées. Par exemple, il obtient un regret cumulé de $22,39$ avec $w = 0,5$, tandis que pour $w = 1$ celui-ci est de $34,98$ (avec un minimum pour $w = 0,45$ avec $21,94$ de regret cumulé).

5 Conclusion

Nous avons proposé le concept de solutions λ-cœur fondé sur un équilibre exploration-exploitation par l'intégration d'une notion d'intérêt pour les agents. En fixant cet intérêt à un biais d'exploration *UCB-V* et définissant la composition comme étant une addition, nous avons montré que ce concept de solutions est performant sur les jeux de coalitions stochastiques répétés, notamment lorsque les fonctions caractéristiques ne sont pas très fortement structurées. Toutefois, le calcul du λ-cœur est long en raison de biais d'exploration. En effet, ce biais conduit le dernier cœur à avoir une valeur λ élevée, et donc à davantage parcourir l'espace des solutions car une approche naïve de ce calcul consiste à chercher des λ-cœur en incrémentant itérativement la valeur de λ. Il serait pertinent de travailler sur une approche distribuée ou décentralisée, mais qui introduit une nouvelle problématique : celle de l'observabilité partielle, propre à chaque agent, de la fonction caractéristique.

Références

[1] R. Agrawal. 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] J.Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19) :1876–1902, 2009.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2) :235–256, 2002.

[4] J.-P. Benoit and V. Krishna. Finitely repeated games. *Foundations in Microeconomic Theory*, pages 195–212, 1984.

[5] M. Bettinelli, M. Occello, and D. Genthial. ABSG : un modèle d'agent socialement inspiré pour la formation de coalitions. In *29e JFSMA*, pages 53–60, 2021.

[6] B. Blankenburg, R. K. Dash, S. D. Ramchurn, M. Klusch, and N. R. Jennings. Trusted kernel-based coalition formation. In *4th AAMAS*, pages 989–996, 2005.

[7] G. Chalkiadakis and C. Boutilier. Bayesian reinforcement learning for coalition formation under uncertainty. In *3rd AAMAS*, pages 1090–1097, 2004.

[8] G. Chalkiadakis and C. Boutilier. Sequential decision making in repeated coalition formation under uncertainty. In *7th AAMAS*, pages 347–354, 2008.

[9] G. Chalkiadakis, E. Elkind, and M. Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6) :1–168, 2011.

[10] G. Chalkiadakis, E. Markakis, and C. Boutilier. Coalition formation under uncertainty : Bargaining equilibria and the Bayesian core stability concept. In *6th AAMAS*, pages 1–8, 2007.

[11] A. Charnes and D. Granot. Prior solutions : Extensions of convex nucleus solutions to chance-constrained games. Technical report, Texas Univ., 1973.

[12] A. Charnes and D. Granot. Coalitional and chance-constrained solutions to n-person games. i : The prior satisficing nucleolus. *SIAM J. Appl. Math.*, 31(2) :358–367, 1976.

[13] J. Guéneron and G. Bonnet. De la diversité des jeux de coalitions à utilité transférable. In *29e JFSMA*, pages 149–158, 2021.

[14] J. Guéneron and G. Bonnet. Un protocole de concessions monotones pour la formation distribuée de coalitions. In *30e JFSMA*, pages 31–40, 2022.

[15] S. Ieong and Y. Shoham. Bayesian coalitional games. In *23rd AAAI*, pages 95–100, 2008.

[16] H. Konishi and D. Ray. Coalition formation as a dynamic process. *Journal of Economic theory*, 110(1) :1–41, 2003.

[17] A. Louati, J. El Haddad, and S. Pinson. Formation de coalitions pour une composition de services web fondée sur la confiance dans les réseaux sociaux. In *25e JFSMA*, pages 149–158, 2017.

[18] R. Mochaourab and E. A. Jorswieck. Coalitional games in miso interference channels : Epsilon-core and coalition structure stable set. *IEEE Transactions on Signal Processing*, 62(24) :6507–6520, 2014.

[19] T. Rahwan, S. D. Ramchurn, N. R. Jennings, and A. Giovannucci. An anytime algorithm for optimal coalition structure generation. *Journal of Artificial Intelligence Research*, 34 :521–567, 2009.

[20] L. S. Shapley and M. Shubik. Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica : Journal of the Econometric Society*, pages 805–827, 1966.

[21] L. Tran-Thanh, A. Chapman, E. M. De Cote, A. Rogers, and N. R. Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *24th AAAI*, pages 1211–1216, 2010.