Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 1 Capitole (UT1 Capitole)

Présentée et soutenue par :
Xavier Venel
Le jeudi 12 juillet 2012

Titre :

Existence de la Valeur uniforme dans les Jeux répétés

ED MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de recherche :
GREMAQ - TSE

Directeur(s) de Thèse :

Jérôme RENAULT	Professeur	Université Toulouse 1 Capitole

Rapporteurs :

Eilon SOLAN	Professeur	Université de Tel-Aviv
Nicolas VIEILLE	Professeur	H.E.C.

Autre(s) membre(s) du jury :

Pierre CARDALIAGUET	Professeur	Université Paris Dauphine
Sylvain SORIN	Professeur	Université Paris VI
Stéphane VILLENEUVE	Professeur	Université Toulouse 1 Capitole

À mes parents,

Remerciements.

Je tiens tout d'abord à remercier mon directeur de thèse Jérôme Renault. Je le remercie d'avoir accepté d'encadrer mon travail. J'ai apprécié l'enthousiasme qu'il a manifesté depuis notre premier rendez-vous dans son bureau à Dauphine, puis plus tard sa disponibilité et son attention, en particulier lorsque je l'ai suivi à Toulouse. Finalement je le remercie pour sa patience et ses très nombreuses explications qui m'ont beaucoup aidé.

Je remercie Eilon Solan et Nicolas Vieille d'avoir accepté d'être les rapporteurs de cette thèse et d'être venus assister à la soutenance. Je leur suis reconnaissant de s'être montré particulièrement intéressés lors de mes présentations en conférence et surtout du temps qu'ils ont consacré à leur travail de rapporteur. Ils sont tous les deux les auteurs, de nombreux articles cités dans cette thèse et je suis honoré de les avoir comme rapporteurs. Je remercie également Pierre Cardaliaguet, Sylvain Sorin et Stéphane Villeneuve qui ont accepté de faire partie de mon jury.

Je tiens particulièrement à remercier Sylvain Sorin qui m'a fait découvrir la théorie des jeux en Master 2 et qui m'a présenté à Jérôme lorsque j'ai décidé de commencer une thèse. Il m'a toujours accueilli chaleureusement au sein du laboratoire de l'équipe Combinatoire et Optimisation et a beaucoup contribué au développement de mon intérêt pour la théorie des jeux.

J'adresse toutes mes amitiés aux nombreux chercheurs rencontrés en conférences qui m'ont accordé de leur temps entre autre Jérome Bolte, Janòs Flesch, Hugo Gimbert , Marc Quincampoix, Marco Scarsini, Tristan Tomala, ainsi que beaucoup d'autres qui ne sont pas nommément cités ici aujourd'hui.

Je tiens aussi à remercier Stéphane Loisel et Rida Laraki qui m'ont accueilli lors de stages de recherche et qui m'ont donné l'envie de poursuivre dans cette voie.

Je remercie tous les chercheurs de l'équipe du groupe MAD que j'ai cotoyés en séminaire, au déjeuner et pendant mes pauses café sans café. Je remercie en particulier ceux pour qui j'ai animé des TDs. Adrien Blanchet, Clément Bruche, Olivier Faugeras et Jean Paul Décamps. Ce dernier était en première ligne lors de ma première année en tant que chargé de TDs, je le remercie de tout mon coeur pour son soutien et ses conseils.

Je me suis beaucoup déplacé pendant cette thèse commencée à moitié à l'École Polytechnique et à moitié à Paris 6, et terminée à l'université Toulouse 1 Capitole au sein du GREMAQ. Je tiens à remercier toutes les équipes d'encadrement pour leur accueil, en particulier les secrétaires Corinne Vella et Aline Soulié et la secrétaire de l'école doctorale EDMITT, Agnès Requis. Je remercie également Benedicte Alziary, en tant que directrice du département de mathématiques et la secrétaire de l'école doctorale d'économie Aude Schoesling pour l'aide qu'elles

m'ont apportée à mon arrivée à Toulouse.

Je remercie tout le groupe des thésards en théorie des jeux que j'ai toujours plaisir à rencontrer en conférence à Paris, à Toulouse ou ailleurs et dont les nombreuses questions, en particulier lors du séminaire des thésard à Paris, m'ont fait progresser : Antoine, Cheng, Fabien, Guillaume, Luis, Maël, Marie, Mario, Mathieu, Miquel, Pablo, Vianney, Miquel et Xiaoxi. Une mention spéciale pour Fabien pour nos nombreuses discussions au cours de cette dernière année de thèse quand il a rejoint Toulouse.

Je remercie aussi tous les thésards que j'ai vu passer dans le bureau MF007 et ceux qui n'y seront qu'à partir de l'année prochaine, pour tous les bons moments passés ensemble : Ahmat, Alice, Anastasia, Antonio, Aurelija, Baptiste, Chiara, Caspar, Elodie, Fanny, Juan-Felipe, Julia, Janna, Kyriacos, Loic, Loic, Marc, Nicolas, Olga, Paula, Renaud, Samuele, Thibault, Vanessa.

Merci aussi à mes amis de Cachan pour leur amitié et leur soutien depuis 7 ans : Aude et Antoine, Camille et Antoine, Florent, Barbara et Gregory, Guillaume, Julien et Nassradine.

Merci également à Jean-Philippe pour nos longues discussions, j'espère qu'on trouvera un article à écrire ensemble.

Enfin je souhaite remercier ma famille pour leur soutien permanent tout au long de ces quatres années, mes parents, mes frères et soeurs et leur famille, mon cousin et même mon petit neveu, Thomas, qui m'a fait réviser les entiers naturels. C'est très important pour moi de les voir présents ici, aujourd'hui.

Table des matières

Chapitre 1

Introduction

1.1 Introduction

Les jeux stochastiques à deux joueurs et à somme nulle ont été introduits par Shapley en 1953 [Sha53] afin d'étudier les interactions répétées entre plusieurs joueurs.

À chaque étape, les joueurs prennent des décisions qui génèrent un paiement courant et influent aléatoirement sur l'évolution de l'état qui définit leur environnement. Ensuite ils observent les décisions prises par tous les joueurs et le nouvel état qui en résulte. Les deux joueurs ont des objectifs opposés : ce que gagne le joueur 1 est perdu par le joueur 2, et réciproquement. Ainsi le joueur 1 cherche à maximiser son paiement alors que le joueur 2 cherche à minimiser le paiement du joueur 1 (et ainsi maximiser le sien). Les joueurs doivent choisir leurs décisions afin d'obtenir un bon paiement à chaque étape mais aussi s'assurer qu'ils pourront avoir de bons paiements dans le futur.

La première question est de définir comment les joueurs évaluent une suite de paiements. Si l'on considère un jeu avec un seul joueur, appelé Processus de Decision Markovien (MDP), différents critères d'évaluation sont utilisés dans la littérature et pour chaque critère, on appelle valeur le paiement maximum obtenu : la valeur escomptée (notée v_λ), la valeur des jeux avec un nombre fini d'étapes (notée v_n), la valeur *limsup*. Les chercheurs se sont posés de nombreuses questions sur le lien entre ces différentes notions. Est-ce que les valeurs des jeux avec un nombre fini d'étapes convergent ? Si oui, on dit que le jeu a une valeur limite. Est-ce que les valeurs des jeux escomptés convergent lorsque les joueurs deviennent plus patients ? Est-ce que les limites sont égales entre elles ? Sont-elles égales à la valeur *limsup* ? Lorsqu'il y a deux joueurs, se pose en plus le problème de la définition et de l'existence de ces différentes valeurs (fini, escompté, *limsup*) puis de la limite. D'autre part certaines notions comme la valeur *limsup* ne sont pas symétriques entre les joueurs, on préférera donc étudier la notion de valeur uniforme. Le jeu a une valeur uniforme si les deux joueurs peuvent garantir, indépendamment de la longueur du jeu, le même paiement.

Observer exactement les décisions prises par l'autre joueur et connaître l'état courant sont des hypothèses fortes. Il peut être intéressant de supposer que les joueurs ne sont pas parfaitement informés. À chaque étape, au lieu d'observer les actions jouées et le nouvel état résultant, les joueurs observent uniquement un signal qui dépend de l'état précédent, de l'état courant et des actions jouées.

Lorsqu'il n'y a qu'un seul joueur on parle de Processus de Décision Markoviens partiellement observables (POMDP). Une méthode classique pour étudier les POMDPs consiste à étudier un MDP auxiliaire, où l'unique joueur observe tout, et en déduire des résultats sur le problème initial. L'état de ce nouveau MDP est la croyance du joueur sur l'état du POMDP.

Lorsqu'il y a deux joueurs, on parlera de jeux répétés. Les jeux répétés avec information incomplète étudiés par Aumann et Maschler (voir référence de 1995 [AMS95]) sont un cas particulier de ce modèle.

Dans la littérature ces jeux ont été classés sous différents noms selon l'aspect privilégié :

jeux stochastiques avec observation imparfaite lorsque les joueurs ont connaissance de l'état mais reçoivent seulement des signaux sur les actions ou jeux stochastiques avec information incomplète lorsque les joueurs ont une information imparfaite sur l'état. Nous allons nous intéresser particulièrement à cette deuxième classe de jeux. Comme pour les POMDPs, nous allons voir qu'une technique de résolution est d'introduire un jeu stochastique auxiliaire où les joueurs observent tout. On trouve cette idée dans les travaux d'Aumann et Maschler, mais exprimée différemment : par une formule de récurrence utilisant comme variables les croyances des joueurs sur l'état.

Le problème général est beaucoup plus compliqué que pour les POMDPs. Afin de bien jouer les joueurs doivent prendre en considération non seulement leurs croyances sur l'état mais aussi leurs croyances sur les croyances de l'autre joueur, leurs croyances sur les croyances de l'autre joueur sur leurs croyances, etc ... Les travaux de Harsanyi [Har67] et Mertens et Zamir [MZ85] ont montré qu'il existe un espace, appelé espace universel des types, où on peut résumer cette hiérarchie infinie de croyances. Néanmoins cette approche présente deux difficultés. D'une part dans les jeux stochastiques définis sur cet espace, on ne sait pas s'il existe une valeur uniforme ou une valeur limite. D'autre part il n'est pas évident que les propriétés de ce jeu stochastique puissent être remontées au niveau du jeu répété original. Étant donné un jeu répété avec un espace d'état fini, on trouve dans la littérature des études de cas particuliers où sont faites deux hypothèses : une hypothèse pour pouvoir exprimer une formule de récurrence et/ou un jeu stochastique sur un espace plus petit que celui de Mertens et Zamir et une hypothèse pour pouvoir étudier cette formule de récurrence et/ou ce jeu stochastique auxiliaire. Selon les cas, les auteurs montrent différents résultats sur le jeu auxiliaire puis sur le jeu original.

Cette thèse est découpée en 5 chapitres. Le premier chapitre présente la théorie des jeux stochastiques et des jeux répétés ainsi que les résultats de la thèse.

Le second chapitre est extrait d'un article écrit en collaboration avec Jérôme Renault. Il commence par l'étude d'une nouvelle distance sur le simplexe. Cette distance a été introduite afin de mieux comprendre le MDP associé à un POMDP ou à un jeu répété avec un contrôleur informé. Elle rend 1-Lipschitz les transitions du MDP auxiliaire. On présente ensuite deux résultats qui montrent l'existence de stratégies garantissant la valeur limite sur n'importe quel intervalle de temps suffisamment long et pas seulement sur les intervalles de temps commençant à l'étape 1. Le premier résultat est formulé dans le cadre des maisons de jeux, à la Dubins et Savage [DS65], et le second résultat est formulé dans le cadre des MDPs. Dans chacun des cas, on donne une caractérisation de la limite. Enfin, on montre que les jeux stochastiques auxiliaires associés à un POMDP ou à un jeu répété avec un contrôleur informé satisfont les conditions du second théorème, et que l'on peut en déduire des résultats sur les POMDPs ou les jeux répétés avec un contrôleur informé.

On s'intéresse dans le troisième et le quatrième chapitres à des classes particulières de jeux répétés. Dans le troisième chapitre, on étudie les jeux commutatifs où les joueurs n'observent pas l'état. D'une part, les transitions sont dites commutatives car pour une suite de couples

d'actions donnée, la distribution de l'état après n étapes ne dépend pas de l'ordre dans lequel les n premiers couples d'actions ont été joués. D'autre part, à chaque étape, les joueurs observent les actions jouées mais pas l'état, ils ont donc une croyance commune sur laquelle on peut définir un jeu stochastique auxiliaire. On montre l'existence d'une valeur uniforme et lorsqu'il n'y a qu'un seul joueur l'existence de stratégies qui ne font pas d'erreurs.

Le quatrième chapitre est consacré à l'existence de la valeur uniforme pour les jeux répétés avec un contrôleur plus informé. Il est extrait d'un article écrit en collaboration avec Fabien Gensbittel et Miquel Oliu-Barton. Lorsqu'un joueur contrôle la transition et est informé à chaque étape de l'état, alors que l'autre joueur reçoit une information partielle, Renault [Ren12b] a montré l'existence d'une valeur uniforme en introduisant un jeu stochastique auxiliaire sur l'espace des croyances du joueur non informé. On montre que si le contrôleur est seulement plus informé,on peut définir un jeu stochastique auxiliaire avec pour espace d'états les croyances du joueur non informé sur les croyances du joueur informé puis en déduire l'existence de la valeur uniforme.

Dans le dernier chapitre constitué d'un article écrit avec Sylvain Sorin et Guillaume Vigeral, on examine d'un point de vue plus général les liens entre la convergence uniforme et le comportement asymptotique des suites (σ_n, τ_n) de stratégie optimales dans le jeu de longueur n. On vérifie que lorsqu'il n'y a qu'un seul joueur, la convergence uniforme implique l'existence d'une suite de stratégies qui garantissent la limite sur n'importe quel intervalle de temps. On montre sur un exemple les difficultés pour obtenir un résultat similaire lorsqu'il y a deux joueurs.

1.2 Jeux stochastiques

1.2.1 Modèle

Dans la suite, étant donné un ensemble non vide K, on note $\Delta_f(K)$ l'ensemble des probabilités à support fini sur K. Excepté dans ce paragraphe, les notations K, I et J seront réservées à des ensembles finis et on écrira alors par exemple $\Delta(K)$ l'ensemble des probabilités sur K. Pour tout $k \in K$, on note δ_k la masse de Dirac en k.

Un jeu stochastique $\Gamma = (K, I, J, q, g)$ à somme nulle est défini par un espace d'états K, deux ensembles d'actions I et J, respectivement pour le joueur 1 et le joueur 2, une transition q de $K \times I \times J$ dans $\Delta_f(K)$, et une fonction de paiement g de $K \times I \times J$ dans $[0, 1]$.

Soit $p \in \Delta_f(K)$ une probabilité initiale, le jeu se déroule comme suit : à l'étape 1, un état k_1 est tiré aléatoirement selon p et les joueurs observent k_1. Ensuite ils choisissent respectivement une action i_1 dans I pour le joueur 1 et j_1 dans J pour le joueur 2. Le joueur 1 gagne alors le paiement $g(k_1, i_1, j_1)$ et le joueur 2 gagne l'opposé, soit $-g(k_1, i_1, j_1)$. Un nouvel état est tiré selon la probabilité $q(k_1, i_1, j_1)$, le triplet (i_1, j_1, k_2) est annoncé aux deux joueurs et le jeu passe à l'étape suivante.

Pour tout $t \geq 1$, on définit $H_t = (K \times I \times J)^{t-1} \times K$ l'ensemble des histoires de longueur t et H_∞ l'ensemble des histoires de longueur infinie. Une stratégie comportementale du joueur

1 est une suite $\sigma = (\sigma_t)_{t\geq 1}$ telle que pour tout $t \geq 1$, σ_t est une application de H_t dans $\Delta_f(I)$. Symétriquement, une stratégie du joueur 2 est une suite $\tau = (\tau_t)_{t\geq 1}$, où pour tout $t \geq 1$, τ_t est une application de H_t dans $\Delta_f(J)$. On note Σ l'ensemble des stratégies du joueur 1 et $\mathcal{T}$ l'ensemble des stratégies du joueur 2. On dit que la stratégie σ est pure, si pour tout $t \geq 1$ et pour toutes les histoires $h_t \in H_t$, le support de σ_t est un singleton. On munit K, I et J de la topologie discrète et de la tribu engendrée par les boréliens. Pour tout $n \in \mathbb{N}$, on munit H_n de la topologie produit et de la tribu borélienne associée $\mathcal{H}_n$. Une probabilité initiale $p \in \Delta_f(K)$ associée à un couple de stratégies (σ, τ) définissent pour tout $t \geq 1$ une probabilité $I\!P^t_{p,\sigma,\tau}$ sur $(H_n, \mathcal{H}_n)$. On munit H_∞ de la tribu engendrée par les cylindres finis et par le théorème d'extension de Kolmogorov, il existe une unique probabilité $I\!P_{p,\sigma,\tau}$ sur l'ensemble des histoires infinies H_∞ telle que, pour tout $t \geq 1$, la restriction de $I\!P_{p,\sigma,\tau}$ aux histoires de longueur t est $I\!P^t_{p,\sigma,\tau}$. On note $I\!E_{p,\sigma,\tau}$ l'espérance sous cette probabilité.

La fonction de paiement a été définie étape par étape mais pas globalement. Il y a différentes façons d'évaluer le paiement donc différentes notions d'optimalité et de valeur. Par exemple, soit $\lambda \in (0, 1]$ un taux d'escompte, dans le jeu escompté $\Gamma_\lambda(p)$, le paiement est donné par

$$\gamma_\lambda(p, \sigma, \tau) = I\!E_{p,\sigma,\tau}\left(\lambda \sum_{t=1}^{+\infty} (1-\lambda)^{t-1} g(k_t, i_t, j_t)\right).$$

On peut aussi considérer simplement un nombre fini d'étapes et calculer les moyennes de Cesàro ; soit $n \geq 1$, dans le jeu $\Gamma_n(p)$, le paiement est donné par

$$\gamma_n(p, \sigma, \tau) = I\!E_{p,\sigma,\tau}\left(\frac{1}{n} \sum_{t=1}^{n} g(k_t, i_t, j_t)\right).$$

De manière plus générale, une probabilité sur les entiers positifs, $\theta \in \Delta(\mathbb{N}^*)$, est appelée une évaluation et on définit le paiement sous θ par

$$\gamma_\theta(p, \sigma, \tau) = I\!E_{p,\sigma,\tau}\left(\sum_{t\geq 1} \theta_t g(k_t, i_t, j_t)\right).$$

En fait, n'importe quelle application de $\Delta_f(K) \times \Sigma \times \mathcal{T}$ dans $[0, 1]$ définit une fonction de paiement globale et un jeu en un coup. Néanmoins sans hypothèse plus précise sur l'application on perd la structure particulière de jeu répété.

Définition 1.2.1 *Soit $\Gamma(p)$ un jeu stochastique et γ une fonction de paiement globale, c'est à dire une application de $\Delta_f(K) \times \Sigma \times \mathcal{T}$ dans $[0, 1]$. Le jeu a une valeur si*

$$\sup_{\sigma\in\Sigma} \inf_{\tau\in\mathcal{T}} \gamma(p, \sigma, \tau) = \inf_{\tau\in\mathcal{T}} \sup_{\sigma\in\Sigma} \gamma(p, \sigma, \tau)$$

Définition 1.2.2 *Une stratégie σ^* du joueur 1 garantit v dans $\Gamma(p)$ muni de la fonction de*

paiement γ si

$$\inf_{\tau\in\mathcal{T}} \gamma(p,\sigma^*,\tau) \geq v,$$

et le joueur 1 garantit v dans $\Gamma(p)$ muni de la fonction de paiement γ si

$$\sup_{\sigma\in\Sigma}\inf_{\tau\in\mathcal{T}} \gamma(p,\sigma,\tau) \geq v.$$

Symétriquement une stratégie τ^* du joueur 2 garantit v dans $\Gamma(p)$ muni de la fonction de paiement γ si

$$\sup_{\sigma\in\Sigma} \gamma(p,\sigma,\tau^*) \leq v,$$

et le joueur 2 garantit v dans Γ avec la fonction de paiement γ si

$$\inf_{\tau\in\mathcal{T}}\sup_{\sigma\in\Sigma} \gamma(p,\sigma,\tau) \leq v$$

Lorsqu'elles existent, on note respectivement $v_\lambda(p)$, $v_n(p)$ et $v_\theta(p)$ les valeurs associées respectivement aux jeux avec un nombre fini d'étapes, aux jeux escomptés et aux jeux d'évaluation θ. Si la probabilité initiale est une masse de Dirac en un état $k \in K$, on note les valeurs $v_\lambda(k)$, $v_n(k)$ et $v_\theta(k)$ et on vérifie immédiatement que ces fonctions valeurs sont affines sur $\Delta_f(K)$. Lorsque v_n existe pour tout $n \in \mathbb{N}^*$, on s'intéressera à la convergence de la suite v_n lorsque n tend vers $+\infty$ et le lien avec la convergence de v_λ lorsque λ tend vers 0.

Définition 1.2.3 *Le jeu stochastique $\Gamma(p)$ a une valeur limite $v^*(p)$ si la suite $(v_n(p))_{n\in\mathbb{N}}$ converge vers $v^*(p)$.*

Une fois l'évaluation fixée(par exemple n ou λ), le paiement dépend essentiellement d'un nombre fini d'étapes. Ainsi lorsqu'on considère la suite $(v_n(p))_{n\in\mathbb{N}}$, on s'intéresse à un comportement dans les jeux longs mais les joueurs peuvent jouer différemment dans chaque jeu. Une autre approche consiste à définir une fonction de paiement qui dépend de la suite infinie. Par exemple, les deux fonctions de paiement suivantes considèrent la moyenne la plus petite obtenue par le joueur 1 soit trajectoire par trajectoire

$$\gamma_{E(\inf)}(p,\sigma,\tau) = I\!E_{p,\sigma,\tau}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{t=1}^{n} g(k_t,i_t,j_t)\right),$$

soit en espérance

$$\gamma_{\inf E}(p_1,\sigma,\tau) = \liminf_{n\to+\infty} I\!E_{p,\sigma,\tau}\left(\frac{1}{n}\sum_{t=1}^{n} g(k_t,i_t,j_t)\right).$$

Contrairement aux notions précédentes, la symétrie du problème n'est pas conservée : si on considère la limite inférieure, le joueur 1 est pénalisé par comparaison au joueur 2. Ces définitions sont adaptées à des problèmes de contrôle avec un seul joueur où on imagine le pire cas ou à des situations où les rôles des joueurs sont dissymétriques. Pour une étude approfondie

de ce type de fonctions de paiement définies sur les histoires infinies, on pourra se référer au livre de Maitra et Sudderth [MS96]. Dans les jeux stochastiques, nous allons nous intéresser aux notions de *maxmin*, de *minmax* et de valeur uniforme, cette dernière étant symétrique.

Définition 1.2.4 *Le joueur* 1 *garantit* $v \in \mathbb{R}$ *dans le jeu* $\Gamma(p)$ *si pour tout* $\varepsilon > 0$ *il existe une stratégie* $\sigma^* \in \Sigma$ *du joueur* 1 *et un entier* $N \in \mathbb{N}^*$ *tels que*

$$\forall n \geq N,\ \forall \tau \in \mathcal{T},\ \gamma_n(p, \sigma^*, \tau) \geq v - \varepsilon.$$

On appelle max min *le maximum des paiements que le joueur* 1 *peut garantir et on le note* $\underline{v}(p)$

On peut aussi écrire la définition équivalente suivante. Celle-ci introduit la notion de paiement garanti par une stratégie.

Définition 1.2.5 *Le joueur* 1 *garantit* $v \in \mathbb{R}$ *dans le jeu* $\Gamma(p)$ *si pour tout* $\varepsilon > 0$ *il existe une stratégie* $\sigma^* \in \Sigma$ *du joueur* 1 *telle que*

$$\liminf_n \inf_{\tau \in \mathcal{T}} \gamma_n(p, \sigma^*, \tau) \geq v - \varepsilon.$$

On dit alors que la stratégie σ^* *garantit la valeur* $v - \varepsilon$.

Ainsi un joueur peut garantir la valeur v si pour tout $\varepsilon > 0$, il possède une stratégie qui garantit $v - \varepsilon$. De la même manière qu'un supremum peut ne pas être atteint, la valeur peut ne pas être garantie exactement : les joueurs peuvent être obligés de faire de petites erreurs irréversibles afin d'obtenir un bon paiement moyen.

Dans ces deux définitions on considère le plus mauvais paiement du joueur 1 tant sur la longueur que concernant le joueur 2 qui choisit sa stratégie connaissant celle du joueur 1. On remarquera que le joueur 2 peut choisir sa stratégie en fonction de la longueur du jeu et ainsi le *maxmin* existe toujours. Dans la littérature, il existe d'autres définitions, plus fortes, de *maxmin* mais qui n'existent pas toujours.

Ainsi dans Mertens et Neyman [MN81], le joueur 1 doit en plus garantir la valeur du jeu avec comme paiement l'évaluation $liminf$, $v_{E(\inf)}$. Cette propriété est liée à la convergence presque sure des paiements espérés.

Une autre définition classique impose au joueur 2 d'avoir une meilleure réponse indépendante de la longueur du jeu (voir par exemple Mertens, Sorin et Zamir [MSZ94], Sorin [Sor02] ou Rosenberg, Solan et Vieille [RSV03]). Nous allons nous concentrer sur des exemples où il existe une valeur uniforme et les deux joueurs peuvent garantir le même paiement. Dans ce cas les deux définitions sont équivalentes.

Symétriquement, on définit le paiement garanti par le joueur 2 et le *minmax*.

Définition 1.2.6 *Le joueur* 2 *garantit* $v \in \mathbb{R}$ *dans* $\Gamma(p)$ *si pour tout* $\varepsilon > 0$ *il existe une stratégie* $\tau^* \in \mathcal{T}$ *du joueur* 2 *telle que*

$$\limsup_n \sup_{\sigma \in \Sigma} \gamma_n(p, \sigma, \tau^*) \leq v + \varepsilon.$$

On appelle min max *le minimum des paiements que le joueur* 2 *peut garantir et on le note* $\overline{v}(p)$.

Par construction le *maxmin* est toujours plus petit que le *minmax*. S'il y a égalité, les deux joueurs peuvent garantir le même paiement dans tous les jeux suffisamment longs. On obtient ainsi une nouvelle notion de valeur mais, contrairement aux valeurs des jeux escomptés ou des jeux avec n-étapes, cette valeur ne s'exprime pas comme la valeur d'un jeu en un coup où les joueurs choisissent respectivement des actions dans Σ et dans $\mathcal{T}$.

Définition 1.2.7 *Soit* $p \in \Delta_f(K)$, *si* $\underline{v}(p) = \overline{v}(p)$ *on dit que le jeu a une valeur uniforme, notée* $v^*(p)$. *De plus la suite* $v_n(p)$ *converge vers* $v^*(p)$.

La convergence de la suite $v_n(p)$ est évidente à partir des définitions car $\limsup v_n(p)$ est plus petite que $\overline{v}(p)$ et $\liminf v_n(p)$ est plus grande que $\underline{v}(p)$. Pour tout $\varepsilon \geq 0$, une stratégie du joueur 1 qui garantit $v^*(p) - \varepsilon$ est dite ε-optimale.

De nombreuses classes de jeux stochastiques ont été étudiées depuis la définition du modèle. Lorsque les deux ensembles d'actions sont des singletons, on retrouve une chaine de Markov. Lorsque seulement un ensemble d'action est un singleton, on obtient un Processus de Décision Markovien (MDP). Dans le cas des MDPs, on simplifie la notation du jeu en oubliant complètement le joueur 2 et un MDP est noté $\Gamma = (K, I, q, g)$. Ces modèles ont été étudiés dans des cadres plus généraux que ceux présentés ici où nous nous sommes restreints à des probabilités à support fini. Nous nous concentrerons sur les travaux liés aux problèmes posés dans les jeux stochastiques avec deux joueurs.

Dans les jeux absorbants introduits par Kohlberg [Koh74], il n'y a qu'un seul état où les joueurs ont de l'influence mais les paiements dans cet état peuvent être quelconques.

Définition 1.2.8 *Soit* $\Gamma = (K, I, J, q, g)$ *un jeu stochastique, un état* $k \in K$ *est dit absorbant si le paiement en* k *est constant et si l'état* k *ne peut pas être quitté :*

$$\forall (i,j), (i',j') \in I \times J,\ g(k,i,j) = g(k,i',j') \text{ et } q(k,i,j) = \delta_k.$$

Si tous les états du jeu Γ *sauf un seul sont absorbants alors le jeu est dit absorbant.*

A l'inverse les jeux récursifs, introduits par Everett [Eve57] ont une structure de transition potentiellement aussi compliquée que l'on veut mais seule l'influence sur l'état est importante. Pour un état donné, toutes les actions donnent le même paiement, et ce paiement est égal à 0 tant que l'état n'est pas absorbant.

Définition 1.2.9 *Un jeu stochastique est récursif si le paiement est nul en dehors des états absorbants :*

$$\forall k \in K \text{ soit } k \text{ est absorbant, soit } \forall i,j \in I \times J,\ g(k,i,j) = 0.$$

A cause de cette convention, les paiements sont souvent choisis pour les jeux récursifs dans $[-1,1]$ *plutôt que dans* $[0,1]$ *afin d'obtenir un modèle symétrique.*

1.2.2 Cas général

Sans aucune hypothèse sur les ensembles d'états, de régularité sur le paiement et de régularité sur la transition, la valeur sous une évaluation θ n'existe par forcément. De même la valeur uniforme peut ne pas exister. Néanmoins, si on considère le cas d'un joueur, voire le cas des suites réelles, pour chaque évaluation θ, la valeur sous θ est bien définie. Une suite est formellement un jeu stochastique avec un nombre dénombrable d'états, une transition déterministe et des ensembles d'actions qui sont des singletons. Or l'étude des suites montre qu'il y a des liens possibles entre les valeurs des jeux escomptés et les valeurs des jeux stochastiques avec un nombre fini d'étapes.

Étant donnée une suite réelle $r = (r_t)_{t\geq 1}$, si on note

$$\gamma_\lambda(r) = \lambda \sum_{t=1}^{+\infty} (1-\lambda)^{t-1} r_t,$$

et

$$\gamma_n(r) = \frac{1}{n} \sum_{t=1}^{n} r_t,$$

alors on a

$$\liminf_n \gamma_n(r) \leq \liminf_{\lambda\to 0} \gamma_\lambda(r) \leq \limsup_{\lambda\to 0} \gamma_\lambda(r) \leq \limsup_n \gamma_n(r).$$

Ainsi, la convergence de $\gamma_n(r)$ implique la convergence de $\gamma_\lambda(r)$ vers la même limite.

En fait, par un théorème tauberien d'Hardy et Littlewood, on a l'implication inverse lorsque les paiements sont bornés (voir Filar et Sznajder [SF92] pour une preuve dans ce cadre). Soit $r = (r_t)_{t\geq 1}$ une suite dans $[0,1]$, alors $\gamma_n(r)$ converge lorsque n tend vers $+\infty$, si et seulement si, $\gamma_\lambda(r)$ converge lorsque λ tend vers 0. De plus, les deux suites convergent vers la même limite.

Avec un joueur la situation est différente. Pour différentes valeurs des paramètres n ou λ, le joueur n'utilise pas les mêmes stratégies donc les suites de paiement considérées ne sont pas les mêmes. Lehrer et Sorin [LS92] ont prouvé qu'il n'y a pas d'équivalence entre la convergence de v_λ et de v_n mais il y équivalence si la convergence est uniforme par rapport à l'état.

Exemple 1.2.10 (Lehrer-Sorin) *On considère $\Gamma = (K, I, q, g)$, un MDP, tel que $K = \mathbb{N}\times\mathbb{N}$, $I = \{R, T\}$, la fonction de paiement est donnée par*

$$\begin{aligned} \forall (m,l) \in \mathbb{N}\times\mathbb{N}^*, \qquad g((m,0),R) &= 0 \\ g((m,0),T) &= 1 \\ g((m,l),R) = g((m,l),T) &= 1 \text{ si } l \leq m \\ g((m,l),R) = g((m,l),T) &= 0 \text{ si } l > m \end{aligned}$$

et la transition, déterministe, est donnée par

$$\begin{aligned} \forall (m,l) \in \mathbb{N} \times \mathbb{N}^*, \qquad q((m,0),R) &= (m+1,0) \\ q((m,1),L) &= (m,1) \\ q((m,l),R) = q((m,l),T) &= (m,l+1) \end{aligned}$$

Dès que le joueur 1 *joue* T*, la première coordonnée est fixée pour le reste du processus et la seconde coordonnée augmente de* 1 *à chaque étape. Pour tout entier pair* $n \geq 1$*, la valeur du jeu fini* $v_n(0,0)$ *est* $\frac{1}{2}$*. Comme les paiements sont bornés, il existe donc une valeur limite égale à* $\frac{1}{2}$*, alors que les valeurs escomptées* $v_\lambda(0,0)$ *convergent vers* $\frac{1}{4}$*.*

La convergence uniforme ne suffit pas pour impliquer l'existence d'une valeur uniforme (Monderer et Sorin [MS93]). Dans le cas de deux joueurs, la question de l'équivalence entre les convergences uniformes des valeurs de Cesàro et des valeurs escomptées est une question ouverte.

En conclusion, citons deux hypothèses sur les familles de fonctions qui impliquent la convergence de v_n et de v_λ.

Mertens et Neymann [MN81] ont prouvé que si la famille $(v_\lambda)_\lambda$ est à variation bornée :

$$\forall (\lambda_i)_{i\in\mathbb{N}} \text{ décroissante}, \ \sum_{i\in\mathbb{N}} \|v_{\lambda_{i+1}} - v_{\lambda_i}\|_\infty < +\infty,$$

propriété plus forte que la convergence uniforme de v_λ, alors v_λ et v_n convergent toutes les deux vers la même limite. En fait, il existe même une valeur uniforme. On trouve une preuve simple de la convergence de v_n dans Neyman [Ney03].

D'autre part, Renault [Ren11] montre que lorsqu'il n'y a qu'un seul joueur, la suite v_n converge uniformément, si et seulement si, la famille $\{v_n, n \in \mathbb{N}^*\}$ est précompacte pour la norme uniforme. Dans le même article il montre que la précompacité d'une famille de fonctions auxiliaires implique l'existence de la valeur uniforme.

1.2.3 Cas fini

Dans le cas d'espaces finis (états et actions), les jeux avec n étapes et les jeux escomptés ont une valeur. On définit l'ensemble des stratégies mixtes comme l'ensemble de probabilité sur l'ensemble des stratégies pures munis de la topologie produit et de la tribu borélienne. Par le théorème de Kuhn [Kuh53], l'ensemble des probabilités générées par des stratégies comportementales est égal à l'ensemble des probabilités générées par des stratégies mixtes. Le jeu $\Gamma_n(p)$, où seules les n premières étapes comptent, a un nombre fini de stratégies pures différentes. Il a une valeur par le théorème de Von Neumann [VN28] sur les jeux matriciels finis. La fonction de paiement γ_λ est, quant à elle, linéaire par rapport aux stratégies des joueurs et continue si les espaces de stratégies sont munis de la topologie produit. Ainsi v_λ existe, par exemple par le théorème de Sion [Sio58]. Ce résultat a été prouvé par Shapley [Sha53] en approximant le jeu escompté par une succession de jeux finis.

On définit l'extension multilinéaire de g et q à $K \times \Delta(I) \times \Delta(J)$. Soit $k \in K$, $a \in \Delta(I)$ et $b \in \Delta(J)$, on pose

$$g(k,a,b) = \sum_{i \in I, j \in J} a(i)b(j)g(k,i,j),$$

et

$$q(k,a,b) = \sum_{i \in I, j \in J} a(i)b(j)q(k,i,j).$$

La suite v_n satisfait alors l'équation de récurrence suivante

$$\begin{aligned} v_{n+1}(k) &= \sup_{a \in \Delta(I)} \inf_{b \in \Delta(J)} \frac{1}{n+1} g(k,a,b) + \frac{n}{n+1} I\!E_{q(k,a,b)}[v_n], \\ &= \inf_{b \in \Delta(J)} \sup_{a \in \Delta(I)} \frac{1}{n+1} g(k,a,b) + \frac{n}{n+1} I\!E_{q(k,a,b)}[v_n], \end{aligned}$$

et v_λ l'équation de point fixe

$$v_\lambda(k) = \sup_{a \in \Delta(I)} \inf_{b \in \Delta(J)} \lambda g(k,a,b) + (1-\lambda) I\!E_{q(k,a,b)}[v_\lambda], \tag{1.1}$$

$$= \inf_{b \in \Delta(J)} \sup_{a \in \Delta(I)} \lambda g(k,a,b) + (1-\lambda) I\!E_{q(k,a,b)}[v_\lambda]. \tag{1.2}$$

Notons qu'à chaque fois le second opérateur peut être remplacer par un infimum (resp. maximum) sur les actions pures car la fonction de paiement optimisée est linéaire. En particulier, lorsqu'il n'y a qu'un seul joueur, il a une stratégie optimale pure.

Le jeu $\Gamma = (K, I, J, q, g)$ admet une valeur limite et même une valeur uniforme quelque soit la probabilité initiale. L'existence de la valeur uniforme a été prouvée en plusieurs étapes. Blackwell [Bla62] a mis en évidence l'existence de la valeur uniforme dans le cadre des MDPs en montrant que l'unique joueur appelé le décideur, peut se restreindre a des stratégies pures et stationnaires (qui ne dépendent que de l'état courant). Étant donnée une stratégie stationnaire σ, le paiement des jeux escomptés $\gamma_\lambda(\sigma)$ est une fonction bornée rationnelle en λ. En particulier il existe une fonction plus grande que les autres au voisinage de 0 : il existe λ^* et σ^* pure et stationnaire, tels que pour tout $\lambda < \lambda^*$:

$$\forall \sigma \in \Sigma,\ \gamma_\lambda(\sigma^*) \geq \gamma_\lambda(\sigma).$$

Lorsque l'on considère la limite de v_λ quand λ tend vers 0, la suite de paiements est fixe au voisinage de 0 et est générée par une chaîne de Markov avec un nombre d'états fini donc v_λ converge et v_n converge vers la même limite.

Dans le cas général avec deux joueurs, la valeur n'est plus une fonction rationnelle en λ mais Bewley et Kohlberg [BK76b], [BK76a] ont montré que v_λ est une fonction semi-algébrique bornée au voisinage de 0 et est donc convergente. En effet par compacité des ensembles d'actions et continuité de la fonction de paiement, l'équation de point fixe implique que le joueur 1 a une

stratégie optimale stationnaire dans Γ_λ, notée a_λ, et que le joueur 2 a une stratégie optimale stationnaire, notée b_λ. Ainsi la formule de point fixe 1.1 peut être réécrite comme un nombre fini d'inégalités polynomiales, dont $(\lambda, v_\lambda, a_\lambda, b_\lambda)$ pour $\lambda \in (0,1]$ est solution. Cet ensemble est donc semi-algébrique et par projection, pour tout $k \in K$, la fonction qui associe $v_\lambda(k)$ à λ est aussi semi-algébrique. Cette fonction est bornée au voisinage de 0, elle converge donc vers une limite $v(k)$. Comme v_λ est à variation bornée, v_n converge vers la même limite.

Mertens et Neyman [MN81] ont ensuite utilisé la semi-algébricité de la fonction v_λ pour prouver l'existence de la valeur uniforme. Chaque joueur peut garantir la limite en jouant à chaque étape une stratégie optimale dans un jeu escompté de paramètre λ_n, où λ_n est calculé en fonction des paiements observés. A partir de la variation bornée de la famille de fonctions v_λ, ils construisent de manière subtile la famille λ_n telle que v_{λ_n} soit presque une surmartingale reliée aux paiements. Ils déduisent que cette stratégie est effectivement ε-optimale. Les auteurs montrent aussi que ces stratégies garantissent le paiement v dans le jeu avec pour fonction de paiement globale l'espérance de la *liminf* pour le joueur 1 et l'espérance de la *limsup* pour le joueur 2.

Entre l'introduction du modèle des jeux stochastiques et la résolution par Mertens et Neyman, il a été publié de nombreux articles sur des classes particulières de jeux stochastiques. Aujourd'hui, ces classes servent encore de modèles simples pour comprendre les jeux répétés. Dans certains cas, les résultats sont plus forts notamment vis à vis des stratégies optimales : existence de stratégies 0-optimales, pures et/ou stationnaires. On trouvera des preuves spécifiques de l'existence de la valeur limite pour les jeux récursifs dans Everett [Eve57] et de l'existence de la valeur uniforme dans Thuisjman et Vrieze [TV92]. Liggett et Lippman [LL69] ont étudié les jeux où les joueurs jouent tour à tour et Kohlberg [Koh74] a montré l'existence de la valeur uniforme pour les jeux absorbants. Enfin, citons la résolution par Blackwell et Ferguson [BF68], du cas particulier du Big Match introduit par Gilette [Gil57] qui a a servi d'exemple pour le cas général.

Exemple 1.2.11 ([Gil57]) *Le Big Match est un jeu stochastique tel que l'espace d'états K est constitué de 3 éléments : un état non absorbant α, un état k_0 absorbant de paiement 0 et un état k_1 absorbant de paiement 1. Le joueur 1 a deux actions $\{T, B\}$ et le joueur 2 a deux actions $\{L, R\}$. La matrice de paiement/transition en α est donnée par*

$$\begin{array}{c} \\ T \\ B \end{array} \begin{array}{c} \begin{array}{cc} L & R \end{array} \\ \begin{pmatrix} 1_{\to k_1} & 0_{\to k_0} \\ 0_{\circlearrowleft} & 1_{\circlearrowleft} \end{pmatrix} \end{array}.$$

On résume ces informations en indiquant d'une étoile les paiements d'absorption

$$\begin{pmatrix} 1^* & 0^* \\ 0 & 1 \end{pmatrix}.$$

Ce jeu a une valeur uniforme égale à $\frac{1}{2}$. Le joueur 2 peut garantir $\frac{1}{2}$ en jouant i.i.d. $(1/2, 1/2)$. Le joueur 1 n'a pas de stratégie simple qui lui garantisse la valeur uniforme. Il n'a pas de

stratégies à mémoire bornée et doit utiliser une stratégie qui dépend d'une statistique des coups passés.

1.2.4 Extensions à des espaces d'actions compacts

La preuve de l'existence de la valeur uniforme dans le paragraphe précédent repose essentiellement sur les hypothèses que l'espace d'états et les espaces d'actions sont finis. Si ce n'est plus le cas la fonction valeur n'est plus, ni rationnelle pour le cas d'un joueur, ni semi-algébrique lorsqu'il y a 2 joueurs.

Néanmoins, ces résultats ont été généralisés partiellement en relâchant l'hypothèse sur les espaces d'actions : plus précisément lorsque l'espace d'états est fini, les espaces d'actions sont compacts, et les fonctions de paiement et de transition sont continues par rapport aux actions. Rosenberg et Sorin [RS01] ont montré l'existence de la valeur limite pour les jeux absorbants et Sorin [Sor03] dans les jeux récursifs. Mertens, Neyman et Rosenberg [MNR09] ont ensuite complété le résultat sur les jeux absorbants en montrant l'existence de la valeur uniforme. Concernant les MDPs, l'existence de la valeur uniforme a été montrée par Dynkin et Yushkevitch [DJ79]. Renault [Ren11] a ensuite prouvé l'existence de la valeur uniforme dans les MDPs sans hypothèses ni sur l'ensemble d'actions ni sur les fonctions de paiement et de transition. Par contre la stratégie décrite ne garantit pas à priori la valeur limite dans le jeu où l'évaluation globale est donnée par l'espérance de la lim inf des moyennes de Cesàro des paiements sur chaque trajectoire. L'existence d'une stratégie optimale qui vérifie cette propriété est une question ouverte.

La généralisation de ces résultats à un espace d'état compact reste un problème. Il représente une des clefs pour l'étude des jeux répétés, où les joueurs ne sont plus parfaitement informés des états passés et des actions jouées par l'autre joueur.

1.3 Un modèle général de jeu répété

1.3.1 Modèle

Un jeu répété à somme nulle $\Gamma = (K, I, J, C, D, q, g)$, où I, J, C et D sont non vides, est défini par un espace d'états K, par des espaces d'actions I et J respectivement pour le joueur 1 et le joueur 2, par des espaces de signaux C et D respectivement pour le joueur 1 et le joueur 2, par une fonction de transition q de $K \times I \times J$ dans $\Delta_f(K \times C \times D)$ et par une fonction de paiement g de $K \times I \times J$ dans $[0, 1]$. L'espace d'états sera toujours supposé fini sauf mention contraire.

A l'étape initiale, on va supposer que les signaux ne sont pas à valeur dans C et D mais dans les entiers naturels $\mathbb{N}$ et $\mathbb{N}$. Nous allons voir que le jeu répété à partir de l'étape t peut se formuler comme un jeu répété à l'étape 1, avec pour probabilité initiale, la probabilité sur

les histoires de longueur t. Lorsque t augmente, les histoires sont de plus en plus longues d'où la nécessité de définir les jeux répétés avec un nombre fini mais quelconque de signaux et l'utilisation des entiers naturels.

Étant donné une probabilité initiale $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$, le jeu $\Gamma(\pi)$ se déroule de la manière suivante : à l'étape 1, un triplet (k_1, c', d') est tiré aléatoirement selon π, le joueur 1 observe le signal c' et le joueur 2 observe le signal d'. Puis le joueur 1 choisit une action i_1 dans I et le joueur 2 choisit une action j_1 dans J. Le joueur 1 reçoit le paiement $g(k_1, i_1, j_1)$ sans l'observer, le joueur 2 reçoit l'opposé et un nouveau triplet (k_2, c_1, d_1) est tiré selon la loi $q(k_1, i_1, j_1)$. Le joueur 1 observe le signal c_1, le joueur 2 observe le signal d_1 et le jeu passe à l'étape suivante, etc ...

Contrairement au modèle des jeux stochastiques, les joueurs n'observent ni le nouvel état, ni l'action jouée par l'autre joueur mais seulement un signal qui dépend de l'état courant et des actions jouées. Dans cette formulation on ne distingue pas ces deux types d'informations qui sont toutes les deux exprimées par l'unique signal. Ce modèle contient donc des modèles dit "à information incomplète" (, où les joueurs observent parfaitement les actions mais ont une information partielle sur l'état), des modèles dit "à observation imparfaite" (, où les joueurs connaissent l'état mais ont une information partielle sur les actions), et des modèles, où les deux aspects sont présents. Les exemples de jeux stochastiques que nous allons étudier correspondent à l'aspect "information incomplète".

Pour tout $t \geq 1$, on définit $H_t^1 = \mathbb{N} \times (I \times C)^{t-1}$ l'ensemble des histoires du joueur 1 à l'étape t et $H_t^2 = \mathbb{N} \times (J \times D)^{t-1}$ celles du joueur 2. L'ensemble des histoires complètes $K \times \mathbb{N} \times \mathbb{N} \times (I \times J \times K \times C \times D)^{t-1}$ sera noté H_t et l'ensemble des histoires infinies $K \times \mathbb{N} \times \mathbb{N} \times (I \times J \times K \times C \times D)^{\infty}$ sera noté H_∞. Une stratégie comportementale du joueur 1 est une suite $\sigma = (\sigma_t)_{t \geq 1}$ d'applications telle que pour tout $t \geq 1$, σ_t est une application de ses histoires de longueur t, H_t^1, dans les probabilités à support fini sur ses actions, $\Delta_f(I)$. Une stratégie comportementale du joueur 2 est une suite d'applications $\tau = (\tau_t)_{t \geq 1}$ telle que pour tout $t \geq 1$, τ_t est une application de $\mathbb{N} \times (J \times D)^{t-1}$ dans $\Delta_f(J)$. En général on notera Σ l'ensemble des stratégies du joueur 1 et $\mathcal{T}$ l'ensemble des stratégies du joueur 2.

Un triplet (π, σ, τ) définit pour tout $t \geq 1$ une probabilité $I\!P^t_{\pi,\sigma,\tau}$ sur un sous-ensemble dénombrable de H_t. Par le théorème d'extension de Kolmogorov, il existe une unique probabilité $I\!P_{\pi,\sigma,\tau}$ sur les histoires de longueur infinie telle que pour tout $t \geq 1$ la restriction de $I\!P_{\pi,\sigma,\tau}$ aux histoires de longueur t soit la probabilité $I\!P^t_{\pi,\sigma,\tau}$ et on notera $I\!E_{\pi,\sigma,\tau}$ l'espérance sous cette probabilité.

Afin d'évaluer le paiement, on dispose des mêmes critères que pour les jeux stochastiques : jeux escomptés, jeux finis, jeux avec comme paiement la lim inf de l'espérance, ... et les définitions des valeurs associées. On peut aussi définir la notion de valeur uniforme de la même manière avec ces nouveaux ensembles de stratégies.

Il existe deux questions centrales. Est ce que le jeu répété a une valeur limite ? Est ce qu'il a une valeur uniforme ? Dans tous les cas particuliers étudiés depuis les années 1965, où les espaces d'états et les espaces d'actions sont finis, soit il existe une valeur limite soit on ne sait pas. Par contre la valeur uniforme n'existe pas forcément. Mertens, Sorin et Zamir [MSZ94] ont formalisé deux conjectures. La première conjecture dit que lorsque les ensembles d'états, d'actions et de signaux, sont finis la valeur limite existe. La seconde conjecture dit que si la valeur limite existe et le joueur 1 est plus informé que le joueur 2 alors elle doit être égale au *maxmin* (ceci est nécessairement le cas lorsque la valeur uniforme existe).

Dans un premier temps on considère le cas des processus de décision Markoviens partiellement observables (POMDPs) où il n'y a qu'un joueur. Ensuite on présente l'espace universel des croyances de Mertens et Zamir [MZ85] et les limites actuelles de la théorie sans hypothèses supplémentaires sur la transition.

1.3.2 Étude du cas d'un joueur : MDPs partiellement observables

a) Réduction

Soit $\Gamma = (K, I, C, q, g)$ un POMDP avec K et I deux ensembles finis. Étant donnée une stratégie σ et une histoire observée h_t^1, le décideur peut calculer une croyance sur l'état $p_t(h_t^1) = I\!P_{\pi,\sigma}(k_t|h_t^1) \in X = \Delta(K)$. On va montrer que cette croyance joue le rôle de statisiqute suffisante et que l'on peut associer à ce POMDP, un MDP sur l'espace $\Delta_f(K)$. On présente une réduction explicite inspirée de Renault [Ren11], où l'hypothèse de transition avec support fini nous assure que l'on définit bien un MDP. Pour le cas de transitions avec des supports non finis, il faut en plus des hypothèses de mesurabilité sur la transition (Astrom, K.J. [Ast65], Sawaragi et Yoshikawa [SY70] et Rhenius [Rhe74]).

À l'étape 1, on note $\psi_{\mathbb{N}}$ l'application de $\Delta_f(K \times \mathbb{N})$ dans $\Delta_f(X)$ qui associe, à la probabilité initiale π, la loi initiale des croyances du décideur sur l'état

$$z_1 = \psi_{\mathbb{N}}(\pi) = \sum_{c' \in \mathbb{N}} \pi(c')\delta_{p_1(c')}.$$

Pour chaque évaluation $\theta \in \Delta(\mathbb{N}^*)$, la valeur $v_\theta(\pi)$ ne dépend que de la projection de π par $\psi_{\mathbb{N}}$ sur $Z = \Delta_f(X)$ car

$$v_\theta(\pi) = \sup_{\sigma \in \Sigma} \sum_{c' \in \mathbb{N}} \pi(c')\gamma_\theta(p_1(c'), \sigma(c')) = \sum_{c' \in \mathbb{N}} \pi(c') \sup_{\sigma \in \Sigma} \gamma_\theta(p_1(c'), \sigma),$$

qui ne dépend que de la désintégration de π. Inversement si $z \in \Delta_f(\Delta(K))$, il existe p_1,...,p_l tel que $z = \sum_{i=\{0,..,l\}} z(p_i)\delta_{p_i}$ et on note $\widetilde{v}_\theta(z)$ la valeur du jeu $\Gamma(\pi)$ avec $\pi(k, i) = p_i^k$ pour tout $(k, i) \in K \times \{0, ..., l\}$. Lorsque z est une masse de Dirac en p, on notera la valeur directement

$\widetilde{v}_\theta(p)$. L'équation précédente devient donc après projection

$$\widetilde{v}_\theta(z) = \sum_{p\in\Delta(K)} z(p)\widetilde{v}_\theta(p).$$

D'autre part on peut écrire un principe de programmation dynamique. Soit θ une évaluation, si on note θ^+ l'évaluation donnée pour tout $t \geq 1$ par $\theta_t^+ = \frac{\theta_t}{1-\theta_1}$ si $\theta_1 < 1$ et $\theta_t^+ = 0$ sinon, alors la valeur satisfait l'équation

$$v_\theta(\pi) = \sup_{a\in I^{\mathbb{N}}} \theta_1 \left(\sum_{k,c'\in K\times\mathbb{N}} \pi(k,c')g(k,a(c')) \right) + (1-\theta_1)v_{\theta^+}(\pi.q(\pi,a)),$$

où $\pi.q(\pi,a)$ est la loi sur les histoires de longueur 2, si la probabilité initiale est π et le décideur choisit la stratégie a. À priori, la fonction v_{θ^+} n'est pas définie sur $\pi.q(\pi,a)$ qui n'est pas dans $\Delta_f(K\times\mathbb{N})$. Néanmoins si on énumère les différents histoires observées par le décideur, on peut réécrire $\pi.q(\pi,a)$ comme une probabilité sur $\Delta_f(K\times\mathbb{N})$ et la valeur a bien un sens. Avec les notations réduites, on obtient une équation équivalente à un jeu stochastique auxiliaire. On définit $\widetilde{g}$, la fonction de $X\times I$ dans $[0,1]$, par

$$\widetilde{g}(p,i) = \sum_{k'\in K} p(k')g(k',i),$$

et $\widetilde{q}$, une application de $X\times I$ dans $\Delta_f(X)$, par

$$\widetilde{q}(p,i) = \sum_{c\in C} q(p,i)(c)\delta_{\hat{q}(p,i)(.|c)},$$

où $q(p,i)(c) = \sum_{k,k'\in K} p^{k'}q(k',i)(k,c)$, $\hat{q}(p,i)(k|c) = \frac{\sum_{k'\in K} p^{k'}q(k',i)(k,c)}{q(p,i)(c)}$ et

$$\hat{q}(p,i)(.|c) = (\hat{q}(p,i)(k|c))_{k\in K}.$$

Ainsi $\hat{q}(p,i)(k|c)$ est la probabilité que l'état soit k si le décideur a joué l'action i et observé le signal c. En utilisant les notations réduites précédentes, on obtient pour tout $p\in\Delta(K)$ (jeu où le joueur 1 ne reçoit aucune information et donc choisit une seule action) :

$$\begin{aligned}\widetilde{v}_\theta(p) &= \sup_{i\in I} \theta_1\widetilde{g}(p,i) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(p,i)).\\ &= \sup_{i\in I} \theta_1\widetilde{g}(p,i) + (1-\theta_1)I\!E_{\widetilde{q}(p,i)}\left[\widetilde{v}_{\theta^+}\right].\end{aligned}$$

Cette équation est exactement l'équation associée au MDP, $\Psi = (X, I, \widetilde{q}, \widetilde{g})$, et les fonctions

valeurs coïncident lorsque $\theta_1 = 1$ donc les valeurs sont égales pour toutes les évaluations. Ainsi $\Psi(\psi_{\mathbb{N}}(\pi))$ a une valeur limite, si et seulement si, $\Gamma(\pi)$ a aussi une valeur limite. Si σ est une stratégie dans le MDP $\Psi(\psi_{\mathbb{N}}(\pi))$, il existe alors σ' une stratégie dans le jeu original telle que pour toutes les évaluations θ, les paiements dans Ψ et dans Γ sont égaux. Ainsi si Ψ admet une valeur uniforme pour chaque probabilité initiale alors Γ a aussi une valeur uniforme pour chaque probabilité initiale. Par contre l'étude de Ψ n'indique rien pour les fonctions de paiements, comme l'espérance de la $liminf$, qui sont calculées trajectoire par trajectoire.

b) Résultats

On peut donc déduire des résultats sur les POMDPs à partir des théorèmes sur les MDPs avec espace d'états Borélien. Pour l'étude des MDPs avec un espace d'états Borélien, une propriété supplémentaire est nécessaire afin de prouver l'existence de la limite ou de la valeur uniforme comme dans Schal [Sch93] ou Borkar [Bor00] [Bor07]. En particulier, lorsque le processus est ergodique et les valeurs escomptées v_λ convergent suffisamment rapidement, on peut montrer que la valeur satisfait l'ACOE (Average Cost Optimality Equation).

Si l'on s'intéresse uniquement aux MDPs définis à partir des processus de décision Markoviens partiellement observables, on peut utiliser des preuves spécifiques liées à leur structure. Ainsi si on munit $X = \Delta(K)$ de la métrique induite par $\|.\|_1$, les fonctions valeurs pour chaque évaluation sont équicontinues.

Rosenberg, Solan et Vieille [RSV02] montrent que lorsque les ensembles d'états, d'actions et de signaux sont finis alors le POMDP a une valeur uniforme. De plus, pour tout $\varepsilon > 0$, il existe un taux d'escompte λ^* tel que pour tout $\lambda \in (0, \lambda^*]$ et $p \in \Delta(K)$, il existe une stratégie σ^* où

$$\gamma_\lambda(p, \sigma^*) \geq v_\lambda(p) - \varepsilon.$$

Contrairement à un MDP, il n'existe pas forcément de stratégie qui garantisse exactement la valeur uniforme, mais à une erreur ε fixée, le décideur peut utiliser la même stratégie pour tous les taux d'escomptes suffisamment petits. Lorsque l'ensemble des signaux est réduit à un singleton, on parle de MDP dans le noir car le décideur n'observe rien et le décideur n'a pas besoin d'utiliser de stratégies comportementales. Il peut donc garantir la limite avec une stratégie pure.

Renault [Ren11] a étendu ce résultat au cas où les ensembles d'actions A et de signaux C sont quelconques et pour tous les couples $(k, a) \in K \times A$, $q(k, a)$ est à support fini, en introduisant un problème de programmation dynamique. La preuve repose sur l'étude de la famille de fonctions auxiliaires

$$w_{m,n}(z) = \sup_{\sigma \in \Sigma} \inf_{t \in \{0,\dots,n\}} \gamma_{m,t}(z, \sigma),$$

où pour tout $m \in \mathbb{N}$ et $n \geq 1$, $\gamma_{m,n}$ est le paiement avec pour évaluation la moyenne de Cesàro

entre l'étape $m+1$ et l'étape $m+n$. Dans cette formule, on considère pour chaque stratégie du joueur 1, le paiement le plus mauvais parmi les jeux finis commençant à la date m et finissant avant n. Renault montre que cette famille est précompacte et que cela implique l'existence d'une valeur uniforme. Dans le même article, il montre que la famille $\{v_n, n \in \mathbb{N}^*\}$ est précompacte, si et seulement si, $(v_n)_{n\in\mathbb{N}^*}$ converge uniformément.

Ainsi chez Rosenberg, Solan et Vieille [RSV02] et Renault [Ren11], la valeur uniforme existe mais on ne sait pas si les stratégies décrites dans ces deux articles garantissent aussi la limite dans le jeu où le paiement est l'espérance de la lim inf. De plus les résultats précédents contrastent avec le cas des MDPs où le décideur peut obtenir le paiement maximum avec des stratégies pures. L'utilisation de stratégies comportementales est important lorsqu'il y a un adversaire, afin de lui cacher sa stratégie, mais semble superflue lorsque le décideur est tout seul. Les deux démonstrations nécessitent l'utilisation de stratégies comportementales pour jouer contre un joueur fictif : "le temps". Dans la première démonstration, étant donné une stratégie optimale à partir d'une distribution $z \in \Delta_f(X)$, l'utilisation des probabilités est nécessaire pour imiter cette stratégie à partir d'une distribution proche $z' \in \Delta_f(X)$. Dans la seconde démonstration, la nécessité de considérer des stratégies comportementales apparait sous la forme de la convexité de l'ensemble Σ qui permet d'utiliser un théorème de Sion dans la définition des fonctions $w_{m,n}$ et ainsi montrer que la famille $w_{m,n}$ est équicontinue. Comme $\Delta_f(X)$ est précompact, cela implique que c'est une famille précompacte.

1.3.3 Étude des jeux répétés

Soit $\Gamma = (K, I, J, C, D, q, g)$ un jeu répété tel que tous les ensembles (d'états, d'actions et de signaux) soient finis. Étant donné une évaluation $\theta \in \Delta_f(\mathbb{N}^*)$, on définit le paiement pour une probabilité initiale π et un couple de stratégies (σ, τ) par

$$\gamma_\theta(p, \sigma, \tau) = I\!E_{p,\sigma,\tau}\left(\sum_{t\geq 1} \theta_t g(k_t, i_t, j_t)\right).$$

Le joueur 1 et le joueur 2 peuvent alors garantir la même quantité et le jeu a une valeur notée v_θ :

$$v_\theta(\pi) = \sup_{\sigma\in\Sigma}\inf_{\tau\in\mathcal{T}} \gamma_\theta(\pi, \sigma, \tau) = \inf_{\tau\in\mathcal{T}}\sup_{\sigma\in\Sigma} \gamma_\theta(\pi, \sigma, \tau).$$

alors la valeur v_θ existe pour chaque évaluation $\theta \in \Delta(\mathbb{N}^*)$.

on dispose d'une équation de récurrence similaire à celle pour les POMDPs pour chaque probabilité initiale. Pour tout $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$, $a \in \Delta(I)^{\mathbb{N}}$ et $b \in \Delta(J)^{\mathbb{N}}$, on définit

$$g(\pi, a, b) = \sum_{c'\in\mathbb{N}, d'\in\mathbb{N}, i\in I, j\in J} a(c')(i)b'(j)\pi(k, c', d')g(k, a(c')(i), b(d')(j))$$

et $\pi.q(\pi, a, b)$, la loi sur les histoires de longueur 2 si la loi initiale est π, le joueur 1 joue a et

le joueur 2 joue b. Formellement, cette loi est définie par : pour tout $(k', c', d', i, j, k, c, d) \in H_2$,

$$(\pi.q(\pi, a, b))\,(k', c', d', i, j, k, c, d) = \pi(k', c', d')a(c')(i)b(d')(j)q(k', i, j)(k, c, d).$$

La valeur satisfait alors la formule de récurrence suivante

$$\begin{aligned} v_\theta(\pi) &= \sup_{a\in\Delta(I)^{\mathbb{N}}} \inf_{b\in\Delta(J)^{\mathbb{N}}} \theta_1 g(\pi, a, b) + (1-\theta_1) v_{\theta^+}(\pi.q(\pi, a, b)), \\ &= \inf_{b\in\Delta(J)^{\mathbb{N}}} \sup_{a\in\Delta(I)^{\mathbb{N}}} \theta_1 g(\pi, a, b) + (1-\theta_1) v_{\theta^+}(\pi.q(\pi, a, b)). \end{aligned}$$

Comme dans les cas des POMDPs, en choisissant une énumération de K, C et D, la loi $\pi.q(\pi, a, b)$ peut être interprétée comme une loi sur $K \times \mathbb{N} \times \mathbb{N}$. On agrège les signaux du joueur 1, c'est à dire son signal initial, son action i et son signal c à la fin de l'étape 1, en un seul signal dans $\mathbb{N}$ et on fait de même pour le joueur 2. Ainsi $v_{\theta^+}(\pi.q(\pi, a, b))$ est bien définie.

Mais $\Delta_f(K \times \mathbb{N} \times \mathbb{N})$ n'a pas de structure adaptée à l'étude des fonctions v_θ pour au moins deux raisons : de nombreuses distributions sont équivalentes en terme de jeu et les distances classiques ne sont pas adaptées. Par exemple la distance

$$d(\pi, \pi') = \sum_{k,c,d} |\pi(k, c, d) - \pi'(k, c, d)|,$$

rend les fonctions valeurs 1-Lipschitz mais ne rend pas l'espace compact. Il est donc naturel de chercher un autre espace où serait résumée toute l'information stratégique et où serait exprimée la formule de récurrence. Dans le cas d'un seul joueur, on a vu que l'espace $\Delta(\Delta(K))$ convient.

Lorsqu'il y a deux joueurs, il y a deux types d'informations apportées par les signaux : une information sur l'état et une information sur le signal de l'autre joueur, et donc implicitement sur l'information que le second joueur a obtenu sur l'état et sur notre propre signal. Ainsi mon signal induit une croyance sur le signal du second joueur qui, lui même, implique une croyance sur mon signal. À cause de cet aspect récursif, il est nécéssaire de considérer un nombre infini de niveaux de croyances : les croyances sur l'état, les croyances des joueurs sur les croyances de l'autre joueur sur l'état et ainsi de suite. Cette idée a été formalisée par les travaux d'Harsanyi [Har67] et de Mertens et Zamir [MZ85]. Ces derniers ont prouvé l'existence d'un espace universel des types Θ (dont on considérera deux copies Θ_1 et Θ_2) et un espace universel des croyances $\Omega = K \times \Theta_1 \times \Theta_2$.

L'espace universel des types Θ est compact, ne dépend que de l'ensemble d'états K et il existe un homéomorphisme ϕ de Θ dans $\Delta(K \times \Theta)$. Le but de cet espace est de représenter toute l'information d'un joueur : si un joueur est de type θ alors sa croyance sur l'état et sur le type de l'autre joueur 2 est $\phi(\theta)$.

Étant donnée une probabilité initiale $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$, Mertens et Zamir ont montré l'existence d'une probabilité π' sur Ω telle que pour chaque évaluation $\theta \in \Delta(\mathbb{N}^*)$, $v_\theta(\pi') = v_\theta(\pi)$. De plus sous la probabilité π', lorsque le joueur 1 reçoit le signal, ou type, θ_1 alors sa croyance sur l'état et le type du joueur 2 donnée par la probabilité conditionnelle $\pi'(.|\theta_1)$ et

$\phi(\theta_1)$ donné par l'homéomorphisme sont égales. On dit que la probabilité est consistante. Ils montrent de plus que, pour chaque évaluation, la fonction v_θ est continue sur l'ensemble des probabilités consistantes.

Ainsi la formule de récurrence sur l'espace $\Delta_f(K \times \mathbb{N} \times \mathbb{N})$ induit une formule de récurrence sur l'ensemble des probabilités consistantes sur Ω qui est compact. Néanmoins on ne sait pas travailler directement sur cet espace. Par exemple, contrairement aux POMDPs on ne connait pas de distance sous laquelle l'ensemble des probabilités consistantes soit compact et les fonctions v_θ soient équicontinues. Ceci impliquerait l'existence d'une valeur d'adhérence pour la norme uniforme par le théorème d'Ascoli, et serait une première étape vers l'existence d'une valeur limite.

Dans le paragraphe suivant, on présente quelques classes de jeux répétés où on peut écrire explicitement la formule de récurrence sur un espace plus petit que Ω. On peut étudier directement les fonctions valeurs sur cet espace ou introduire un jeu stochastique auxiliaire qui satisfait la même équation de récurrence. Dans ce dernier cas, si le jeu auxiliaire a une valeur limite, alors le jeu initial a aussi une valeur limite. Dans certains cas, il est possible de montrer l'existence de la valeur uniforme dans le jeu stochastique puis de déduire l'existence de la valeur uniforme dans le jeu répété original. Contrairement aux cas des MDPs, il peut exister une valeur uniforme dans le jeu stochastique sans que le jeu répété ait une valeur uniforme.

1.4 Exemples de jeux répétés

1.4.1 Jeux répétés avec information incomplète d'un coté

Le premier modèle de jeu répété a été introduit par Shapley [Sha53]. Aumann et Maschler (voir référence de 1995 [AMS95] sur leur travaux de 1968-69) ont introduit un autre modèle, afin d'étudier la transmission d'information. En particulier il n'y a pas d'aspect stochastique dans la définition du problème mais une asymétrie d'information entre le joueur 1 et le joueur 2. L'aspect stochastique apparaît lorsqu'on exprime la formule de récurrence. On dit que $\Gamma = (K, C, D, I, J, q, g)$ est un jeu répété avec information incomplète lorsque l'état est fixé à l'étape 1 pour tout le jeu : pour tout $k \in K$, $i \in I$, $j \in J$, la marginale de $q(k, i, j)$ sur K est la masse de Dirac en k. Ainsi à chaque étape, les joueurs jouent un jeu matriciel, donné par la matrice $G^k = (q(k, i, j))_{i \in I, j \in J}$, fixé pour tout le jeu par la probabilité initiale π mais les joueurs n'observent pas quelle matrice est jouée.

Aumann et Maschler se sont d'abord intéressés au cas où le joueur 1 est informé de l'état à l'étape 1 alors que le joueur 2 n'a aucune information. Au cours du jeu, les joueurs voient les actions jouées mais pas les paiements, on parle alors de jeu avec information incomplète d'un coté. Formellement les ensembles de signaux sont donnés par $C = I \times J$, $D = I \times J$, pour tout $(k, i, j) \in K \times I \times J$, le paiement est donné par $g(k, i, j) = G^k(i, j)$ et la transition par $q(k, i, j) = \delta_{k,(i,j),(i,j)}$. Ils se restreignent aux probabilités initiales à supports dans $\{(k, k, 1), k \in K\}$ où le signal du joueur 2 ne lui apprend rien et le joueur 1 déduit l'état de son signal. Comme

cet ensemble est en bijection avec $\Delta(K)$, pour tout $p \in \Delta(K)$, on note $\Gamma(p)$ le jeu où l'état k est tiré suivant p. Dans ce jeu, pour une probabilité initiale p, le joueur 2 doit choisir une action alors que le joueur 1 doit choisir une action pour chaque état possible. On définit donc $A = \Delta(I)^K$ et pour tout $p \in \Delta(K)$, $a \in A$, $j \in J$

$$\widetilde{g}(p, a, j) = \sum_k p^k G^k(a^k, j).$$

$\hat{p}(a, i)$ est la nouvelle croyance sur l'état, calculée par la relation de Bayes, lorsque le joueur 2 a une croyance initiale p, observe l'action i et sait que le joueur 1 utilise la stratégie $a \in \Delta(I)^K$ à l'étape 1 :

$$\forall k' \in K,\ \hat{p}(a, i)(k') = \frac{p^{k'} a^{k'}(i)}{\sum_k p^k a^k(i)},$$

si le dénominateur est non nul. Si la probabilité que l'action i soit jouée est nulle on définit $\hat{p}(a, i)$ de manière arbitraire.

Les auteurs montrent que la formule de récurrence pour les jeux $\Gamma_n(p)$ s'écrit dans ce contexte

$$v_n(p) = \max_{a \in \Delta(I)^K} \min_{j \in J} \left(\frac{1}{n} \widetilde{g}(p, a, j) + \frac{n-1}{n} \sum_{k \in K, i \in I} p^k a^k(i) v_{n-1}(\hat{p}(a, i)) \right).$$

Il a été montré ensuite que cette équation est l'équation associée au jeu stochastique $\Psi = (X, A, J, \widetilde{g}, \widetilde{q})$ joué en stratégies pures, où $X = \Delta(K)$, $A = \Delta(I)^K$ et $\widetilde{q}$ est une application de $X \times A \times B$ dans $\Delta_f(X)$ définie par

$$\forall\ (p, a, b) \in X \times A \times B,\quad \widetilde{q}(p, a, b) = \sum_{i \in I, k \in K} p^k a^k(i) \delta_{\hat{p}(a,i)}.$$

Le problème du joueur 1 est donc de choisir entre deux options : ne pas utiliser son information et donc ne pas la révéler, ou l'utiliser pour gagner plus à l'étape courante mais en révéler une partie au joueur 2. Aumann et Maschler ont montré que $\Gamma(p)$ a une valeur limite et une valeur uniforme v^* caractérisée par

$$v^* = cav f^* = \inf\{w : \Delta(K) \to [0, 1], w \text{ concave } w \geq f^*\},$$

où $f^*(p) = Val\left(\sum_k p^k G^k\right)$ pour tout $p \in \Delta(K)$. La fonction f^* est la valeur du jeu, appelé non révélateur, où le joueur 1 ne révèle pas son information. Comme le joueur 1 utilise la même stratégie quel que soit l'état, la croyance du joueur 2 ne change pas.

La particularité importante de ce modèle est l'irréversibilité de la révélation d'information. Étant donné une stratégie du joueur 1, le processus des croyances du joueur 2 forme une martingale bornée et donc convergente. D'autre part, l'écart des paiements entre une stratégie σ, qui utilise l'information, et une stratégie optimale dans le jeu non révélateur, qui garantit le paiement d'étape f^*, est contrôlé par la variation L^1 de cette martingale. Ainsi la valeur

converge vers $cavf^*$ (pour une présentation détaillée de leur preuve, le lecteur peut se référer à Sorin [Sor02]). Enfin, les deux joueurs peuvent garantir $cavf^*$ dans le jeu initial. Le joueur 1 peut garantir f^* en jouant non révélateur et $cavf^*$ en choisissant à la première étape de révéler de l'information pour obtenir la concavification. Le joueur 2, quant à lui, peut la garantir en jouant par blocs de plus en plus longs et en oubliant à chaque fois le passé. Comme il n'influence pas la transition, il garantit ainsi la limite de v_n. Les stratégies optimales des deux joueurs sont très différentes car seul le joueur 1 peut utiliser une stratégie du jeu auxiliaire Ψ dans le jeu Γ. Lorsqu'il joue le jeu répété Γ, le joueur 2 ne connait pas la stratégie du joueur 1 et donc ne peut pas calculer de croyance sur l'état et utiliser une stratégie optimale dans le jeu stochastique auxiliaire.

1.4.2 Jeux répétés avec information incomplète des deux cotés

Lorsque les deux joueurs ont une information privée, Aumann et Maschler [AMS95] ont montré que la valeur uniforme n'existe pas. Si chaque joueur a une incitation à révéler son information après l'autre, le $maxmin$ et le $minmax$ sont différents. Ils considèrent un modèle où l'état a deux composantes $K \times L$ et le joueur 1 est informé de la première coordonnée, k, alors que le joueur 2 est informé de la seconde coordonnée, l. Si la valeur uniforme n'existe pas, la valeur limite existe encore dans ce modèle. L'équation de récurrence s'écrit désormais sur l'espace produit de la croyance du joueur 2 sur la coordonnée k, $\Delta(K)$, et de la croyance du joueur 1 sur la coordonnée l, $\Delta(L)$, soit $\Delta(K) \times \Delta(L)$. Les espaces d'actions, quant à eux, deviennent $A = \Delta(I)^K$ et $B = \Delta(J)^L$. Soit $(p, r) \in \Delta(K) \times \Delta(L)$, $a \in A$ et $b \in B$, alors on note

$$\widetilde{g}((p, r), a, b) = \sum_{k,l} p^k r^l g((k, l), a^k, b^l),$$

tandis que les nouvelles croyances sont données par

$$\forall k' \in K,\ \hat{p}(a, i)(k') = \frac{p^{k'} a^{k'}(i)}{\sum_k p^k a^k(i)} \text{ et } \forall l' \in L,\ \hat{r}(b, j)(l') = \frac{r^{l'} b^{l'}(j)}{\sum_l r^l b^l(j)}.$$

La formule de récurrence est

$$v_n(p, r) = \max_{a \in \Delta(I)^K} \min_{b \in \Delta(J)^L} \left(\frac{1}{n} \widetilde{g}\left((p, r), a, b\right) + \frac{n-1}{n} \sum_{k,l,i,j} p^k a^k(i) r^l b^l(j) v_{n-1}\left(\hat{p}(a, i), \hat{r}(b, j)\right) \right).$$

Formellement, les jeux répétés avec information incomplète des deux cotés sont donnés par la même transition que pour les jeux répétés avec information incomplète d'un coté mais la probabilité initiale est différente et il en découle une expression plus complexe de l'équation de récurrence et du jeu stochastique auxiliaire. Comme précédemment, si on note $\widetilde{q}$ la transition

de $X \times A \times B$ dans $\Delta_f(X)$, où $X = \Delta(K) \times \Delta(L)$, donnée par

$$\forall (p,r) \in X,\ a \in A,\ b \in B,\ \widetilde{q}((p,r),a,b) = \sum_{k,l,i,j} p^k a^k(i) r^l b^l(j) \delta_{(\hat{p}(a,i),\hat{r}(b,j))},$$

alors l'équation de récurrence précédente est l'équation associée au jeu $\Psi = (X, A, B, \widetilde{q}, \widetilde{g})$. Dans ce modèle, on note que la croyance du joueur 1 sur la composante L est indépendante de la composante k, on parle d'information indépendante.

Mertens et Zamir [MZ72], [MZ80] ont montré que le jeu répété Γ a une valeur limite et que c'est l'unique solution w de

$$(a)\ w \geq Cav_p Vex_r max(f^*, w),$$
$$(b)\ w \leq Vex_r Cav_p min(f^*, w),$$

ou de

$$(a')\ w = Vex_r max(f^*, w),$$
$$(b')\ w = Cav_p min(f^*, w).$$

avec f^* la valeur du jeu non révélateur.

Leurs résultats s'appliquent en fait au cas plus général où l'ensemble d'état est K, l'information du joueur 1 est une partition K^1 de K et l'information du joueur 2 est une partition K^2 de K. A l'étape initiale, l'état est tiré suivant une probabilité p sur K, le joueur 1 observe l'ensemble de sa partition contenant k et le joueur 2 de même. De plus à chaque étape, les joueurs n'observent pas les actions mais des signaux dépendants des actions jouées mais indépendants de l'état. Dans ce cadre, les notions de concavifié et de convexifié sont alors à considérer par rapport à ces partitions et la notion de jeu non révélateur par rapport aux signaux

1.4.3 Famille de jeux stochastiques avec information incomplète

Afin d'étudier l'aspect information incomplète et l'aspect jeu stochastique ensemble, une première étape est de considérer une famille de jeux stochastiques. A l'étape initiale, un jeu stochastique est tiré aléatoirement selon une probabilité p, puis le jeu stochastique est joué normalement. Les joueurs observent l'état du jeu stochastique. La littérature s'est concentrée sur le cas où le joueur 1 est informé du jeu stochastique joué alors que le joueur 2 ne le sait pas. Comme il y a un nombre fini d'états et d'actions, on peut supposer que tous les jeux stochastiques ont le même ensemble d'états et les mêmes ensembles d'actions. Ils diffèrent donc seulement par leur fonction de paiement g^k et leur fonction de transition q^k. Formellement c'est un jeu répété avec notre modèle général $\Gamma = (K', I, J, C, D, q, g)$ où l'espace d'états est divisé en deux parties K et Ω, et $K' = K \times \Omega$. La coordonnée K ne change plus après l'étape 1, et pour chaque $k \in K$, la transition et la fonction de paiement vérifient $q(k,.) = q^k$ et $g(k,.) = g^k$. Á chaque étape, les joueurs observent les actions jouées et le nouvel état dans Ω de la partie

"jeu stochastique", $C = D = \Omega \times I \times J$. Comme pour les jeux répétés, on se restreint aux probabilités initiales telles que le joueur 1 connait la coordonnée k et le joueur 2 ne la connait pas.

La formule de récurrence peut alors s'exprimer sur l'espace auxiliaire $\Delta(K) \times \Omega$, produit de la croyance du joueur 2 sur le jeu joué et de l'état du jeu stochastique. Si on note $A = \Delta(I)^{K \times \Omega}$ et $B = \Delta(J)^{\Omega}$ alors la formule de récurrence s'écrit pour tout $n \geq 1$,

$$v_n(p,\omega) = \max_{a \in A} \min_{b \in B} \left(\frac{1}{n} \widetilde{g}((p,\omega),a,b)) + \frac{n-1}{n} \sum_{i,j,\omega,\omega'} q(p,a,b)(i,j,\omega,\omega') v_{n-1}(\hat{q}(p,a)(.|i,j,\omega,\omega'),\omega') \right),$$

où $\widetilde{g}((p,\omega),a,b) = \sum_{k \in K} p^k g^k(\omega, a^{k,\omega}, b^{\omega})$, $q(p,a,b)(i,j,\omega,\omega')$ est la probabilité que l'histoire observée par le joueur 2 soit (i,j,ω,ω'),

$$q(p,a,b)(i,j,\omega,\omega') = \sum_{k \in K} p^k a^{k,\omega}(i) b^{\omega}(j) q^k(\omega,i,j)(\omega'),$$

et $\hat{q}(p,a)(.|i,j,\omega,\omega')$ est la nouvelle croyance du joueur 2 calculée par la règle de Bayes,

$$\forall k' \in K,\ \hat{q}(p,a)(.|i,j,\omega,\omega')(k') = \frac{p^{k'} a^{k',\omega}(i) q^{k'}(\omega,i,j)(\omega')}{\sum_k p^k a^{k,\omega}(i) q^k(\omega,i,j)(\omega')}.$$

Les premiers exemples considérés étaient des familles de jeux similaires au Big Match. Le joueur 1 est à la fois informé et choisit si l'absorption peut avoir lieu ou pas ; néanmoins l'état d'absorption dépend de l'action du joueur 2. Sorin [Sor84] [Sor85] puis Sorin et Zamir [SZ91] ont montré l'existence de la valeur limite dans plusieurs de ces cas. Rosenberg et Vieille [RV00] ont ensuite prouvé que lorsque les jeux stochastiques sont récursifs, le *maxmin* peut être défendu uniformément par le joueur 2. Les valeurs des jeux avec n étapes et les valeurs des jeux escomptés, quant à elles, convergent vers ce *maxmin*. Par contre dans ces deux modèles la valeur uniforme n'existe pas.

Le résultat de Rosenberg et Vieille est un résultat sans valeur uniforme, où la valeur limite est égale au *maxmin*. De plus leur preuve montre l'importance d'obtenir une équation de récurrence sur un espace compact tel que les fonctions valeurs soient équicontinues. En effet par le théorème d'Ascoli, ils prouvent l'existence d'une valeur d'adhérence v à la suite v_λ lorsque λ tend vers 0 qui sert de point de départ pour la construction d'une stratégie optimale dans le jeu stochastique. Le joueur 1 choisit λ^* proche de 0 tel que $\|v - v_{\lambda^*}\|_\infty \leq \varepsilon^2$ puis alterne deux stratégies : si la trajectoire atteint un point où $v > \varepsilon$, le joueur 1 suit σ_{λ^*} optimale dans Γ_{λ^*} et si la trajectoire atteint un point où $v < 0$, le joueur 1 suit une stratégie telle que la valeur ne décroit pas en espérance. A chaque changement, le joueur fait une erreur mais par un argument de martingale reposant sur l'absence de paiement en dehors des états absorbants, ils montrent que l'espérance du nombre de changements est bornée et que cette stratégie garantit v. Comme le joueur 2 peut faire la procédure inverse en meilleure réponse, le *maxmin* peut être défendu et v_n converge vers cette limite.

Lorsque la transition des jeux stochastiques ne dépend pas de la coordonnée sur K : l'unique différence entre les différents jeux stochastiques est la fonction de paiement. Dans ce cas la mise à jour de la croyance se simplifie et le joueur 2 n'a plus d'influence sur l'évolution de sa croyance et on trouve un jeu stochastique auxiliaire dont la première coordonnée est controlée par le joueur 1. Rosenberg [Ros00] prouve l'existence de la valeur limite pour les jeux absorbants en étudiant le membre de droite de la formule de récurrence comme un opérateur sur l'espace des fonctions et ayant pour argument v_{n-1}. Rosenberg, Solan et Vieille [RSV04] ont ensuite démontré l'existence de la valeur uniforme pour les jeux où le joueur informé contrôle la transition, ainsi il contrôle totalement la transition du jeu auxiliaire. Si c'est le joueur non informé qui contrôle la transition originale alors la transition du jeu stochastique auxiliaire dépend des deux joueurs. Il n'existe pas de valeur uniforme et l'existence de la valeur limite dans ce dernier cas est encore ouverte.

1.4.4 Jeux avec acquisition d'information pendant le cours du jeu

Dans tous les modèles précédents la révélation d'information exogène sur l'état est faite à l'étape initiale et ensuite les joueurs s'échangent ou non cette information initiale. On peut aussi directement considérer un jeu stochastique et des signaux quelconques. Par exemple, Kohlberg et Zamir [KZ74] ont étudié le modèle des jeux répétés avec information incomplète avec des signaux symétriques. Comme dans le modèle d'Aumann et Maschler, une matrice est tirée à l'étape initiale suivant une probabilité p mais aucun joueur ne reçoit d'information supplémentaire. Par contre durant le cours du jeu, les joueurs observent les actions et un signal public, observé par tous et qui dépend de l'état et des actions jouées. Chaque joueur doit donc tenir compte lorsqu'il joue du paiement courant et de la révélation commune sur l'état que va entraîner son action. Ainsi dans certains cas, le joueur 1 préfère soit que tout le monde connaisse l'état soit que personne ne le connaisse. Formellement c'est un jeu répété avec un espace d'états K, des ensembles d'actions I et J, deux ensembles de signaux $C = D = I \times J \times U$ définis à partir d'un ensemble de signaux publiques U, une fonction de paiement g et une transition q telle que pour tout $k \in K$, $i \in I$ et $j \in J$, la marginale de $q(k, i, j)$ sur K est la masse de Dirac en k (jeu répété à la Aumann et Maschler) et l'information est symétrique

$$\sum_{k' \in K,\ u \in U} q(k, i, j)(k', (i, j, u), (i, j, u)) = 1.$$

Kohlberg et Zamir [KZ74] prouvent l'existence de la valeur uniforme lorsque les signaux sont déterministes. Ce résultat a été ensuite étendu au signaux probabilistes par Forges [For82]. Dans ce cas la formule de récurrence peut s'écrire sur l'espace $\Delta(K)$. Le nouvel état est la croyance commune des joueurs. Les ensembles d'actions sont les mêmes que dans le jeu répété. Pour tout $p \in \Delta(K)$, $a \in \Delta(I)$ et $b \in \Delta(J)$, on définit l'extension multilinéaire de g

$$\widetilde{g}(p, a, b) = \sum_{k \in K} p^k g(k, a, b),$$

$q(p,a,b)(i,j,u) = \sum_{k,k'} p^k a(i) b(j) q(k,i,j)(k',u)$, la probabilité que le signal (i,j,u) soit observé et la nouvelle croyance si (i,j,u) est observé

$$\forall k' \in K,\ \hat{q}(p,a,b|i,j,u)(k') = \frac{\sum_{k\in K} p^k q(k,i,j)(k',u)}{\sum_{k,k'\in K} p^k q(k,i,j)(k',u)}.$$

Comme la croyance précédente ne dépend ni de a ni de b, on l'écrira dans la suite $\hat{q}(p|i,j,u)$. La formule de récurrence s'écrit alors

$$v_n(p) = \sup_{a\in\Delta(I)} \inf_{b\in\Delta(J)} \left(\frac{1}{n}\tilde{g}(p,a,b)) + \frac{n-1}{n} \sum_{i,j,u} q(p,a,b)(i,j,u) v_{n-1}(\hat{q}(p|i,j,u)) \right).$$

Les deux joueurs mettent à jour leur croyances en fonction des actions jouées et du signal public. On définit le jeu stochastique auxiliaire $\Psi = (\Delta(K), I, J, \tilde{g}, \tilde{q})$ où

$$\tilde{q}(p,i,j) = \sum_{s\in S} q(p,i,j)(s) \delta_{\hat{q}(p|i,j,u)}.$$

La nouvelle croyance ne dépend pas des stratégies, les deux joueurs peuvent donc estimer la croyance commune dans le jeu répété sans connaître la stratégie de l'autre joueur. Geitner [Gei02] a étendu ces résultats sur les jeux répétés au cas où un jeu stochastique est tiré à l'étape initiale et montré que le problème se réduit à la résolution d'une succession de MDPs avec un nombre fini d'états et d'actions.

Renault [Ren06] s'est intéressé à une autre généralisation du modèle d'Aumann et Maschler : l'état évolue selon une chaîne de Markov indépendamment des décisions des joueurs. À chaque étape, le joueur 1 observe les actions jouées et le nouvel état, alors que le joueur 2 n'observe que les actions. Contrairement au cas précédent, la révélation d'information n'est plus irréversible. Par exemple si la chaîne est ergodique et apériodique, alors partant de n'importe quelle distribution initiale, la loi converge vers l'unique distribution invariante. Si le joueur 1 révèle son information puis joue non révélateur pendant suffisamment longtemps, la croyance du joueur 2 revient à la mesure invariante. Il démontre l'existence d'une valeur uniforme. Neyman [Ney08] donne une autre preuve de ce résultat qui se généralise lorsque les joueurs n'observent pas parfaitement les coups de l'autre joueur. Renault [Ren12b] unifie l'article de Rosenberg, Solan et Vieille [RSV04] et son article sur les jeux avec une chaîne de Markov, en montrant l'existence de la valeur uniforme dans le cas d'un jeu répété où le joueur 1 contrôle la transition et peut déduire à chaque étape l'état et le signal du joueur 2. En fait la première hypothèse est affaiblie en l'hypothèse suivante : le joueur 2 n'influe pas sa propre croyance ou formellement.

Hypothèse 1.4.1 *La marginale de la transition sur $K \times D$ n'est pas influencée par les actions du joueur 2 . Pour k dans K, i dans I et j dans J, on note $\bar{q}(k,i)$ la marginale de $q(k,i,j)$ sur $K \times D$.*

La seconde hypothèse, que le joueur 1 apprend l'état, et, qu'il est plus informé que le joueur 2, s'écrit de la manière suivante.

Hypothèse 1.4.2 *Il existe deux applications* $\widetilde{k} : C \to K$ *et* $\widetilde{d} : C \to D$ *telles que si* E *représente* $\{(k,c,d) \in K \times C \times D,\ \widetilde{k}(c) = k,\ \widetilde{d}(c) = d\}$, *on a :* $\forall (k,i,j) \in K \times I \times J$, $q(k,i,j)(E) = 1$.

De plus on se restreint à l'ensemble, noté Ω, des probabilités initiales compatibles avec la structure d'information : c'est dire les probabilités $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$ telles qu'il existe $\widetilde{k_0} : \mathbb{N} \to K$ et $\widetilde{d_0} : \mathbb{N} \to \mathbb{N}$ avec $\pi(E) = 1$ où $E = \{(k,c,d) \in K \times \mathbb{N} \times \mathbb{N},\ \widetilde{k_0}(c) = k,\ \widetilde{d_0}(c) = d\}$. Pour une distribution π fixée, quitte à changer les espaces de signaux, on peut supposer que π est une probabilité sur $K \times C \times D$. Renault montre qu'il existe un jeu stochastique auxiliaire sur l'espace $X = \Delta(K)$ où l'espace d'action du joueur 1 est $A = \Delta(I)^K$ et celui du joueur 2 est $B = \Delta(J)$. Dans le cas simple où la transition du jeu répété ne dépend pas du joueur 2, on note

$$\widetilde{g}(p,a,b) = \sum_{(k,i,j) \in K \times I \times J} p^k a^k(i) b(j) g(k,i,j) \in [0,1]$$

et étant donné $p \in X$, $a \in A$ et $i \in J$, on note $\hat{q}(p,a)(.|i) \in X$ la croyance du joueur 2 après une étape

$$\forall k' \in K,\ \hat{q}(p,a)(.|i)(k') = \frac{q(p,a)(k',i)}{\sum_{k \in K} q(p,a)(k,i)},$$

où $q(p,a)(k',i) = \sum_k p^k a^k(i) q(k,i)(k')$ est la probabilité que l'action i soit jouée et l'état k tiré si le joueur 1 joue selon a. La formule de récurrence s'écrit alors

$$\forall n \geq 1,\ \forall p \in X,\quad v_n(p) = \sup_{a \in A} \inf_{j \in J} \left(\frac{1}{n} g(p,a,j) + \frac{n-1}{n} \left(\sum_i p^k a^k(i) v_{n-1}\left(\hat{q}(p,a)(.|i)\right) \right) \right).$$

1.5 Résultats de la thèse

1.5.1 Généralisation de la valeur limite et de la valeur uniforme à des évaluations quelconques.

Ce chapitre est extrait d'un article écrit en collaboration avec Jérôme Renault. Les notions de valeur limite et de valeur uniforme considèrent les moyennes de Cesàro des paiements entre l'étape 1 et une étape $n \geq 1$. Dans les cas simples tels que les MDPs avec des espaces finis, on sait qu'il existe une unique stratégie optimale pour tous les taux d'escomptes suffisamment petits et pour les moyennes de Cesàro entre 1 et n pour tout $n \in \mathbb{N}^*$ suffisamment grand. Cette stratégie garantit aussi la valeur uniforme. Sorin [Sor02] souligne que cette stratégie garantit la valeur limite pour toute une classe d'évaluations.

Soit $\theta \in \Delta(\mathbb{N}^*)$ une évaluation. Si la probabilité initiale est π et si les joueurs jouent

respectivement selon la stratégie σ et la stratégie τ, on rappelle que le paiement selon θ, s'écrit

$$\gamma_\theta(\pi,\sigma,\tau) = I\!E_{\pi,\sigma,\tau}\left(\sum_{t\geq 1}\theta_t g(k_t,i_t,j_t)\right).$$

On définit l'irrégularité de l'évaluation θ par

$$I(\theta) = \sum_{t\geq 1}|\theta_{t+1}-\theta_t|.$$

Plus $I(\theta)$ est proche de zero, plus l'évaluation θ est régulière. En particulier lorsque l'évaluation est décroissante $I(\theta)=\theta_1$ et cette mesure d'irrégularité coïncide avec une mesure de l'impatience des joueurs. L'irrégularité du jeu avec n étapes est $\frac{1}{n}$ et celle du jeu escompté de paramêtre λ est λ. On définit ainsi des notions plus fortes que la valeur limite et la valeur uniforme qui considèrent tous les types d'évaluations lorsque l'irrégularité tend vers 0.

Définition 1.5.1 *Le jeu répété $\Gamma(\pi)=(K,I,J,C,D,q,g,\pi)$ a une valeur limite générale $v^*(\pi)$ si $v_\theta(\pi)$ converge vers $v^*(\pi)$ lorsque $I(\theta)$ tend vers zero :*

$$\forall\varepsilon>0,\exists\alpha>0,\forall\theta,\ \ (\ I(\theta)\leq\alpha\implies(|v_\theta(\pi)-v^*(\pi)|\leq\varepsilon)\ ).$$

Définition 1.5.2 *Le jeu répété $\Gamma(\pi)$ a une valeur uniforme générale s'il existe une valeur limite générale v^* et pour tout $\varepsilon>0$, il existe $\alpha>0$ et un couple de stratégies σ^* et τ^* tel que pour chaque évaluation θ telle que $I(\theta)\leq\alpha$:*

$$\forall\tau\in\mathcal{T},\gamma_\theta(\pi,\sigma^*,\tau)\geq v^*(\pi)-\varepsilon\ \text{ et }\ \forall\sigma\in\Sigma,\gamma_\theta(\pi,\sigma,\tau^*)\leq v^*(\pi)+\varepsilon.$$

Des stratégies optimales pour la valeur uniforme classique peuvent ne pas être optimales pour la notion générale et la valeur limite générale peut ne pas exister alors que la valeur limite existe.

Exemple 1.5.3 *Soit $(r_k)_{k\geq1}$ une suite de $[0,1]$ définie par*

$$\begin{aligned} r_k &= 0 \text{ s'il existe } l\geq 1 \text{ tel que } k\in[2^l,2^l+l],\\ &= 1 \text{ sinon.}\end{aligned}$$

Alors les moyennes de Cesàro convergent vers 1, donc le "jeu" a une valeur limite mais il n'a pas de valeur limite générale. En effet soit $\alpha>0$, si on note $l=[\frac{2}{\alpha}]$ la partie entière de $2/\alpha$, alors l'évaluation uniforme entre la date 2^l et la date 2^l+l a une irrégularité plus petite que α et la valeur, égale à l'unique paiement possible, est 0.

On prouve que les POMDPs avec un nombre fini d'états ainsi que les jeux répétés finis avec un contrôleur informé admettent une valeur uniforme générale.

Théorème 1.5.4 *Soit $\Gamma = (K, A, S, q, g)$ un POMDP tel que K est fini tandis que A et S sont quelconques, alors pour tout $\pi \in \Delta_f(K \times \mathbb{N})$, $\Gamma(\pi)$ admet une valeur uniforme générale. De plus la convergence est uniforme sur $\Delta_f(K \times \mathbb{N})$.*

Pour les jeux répétés, on se restreint aux probabilités initiales compatibles avec l'hypothèse d'un joueur 1 parfaitement informé.

Théorème 1.5.5 *Soit $\Gamma = (K, I, J, C, D, q, g)$ un jeu répété fini qui satisfait les hypothèses 1.4.2 et 1.4.1, i.e. avec un contrôleur informé. Pour toute probabilité initiale π compatible avec l'hypothèse d'un joueur 1 plus informé, $\Gamma(\pi)$ admet une valeur uniforme générale.*

Ces deux résultats sont obtenus à partir de l'introduction d'une nouvelle distance sur l'espace $\Delta_f(\Delta(K))$. Muni de cette distance, on montre l'existence de la valeur uniforme générale pour une classe de MDPs sur $\Delta(K)$. On déduit alors l'existence de la valeur uniforme dans les POMDPs et les jeux répétés avec un contrôleur informé en introduisant un MDP auxiliaire. Dans les cas des jeux répétés avec un contrôleur informé, on introduit d'abord un jeu stochastique. Comme le joueur 2 n'influence pas la transition, il joue à chaque étape uniquement pour le paiement courant et ce jeu stochastique se réduit à un MDP.

Le chapitre présente en plus un autre résultat sur les maisons de jeu où la preuve est similaire, plus simple et pour des espaces d'états métriques compacts mais qui ne suffit pas pour les applications aux POMDPs et aux jeux répétés avec un contrôleur informé.

a) Une distance sur $\Delta_f(\Delta(K))$.

On décrit une nouvelle façon d_* de mesurer l'écart entre des probabilités sur un ensemble X d'un espace vectoriel normé $(V, \|.\|)$. Une manière de mesurer la distance entre deux probabilités sur X est la distance de Kantorovitch Rubinstein qui peut s'écrire de différentes manières (voir Villani [Vil03]). Si u et v sont deux probabilités à support fini, dont les supports sont respectivement notés U et V, alors la distance de Kantorovitch-Rubinstein est donnée par :

$$d_{KR}(u, v) = \sup_{f \in E_1} |u(f) - v(f)| = \min_{\chi \in \Pi(u,v)} \sum_{(x,y) \in U \times V} \|x - y\|_1 \chi(x, y)$$

où E_1 est l'ensemble des fonctions 1-Lipschitz pour $(X, \|.\|)$ et $\Pi(u, v)$ est l'ensemble des probabilités sur $\Delta(X \times X)$ telles que la première marginale est u et la seconde est v, aussi appelées couplages entre u et v.

Étant donnée une évaluation θ, la fonction valeur v_θ restreinte à $(\Delta(K), \|.\|_1)$ est une application 1-Lipschitz et son extension linéaire à $\Delta_f(\Delta(K))$ est 1-Lipschtiz pour la distance de Kantorovitch-Rubinstein. Comme cette distance rend l'espace compact, les fonctions valeurs forment une famille de fonctions équicontinues. Ces propriétés ont été déjà utilisées dans Renault [Ren11] et dans Rosenberg, Solan et Vieille [RSV02]. Néanmoins les transitions du MDP auxiliaire associé au POMDP ne sont pas 1-Lipschitz de $(X, \|.\|_1)$ dans $(\Delta_f(X), d_{KR})$.

La distance d_* est définie en ne considérant qu'un sous-ensemble des fonctions test E_1 ou symétriquement en autorisant des couplages plus généraux. On note $\mathcal{M}(X)$ les mesures boréliennes sur X. Le premier résultat montre que d_* peut être définie de plusieurs manières soit à partir de fonctions tests soit à partir d'une mesure jointe à la manière de la distance de Kantorovitch-Rubinstein.

Théorème 1.5.6 *Soit X un ensemble compact d'un espace vectoriel normé $(V, \|.\|)$. On note E l'ensemble des fonctions continues de X dans $\mathbb{R}$. Soient u et v dans $\Delta(X)$ alors on a*

$$d_*(u,v) = \sup_{f\in D_1} |u(f) - v(f)| = \inf_{\gamma\in\mathcal{M}_3(u,v)} \int_{X^2\times[0,1]^2} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu)$$

où

$$D_1 = \{f \in E_1, \forall x, y \in X, \forall a, b \geq 0,\ af(x) - bf(y) \leq \|ax - by\|\},$$

et

$$\mathcal{M}(u,v) = \Bigg\{ \gamma \in \mathcal{M}(X^2\times[0,1]^2) \text{ } t.q. \text{ } \forall f \in E_1,\ \int_{(x,y,\lambda,\mu)\in X^2\times[0,1]^2} \lambda f(x) d\gamma(x,y,\lambda,\mu) = u(f)$$
$$et\ \int_{(x,y,\lambda,\mu)\in X^2\times[0,1]^2} \mu f(y) d\gamma(x,y,\lambda,\mu) = v(f) \Bigg\}.$$

Lorsque l'ensemble $X = \Delta(K)$ est l'ensemble des probabilités sur un ensemble fini K, vu comme un simplexe dans $\mathbb{R}^K$, alors d_* est effectivement une distance sur l'ensemble des probabilités à support fini $\Delta_f(X)$. Dans ce cas on peut alors préciser la définition par des couplages afin de faire apparaitre des probabilités.

Théorème 1.5.7 *(Formule de dualité) On munit $\mathbb{R}^K$ de la norme $\|.\|_1$. Soit X un sous ensemble du simplexe $\Delta(K)$ et soient u et v dans $Z = \Delta_f(X)$ avec comme supports respectifs U et V.*

$$d_*(u,v) = \sup_{f\in D_1} |u(f) - v(f)| = \min_{(\alpha,\beta)\in\mathcal{M}'(u,v)} \sum_{(x,y)\in U\times V} \|x\alpha(x,y) - y\beta(x,y)\|_1$$

où

$$D_1 = \{f \in E_1, \forall x, y \in X, \forall a, b \geq 0,\ af(x) - bf(y) \leq \|ax - by\|_1\},$$

et

$$\mathcal{M}'(u,v) = \Big\{(\alpha,\beta) \in {\mathbb{R}_+}^{U\times V} \times {\mathbb{R}_+}^{U\times V} \text{ } t.q. \text{ } \forall (x,y) \in U\times V,$$
$$\sum_{y'\in V} \alpha(x,y') = u(x) \text{ } et \text{ } \sum_{x'\in U} \beta(x',y) = v(y) \Big\}.$$

Si on compare les définitions des distances d_{KR} et d_* par les formules avec des supremum, on note que la distance d_* considère un sous-ensemble des fonctions 1-Lipschitz. Elle est donc plus petite que la distance de Kantorovitch-Rubinstein et on vérifie qu'elle métrise la topologie

faible-$*$. Il est naturel de considérer un sous-ensemble des fonctions de E_1 car nous allons nous intéresser uniquement aux fonctions valeurs provenant de jeux aux paiements entre 0 et 1. En fait, l'ensemble D_1 peut être remplacé par l'ensemble D_0 des valeurs des jeux non révélateurs avec paiement dans $[-1, 1]$ définis dans le cadre des jeux répétés avec information incomplète d'un coté.

Une propriété importante de cette distance est de rendre 1-Lipschitz la désintégration, introduite lors de l'étude des POMDPs. Rappelons que ψ_N est définie de $\Delta_f(K \times \mathbb{N})$ dans $\Delta_f(\Delta(K))$ par : pour tout $\pi \in \Delta(K \times \mathbb{N})$,

$$\psi_{\mathbb{N}}(\pi) = \sum_{c' \in \mathbb{N}} \pi(c')\delta_{p_1(c')},$$

où $p_1(c') = I\!P_{\pi,\sigma}(k_1|h_1^1) = \left(\frac{\pi(k,c')}{\sum_{k' \in K} \pi(k',c')}\right)_{k \in K} \in X = \Delta(K)$.

Proposition 1.5.8 *L'application $\psi_{\mathbb{N}}$ est 1-Lipschitz de $(\Delta(K \times \mathbb{N}), \|.\|_1)$ dans $(\Delta_f(X), d_*)$ mais pas pour d_{KR}.*

On déduit de cette proposition que les transitions sont 1-Lipschitz de $(X, \|.\|_1)$ dans $(\Delta_f(X), d_*)$. En effet si on considère deux probabilités initiales p et p' et une action i, alors la distance en norme 1 entre les deux probabilités sur les histoires de longueur 2 est plus petite que $\|p - p'\|_1$. Comme la désintégration est aussi 1-Lipschitz, la transition du MDP auxiliaire est 1-Lipschitz. Ce résultat implique en particulier que les fonctions valeurs sont équicontinues. On va utiliser cette nouvelle régularité sur les transitions pour montrer l'existence d'une valeur limite générale et d'une valeur uniforme générale.

b) Les maisons de jeux "compactes".

On démontre en premier l'existence de la valeur uniforme générale lorsque X est un espace métrique compact quelconque et $\Delta(X)$ est métrisé par la distance de Kantorovitch-Rubinstein. Une maison de jeu est donnée par un ensemble d'état X, une correspondance $F : X \rightrightarrows \Delta_f(X)$ et r une fonction de X dans $[0, 1]$. Ainsi à l'étape $t \geq 1$, le décideur choisit une distribution u_{t+1} dans $F(x_t)$, l'état x_{t+1} est tiré selon u_{t+1} et le décideur gagne le paiement $r(x_{t+1})$. La principale différence entre un MDP et une maison de jeu concerne le paiement qui ne dépend que du nouvel état alors que dans un MDP, deux actions peuvent donner le même état et des paiements différents. Les deux modèles sont en fait équivalents quitte à augmenter l'espace d'états pour y inclure le paiement.

Étant donnée $\Gamma = (X, F, r)$, une maison de jeu, on définit l'extension linéaire de r à $Z = \Delta_f(X)$ et l'extension mixte de F pour tout $u = \sum_{x \in X} u(x)\delta_x \in \Delta_f(X)$ par $r(u) = \sum_{x \in X} r(x)u(x)$ et

$$\hat{F}(u) = \left\{ \sum_{x \in X} u(x)f(x),\ t.q.\ f : X \to Z \text{ et } f(x) \in \text{conv}F(x)\ \forall x \in X \right\}.$$

Par définition r est affine et on vérifie que la correspondance est aussi affine. Une stratégie comportementale σ partant de $x_0 \in X$ pour la maison de jeu $\Gamma = (X, F, r)$ est une suite $(u_t)_{t\in\mathbb{N}^*}$ d'états dans $Z = \Delta_f(X)$ telle que $u_1 \in ConvF(x_0)$ et $u_{t+1} \in \hat{F}(u_t)$ pour tout $t \in \mathbb{N}^*$. Le paiement pour une évaluation $\theta \in \Delta_f(\mathbb{N}^*)$ est, quant à lui, donné par

$$\gamma_\theta(\sigma) = \sum_{t\geq 1} \theta_t r(u_t).$$

La valeur pour l'évaluation θ est le supremum sur toutes les stratégies comportementales du décideur. La notion de valeur limite générale et de valeur uniforme générale sont alors les mêmes que pour les jeux répétés.

S'inspirant de l'étude des chaînes de Markov avec un ensemble d'états fini, on cherche une notion de mesures invariantes. Bien que le processus ne soit pas défini sur $\Delta(X)$ tout entier, on peut quand même définir une notion d'invariance.

Définition 1.5.9 *Une probabilité $u \in \Delta_f(X)$ est une mesure invariante de la maison de jeu $\Gamma = (X, F, r)$ si $(u, u) \in cl(Graph(\hat{F}))$ et on note cet ensemble R.*

Sous l'hypothèse que les transitions soient 1-Lipschitz, ces mesures sont un substitut aux mesures invariantes et pour un état initial dans le voisinage de R, le décideur peut rester dans le voisinage de R. On déduit l'existence de la valeur uniforme générale et une caractérisation de cette valeur. Lorsque X est compact métrique et r est continue, alors l'extension affine de r est continue sur $\Delta(X)$ et on peut écrire le théorème suivant.

Théorème 1.5.10 *Soit $\Gamma = (X, F, r)$, une maison de jeu, telle que X est compact métrique, r est continue et F est non expansive par rapport à la distance de Kantorovitch-Rubinstein :*

$$\forall x \in X, \forall x' \in X, \forall u \in F(x), \exists u' \in F(x') \text{ t.q. } d_{KR}(u, u') \leq d(x, x').$$

Alors la maison de jeu a une valeur uniforme générale v^* caractérisée par :

$$\begin{aligned} \forall x \in X,\ v^*(x) \ = \inf \ \Big\{ & w(x), w : \Delta(X) \to [0,1] \text{ affine } C^0 \text{ t.q.} \\ & (1)\ \forall y \in X, w(y) \geq \sup_{u\in F(y)} w(u) \text{ et } (2) \forall u \in R, w(u) \geq r(u) \Big\}. \end{aligned}$$

Ainsi v^* est la plus petite fonction affine et continue sur X qui est excessive au sens de Choquet [Cho56] et plus grande que r sur les mesures invariantes. De plus elle peut être garantie uniformément en x.

On commence par montrer qu'il existe une valeur limite générale caractérisée par cette formule. On note $w^*(x)$ l'infimum sur toutes les fonctions affine C^0 vérifiant les équations (1) et (2). N'importe quelle valeur d'adhérence de la famille v_θ lorsque $I(\theta)$ tend vers 0 doit vérifier (1) à cause de la formule de récurrence. D'autre part les états de R se comportent comme des

points fixes donc le joueur peut garantir $r(u)$ dans le jeu avec pour état "initial" u quelque soit l'évaluation. Rappelons qu'à priori cela n'a pas de sens de commencer en u car u n'est pas supposée à support fini. Ainsi les valeurs d'adhérence de v_θ vérifient (1) et (2) et $w^*(x)$ est plus petit que n'importe quelle valeur d'adhérence évalué en x.

D'autre part si on considère une suite θ^k telle que $I(\theta^k)$ converge vers 0, et une stratégie $(z_1^k, ..., z_t^k, ...)$, ε-optimale pour l'évaluation θ^k à partir de l'état x, on peut définir une mesure d'occupation :

$$z(k) = \sum_{t\geq 1} \theta_t^k z_t^k.$$

On vérifie que les valeurs d'adhérence de la suite $(z(k))_{k\geq 1}$ sont dans R. En particulier si z est une valeur d'adhérence, la décroissance (équation (1)) implique alors que $w^*(x)$ est plus grand que $w^*(z)$ donc plus grand que $r(z)$ par l'équation (2). D'autre part r est linéaire et z est obtenu à partir d'une stratégie ε-optimale, donc $r(z)$ est plus grand que $\lim_k v_{\theta^k}(x) - \epsilon$. Ceci prouve l'autre inégalité.

On montre ensuite l'existence de la valeur uniforme générale en utilisant une famille de fonctions auxiliaires. Pour une longueur donnée n, on définit $h_{T,n}$ comme le jeu où le décideur choisit une stratégie, la nature choisit une date avant T et le paiement est la moyenne de Cesàro des paiements entre la date T et la date $T+n$. La nature choisit la plus mauvaise date pour le joueur 1. On montre par un théorème de Sion, que $h_{T,n}$ est l'infimum d'une famille de fonctions valeurs dont l'irrégularité est contrôlée uniformément par $\frac{1}{n}$. Ainsi $h_{T,n}$ converge vers $w^*(x)$ uniformément en T puis on construit à l'aide de cette famille une stratégie qui garantit $w^*(x)$ dans le jeu infini.

Néanmoins ce théorème n'est pas suffisant pour les applications aux POMDPs et les jeux répétés avec information incomplète avec un contrôleur informé car les transitions des jeux auxiliaires associés ne satisfont pas les hypothèses. En effet la transition du MDP auxiliaire n'est pas 1-Lipschitz pour la distance de Kantorovitch-Rubinstein.

c) La valeur uniforme générale pour des MDPs "compacts".

Soit X un sous ensemble du simplexe $\Delta(K)$ et g une fonction continue sur X. Afin d'étudier le MDP, $\Gamma = (X, A, q, g)$, on introduit un problème de programmation dynamique Ψ affine sur $Z = \Delta_f(X) \times [0, 1]$. Un problème de programmation dynamique $\Psi = (Z, F, r)$ est une maison de jeu telle que pour tout $z \in Z$, l'ensemble image $F(z)$ est composé uniquement de masses de Dirac ($F : Z \rightrightarrows Z$).

On étend d'abord g et q linéairement par rapport aux actions puis on définit $\Psi = (Z, \hat{F}, r)$ par $Z = \Delta_f(X) \times [0, 1]$ et pour tout $(u, y) \in \Delta_f(X) \times [0, 1]$,

$$r((u, y)) = y,$$

et

$$\hat{F}(u,y) = \left\{ \left(\sum_{x\in X} u(x)q(x,a(x)), \sum_{x\in X} u(x)g(x,a(x)) \right), \text{ où } a : X \to \Delta_f(A) \right\}.$$

Le paiement a été inclus dans l'état. La correspondance $\hat{F}$ et la fonction de paiement r sont affines. Comme pour les maisons de jeu, on définit une notion d'invariance.

Définition 1.5.11 *Un couple $(u,y) \in \Delta(X) \times [0,1]$ est invariant pour le MDP, Γ, si*

$$((u,y),(u,y)) \in cl(Graph(\hat{F})).$$

On note RR l'ensemble des couples invariants pour Γ.

On prouve un résultat d'existence non plus avec la distance de Kantorovitch-Rubinstein mais avec la distance d_* décrite dans le premièr paragraphe qui est la distance adaptée aux applications aux POMDPs et aux jeux répétés.

Théorème 1.5.12 *Soit $\Gamma = (X, A, q, g)$ un MDP où X est un sous ensemble compact d'un simplexe $\Delta(K)$ tel que :*

$$\forall x \in X, \forall y \in X, \forall a \in A, \forall f \in D_1, \forall \alpha \geq 0, \forall \beta \geq 0,$$

$$|\alpha f(q(x,a)) - \beta f(q(y,a))| \leq \|\alpha x - \beta y\|_1 \text{ et } |\alpha g(x,a) - \beta g(y,a)| \leq \|\alpha x - \beta y\|_1,$$

alors Γ a une valeur uniforme générale v^* caractérisée par : pour tout x dans X,

$$\begin{aligned} v^*(x) \quad = \inf \quad & \Big\{ w(x), w : \Delta(X) \to [0,1] \text{ affine } C^0 \text{ t.q.} \\ & (1)\ \forall x' \in X, w(x') \geq \sup_{a\in A} w(q(x',a)) \text{ et } (2)\ \forall (u,y) \in RR, w(u) \geq y \Big\}. \end{aligned}$$

Afin de déduire les théorèmes indiqués au début de la partie, on vérifie que le MDP associé à un POMDP et celui associé à un jeu répété avec un contrôleur informé vérifient ces hypothèses et que l'existence d'une valeur uniforme générale dans le jeu auxiliaire implique l'existence de la valeur uniforme générale dans le jeu original.

1.5.2 Jeux commutatifs

On s'intéresse dans le troisième chapitre à des transitions particulières. La transition d'un jeu stochastique est commutative lorsque l'ordre dans lequel les actions sont jouées n'est pas important pour connaître la distribution sur l'état. Étant donné une transition q de $K \times I \times J$ dans $\Delta(K)$, on définit l'extension linéaire de q par $\tilde{q}$ de $\Delta(K) \times I \times J$ dans $\Delta(K)$

$$\forall p \in \Delta(K),\ \tilde{q}(p,i,j) = \sum_k p^k q(k,i,j).$$

Définition 1.5.13 *Soit $q : K \times I \times J \to \Delta(K)$ la transition d'un jeu stochastique. La transition q est commutative si pour tout $k \in K$, pour tout $(i,j) \in I \times J$ et pour tout $(i',j') \in I \times J$*

$$\widetilde{q}(q(k,i,j),i',j') = \widetilde{q}(q(k,i',j'),i,j).$$

Étant donné un état initial k, si les joueurs jouent le couple d'action (i,j) puis, quel que soit l'état après une étape, jouent le couple d'action (i',j') alors la distribution sur l'état est la même que s'ils avaient joué d'abord (i',j') puis, quel que soit le nouvel état, (i,j). De nombreux jeux peuvent s'exprimer de cette façon : par exemple le jeu stochastique associé à un jeu répété avec information incomplète d'un coté *à la* Aumann et Maschler ou une chaîne de Markov sont des jeux commutatifs. D'autres modèles ne sont pas explicitement des jeux commutatifs mais peuvent être reformulés comme des jeux commutatifs. On montrera comment transformer les jeux absorbants en des jeux commutatifs équivalents.

Si cette définition a du sens pour n'importe quelle transition, elle ne semble pas particulièrement intéressante en général. En effet dans un jeu stochastique, les joueurs à la seconde étape font dépendre leurs actions des informations observées. Or l'hypothèse considère uniquement à l'étape 2, les stratégies qui jouent pareil quelle que soit la réalisation à la première étape. Nous allons voir que pour certaines structures d'information, elle est néanmoins tout à fait adaptée.

a) Jeux répétés symétriques où les joueurs n'observent pas l'état.

Si on considère les jeux répétés où les joueurs observent les actions mais pas les états, la commutation permet d'obtenir de nouveaux résultats, à la fois lorsqu'il y a deux joueurs et dans le cas particulier où il n'y a qu'un seul décideur. Formellement, ces jeux sont le cas particulier des jeux répétés avec information symétrique où il n'y pas de signal public supplémentaire : les ensembles de signaux des joueurs sont réduits à $C = D = I \times J$ et la transition vérifie

$$\sum_{k'} q(k,i,j)(k',(i,j),(i,j)) = 1.$$

Contrairement aux modèles de jeux répétés symétriques présentés précédemment, on considère les modèles généraux où l'état n'est pas fixé à l'étape 1 mais évolue selon les actions jouées par les joueurs. En contrepartie, on se restreint au cas où il n'y a pas de signaux publics et où la transition commute. Dans la suite on omettra les ensembles C et D et on notera ces jeux $\Gamma^{sb} = (K,I,J,q,g)$, "state-blind", afin de les distinguer du jeu stochastique classique $\Gamma = (K,I,J,q,g)$ où les joueurs observent les états et les actions. Pour tout $p \in \Delta(K)$, $\Gamma^{sb}(p)$ est le jeu où l'état initial est tiré selon la probabilité p. Lorsqu'il n'y a qu'un seul joueur, cette classe coïncide avec les MDPs "dans le noir" où le décideur ne reçoit aucune information sur l'état et, où il existe une valeur uniforme (Rosenberg, Solan et Vieille [RSV02]). Dans cet article, les auteurs posent la question de l'existence de stratégies 0-optimales pour ces modèles. En général le résultat est faux mais la commutation est une condition suffisante pour garantir l'existence d'une stratégie 0-optimale.

Exemple 1.5.14 Soit $K = \{\alpha, \beta, k_0, k_1\}$ et $I = \{T, B\}$. Le paiement est 0, excepté dans l'état k_1 où il est égal à 1. Les états k_0 et k_1 sont absorbants et en dehors la transition q est donnée par

$$\begin{aligned} q(\alpha, T) &= \frac{1}{2}\delta_\alpha + \frac{1}{2}\delta_\beta, \\ q(\beta, T) &= \delta_\beta, \\ q(\alpha, B) &= \delta_{k_0}, \\ q(\beta, B) &= \delta_{k_1}. \end{aligned}$$

Une stratégie ε-optimale dans $\Gamma(\delta_\alpha)$ consiste à jouer l'action T jusqu'à ce que la probabilité d'être en β soit assez grande puis B pendant le reste du jeu. La valeur uniforme est 1 mais il n'existe pas de stratégies 0-optimales. Dès que le décideur joue B, il fait une erreur irréversible.

Théorème 1.5.15 *Soit $\Gamma^{sb} = (K, I, q, g)$ un POMDP dans le noir, avec un nombre fini d'états et un nombre fini d'actions tel que la transition est commutative. Pour tout $p \in \Delta(K)$, $\Gamma^{sb}(p)$ admet une valeur uniforme et il existe une stratégie pure 0-optimale.*

Remarque 1.5.16 *Si l'espace d'actions est infini, on sait par Renault [Ren11] que la valeur uniforme existe mais on ne sait pas s'il existe une stratégie 0-optimale quand les transitions sont commutatives.*

Lorsqu'il y a deux joueurs, on montre l'existence de la valeur uniforme mais qu'il n'existe pas forcément de stratégies 0-optimales, en introduisant un jeu simulant le Big Match.

Théorème 1.5.17 *Soit $\Gamma^{sb} = (K, I, J, q, g)$ un jeu répété avec transitions commutatives où les joueurs n'observent pas l'état, avec K, I et J finis. Pour tout $p \in \Delta(K)$, $\Gamma^{sb}(p)$ a une valeur uniforme.*

Exemple 1.5.18 *Soit $\Gamma^{sb} = (K, I, J, q, g)$ défini par $K = \{\alpha, k_{T,R}, k_{T,L}, k_T\}$, $I = \{T, B\}$ et $J = \{L, R\}$ tel que la transition et la fonction de paiement sont données par*

$$\begin{array}{ccc} & \begin{pmatrix} 2_{\circlearrowleft} & 2_{\circlearrowleft} \\ 2_{\circlearrowleft} & 2_{\circlearrowleft} \end{pmatrix} & \\ & k_T & \\ \begin{pmatrix} 1_{\circlearrowleft} & 1_{\nearrow k_T} \\ 1_{\circlearrowleft} & 1_{\circlearrowleft} \end{pmatrix} & \begin{pmatrix} 1_{\leftarrow k_{T,L}} & 0_{\rightarrow k_{T,R}} \\ 0_{\circlearrowleft} & 1_{\circlearrowleft} \end{pmatrix} & \begin{pmatrix} 0_{\nwarrow k_T} & 0_{\circlearrowleft} \\ 0_{\circlearrowleft} & 0_{\circlearrowleft} \end{pmatrix} \\ k_{T,L} & \alpha & k_{T,R} \end{array}$$

Alors le jeu $\Gamma(\delta_\alpha)$ est stratégiquement équivalent au Big Match et le joueur 1 n'a pas de stratégie 0-optimale. Comme la transition est déterministe, l'observation des actions est suffisante pour connaitre l'état. La valeur en $k_{T,L}$ est 1, la valeur en $k_{T,R}$ est 0 et le jeu est équivalent au Big Match.

Les démonstrations de ces deux théorèmes se déroulent en deux temps. On définit le MDP (resp. le jeu stochastique auxiliaire) Ψ sur l'espace $\Delta(K)$ associé au jeu répété et on vérifie que l'existence de la valeur uniforme (resp. d'une stratégie 0-optimale) dans le jeu stochastique implique l'existence de la valeur uniforme (resp. d'une stratégie 0-optimale) dans le jeu répété. Ces problèmes auxiliaires ont un nombre fini d'actions, une transition déterministe et 1-Lipschitz pour la norme $\|.\|_1$. On démontre l'existence d'une stratégie 0-optimale pure dans le cas du MDP et l'existence de la valeur uniforme dans le cas du jeu stochastique puis on déduit le résultat dans le modèle initial.

b) Jeux stochastiques "compact"

Dans un premier temps, on considère le cas où il n'y a qu'un décideur et on montre que si le décideur peut garantir v alors il peut le garantir sans faire d'erreur.

Théorème 1.5.19 *Soit $\Gamma = (X, I, q, g)$ un MDP tel que q est déterministe et commutative, et I est fini.*

1. *Si pour tout $z \in \Delta_f(X)$, $\Gamma(z)$ a une valeur uniforme en stratégies pures alors pour tout $z \in \Delta_f(X)$, il existe une stratégie 0-optimale comportementale.*
2. *De plus si X est un espace précompact métrique, q est 1-Lipschitz et g est uniformément continue alors il existe une stratégie 0-optimale pure.*

Comme la transition est déterministe, si le décideur utilise une stratégie pure alors il existe une unique histoire avec une probabilité positive, appelée la trajectoire. La commutation implique l'existence pour tout $\varepsilon > 0$ de trajectoires ε-optimales telles que la valeur est constante. Le décideur ne fait pas d'erreur irréversible.

On démontre alors de deux manières différentes les deux parties du théorème. Sans hypothèse topologique, le décideur utilise des action mixtes pour concaténer des stratégies de plus en plus précises sans que le paiement ne chute. Ainsi le décideur commence par jouer une stratégie ε-optimale jusqu'à ce que le paiement soit effectivement bon. Puis avec une petite probabilité, il commence à suivre une stratégie $\frac{\varepsilon}{2}$-optimale à partir de l'état courant. Le paiement sur ces histoires peut être mauvais pendant longtemps mais après un certain nombre d'étapes il devient $\frac{\varepsilon}{2}$-optimal. A ce moment là, le décideur peut considérer les histoires qui n'ont pas encore changé et commencer à suivre une stratégie $\frac{\varepsilon}{2}$-optimale avec une petite probabilité. Il peut ainsi passer d'une stratégie ε-optimale à une stratégie $\frac{\varepsilon}{2}$-optimale. Afin d'obtenir une stratégie 0-optimale, il suffit de répéter l'opération.

Avec les hypothèses topologiques, la construction est différente. On définit une suite d'états initiaux x^l récursivement. Soit $x^1 = x_1$ puis pour chaque $l \geq 1$ on considère une stratégie σ^l, ϵ_l-optimale partant de x^l et on choisit x^{l+1} comme une valeur d'adhérence de la suite d'état visité. L'idée est de suivre la stratégie partant de x^l jusqu'à un état proche de x^{l+1} et ainsi de suite. Néanmoins à chaque changement le décideur fait une petite erreur. Afin de la compenser, on construit la stratégie optimale telle qu'elle suit chacune des stratégies σ^l en entier mais

blocs par blocs. En particulier à l'étape où la stratégie commence à suivre σ^l, l'écart entre l'état courant et x^l est de plus en plus petit.

Comme le MDP, associé au POMDP dans le noir, satisfait ces hypothèses on en déduit le corollaire 1.5.15. L'exemple suivant, détaillé dans le chapitre, montre la différence entre les deux résultats du théorème.

Exemple 1.5.20 *Considérons le MDP suivant. L'ensemble d'états est* $\mathbb{N} \times \mathbb{N}$ *et le décideur a deux actions* R *et* T*. L'action* R *incrémente la première coordonnée et l'action* T *incrémente la seconde coordonnée.*

$$q((x,y),R) = (x+1,y),$$
$$q((x,y),T) = (x,y+1).$$

Pour tout $l \in \mathbb{N}$*, on définit la trajectoire* h^l *par sa suite d'actions* $R^{w_l}(TR^{4^{l-1}-1})^\infty$ *avec* $w_l = \sum_{m=1}^{l}(4^{m-1}-1) = \frac{4^l-1}{3} - l$*. Elle est constituée de blocs de* R *entrecoupé de* T *et décrit un escalier avec des marches de hauteur* 1 *et de longueur* $4^{l-1}-1$*. Le paiement est fixé à* $1-\frac{1}{2^l}$ *sur tous les états visités par* h^l *et* 0 *sur les états n'appartenant à aucun* h^l*.*

La transition de ce MDP est déterministe, commutative et il existe une valeur uniforme égale à 1 *quelque soit l'état initial. Elle est de plus garantie par des stratégies pures. D'après le théorème, il existe donc une stratégie* 0*-optimale comportementale. Par contre il n'existe pas de stratégies* 0*-optimales pures. Considérons l'état initial* $(0,0)$*, une stratégie pure* 0*-optimale partant de* $(0,0)$ *doit croiser chaque trajectoire* h^l *et ainsi passer de l'une à l'autre. Or si la stratégie quitte* h^l *à l'étape* n *pour atteindre* h^{l+1} *alors pendant au moins* n *étapes le paiement est* 0 *et cette stratégie garantit au plus* $1/2$*.*

Dans le second théorème, on étudie une classe de jeux stochastiques sur $\mathbb{R}^m$ avec un nombre fini d'actions et une transition commutative et déterministe. On suppose que la transition est de plus 1-Lipschitz pour la norme $\|.\|_1$. L'hypothèse 1-Lipschitz est nécessaire pour espérer obtenir une valeur uniforme d'après les résultats de Renault [Ren11]. Ici on utilise explicitement la norme $\|.\|_1$ dont la boule unité a un nombre fini de points extrêmaux.

Théorème 1.5.21 *Soit* $\Gamma = (X, I, J, q, g)$ *un jeu stochastique tel que* X *est un sous-ensemble compact de* $\mathbb{R}^m$*,* I *et* J *sont des espaces finis,* q *est commutative, déterministe et* 1*-Lipschitz pour la norme* $\|.\|_1$ *et* g *est continue. Alors pour tout* $z \in \Delta_f(X)$*, le jeu stochastique* $\Gamma(z)$ *a une valeur uniforme.*

L'idée de la preuve est de classer les états en fonction de leur nombre de couples d'actions cycliques puis de raisonner par récurrence. Un couple d'actions est cyclique en un état x, si la trajectoire obtenue en itérant ce couple d'action à partir de x revient en x en un nombre fini d'étapes. La commutation implique qu'il existe au moins un état où tous les couples d'actions sont cycliques. L'ensemble des états visités à partir de cet état est fini et il existe une valeur uniforme en appliquant un résultat de Mertens et Neymann [MN81] sur les jeux stochastiques

finis. Fixons un état x_1 avec m couples d'actions cycliques. Afin de prouver l'hérédité de la propriété de récurrence, on définit pour chaque voisinage de l'ensemble des états avec $m+1$ couples d'actions cycliques, un jeu stochastique tel que ce voisinage est absorbant. Le paiement dans ces nouveaux états absorbants est donné par la valeur uniforme d'un état proche avec $m+1$ couples d'actions cycliques. On prouve que ces jeux stochastiques ont un nombre fini d'états et donc admettent une valeur uniforme par le théorème de Mertens et Neymann [MN81]. On en déduit alors que le jeu en x_1 a une valeur uniforme, ce qui achève la preuve.

On montre que la preuve peut être adaptée à des cas plus généraux en particulier à des jeux à somme non-nulle.

1.5.3 Jeux avec un contrôleur plus informé

Ce chapitre est extrait d'un article écrit en collaboration avec Fabien Gensbittel et Miquel Oliu Barton. Le but est de répondre à une remarque de Renault [Ren12b] qui pose la question de la généralisation de son résultat sur les jeux répétés avec un contrôleur informé aux cas où le contrôleur est mieux informé sur l'état que le second joueur mais pas parfaitement informé. On montre que l'on peut définir un jeu stochastique auxiliaire et que l'on peut utiliser certains résultats de Renault [Ren12b] pour prouver l'existence de la valeur uniforme. En plus de démontrer un résultat nouveau, on donne un modèle unifié pour les structures d'information étudiées dans Renault [Ren11] et dans Renault [Ren12b]. Rappelons que dans le premier article, Renault étudie les POMDPs avec un seul joueur partiellement informé alors que dans le second il étudie les jeux répétés avec un contrôleur parfaitement informé et un second joueur partiellement informé.

a) Quel espace d'états auxiliaire ?

Le premier problème est de trouver sur quel espace on peut exprimer la formule de récurrence et, si possible, définir un jeu stochastique. Renault suggère dans [Ren12b] d'introduire un jeu auxiliaire sur les couples de croyances des deux joueurs sur l'état. L'hypothèse d'un joueur 1 plus informé s'exprime alors par l'ordre de Choquet : étant donné deux probabilités μ et ν sur $X = \Delta(K)$, $\mu \leq \nu$ pour l'ordre de Choquet, si et seulement si, pour tout $f : X \to \mathbb{R}$ concave, $\mu(f) \leq \nu(f)$. L'espace proposé est donc

$$\{(\mu, \nu) \in \Delta_f(X) \times \Delta_f(X) \text{ t.q. } \mu \leq \nu\}.$$

On montre dans l'exemple qui suit que, même dans le cas d'une inclusion d'information, il ne suffit pas de connaitre les croyances d'ordre 1. Par contre l'espace des croyances du joueur 2 sur les croyances du joueur 1 est suffisant pour exprimer la formule de récurrence.

Exemple 1.5.22 *On considère $K = \{k_1, k_2\}$, deux signaux publics $U = \{u_1, u_2\}$ observés par le joueur 2 et le joueur 1 et trois signaux privés $S = \{s_1, s_2, s_3\}$ observés uniquement par le joueur 1 (avec les notations précédentes le joueur 1 recoit un signal dans $C = U \times S$ et le joueur*

2 *dans* $D = U$*). On va considérer deux probabilités initiales différentes et donc deux jeux* $\Gamma(\pi)$ *et* $\Gamma(\pi')$*.*

La probabilité initiale $\pi \in \Delta(K \times S \times U)$ *est définie par*

$$\begin{array}{ccc} & u_1 & u_2 \\ s_1 & \frac{8}{24} & 0 \\ s_2 & \frac{1}{24} & \frac{3}{24} \\ s_3 & 0 & 0 \end{array} \qquad \begin{array}{ccc} & u_1 & u_2 \\ s_1 & 0 & 0 \\ s_2 & \frac{1}{24} & \frac{3}{24} \\ s_3 & \frac{2}{24} & \frac{6}{24} \end{array},$$
$$k_1 \qquad\qquad k_2$$

et la probabilité initiale π' *par*

$$\begin{array}{ccc} & u_1 & u_2 \\ s_1 & \frac{6}{24} & \frac{2}{24} \\ s_2 & \frac{3}{24} & \frac{1}{24} \\ s_3 & 0 & 0 \end{array} \qquad \begin{array}{ccc} & u_1 & u_2 \\ s_1 & 0 & 0 \\ s_2 & \frac{3}{24} & \frac{1}{24} \\ s_3 & 0 & \frac{8}{24} \end{array}.$$
$$k_1 \qquad\qquad k_2$$

Le joueur 2 *observe la colonne tandis que le joueur* 1 *observe la case. La croyance du joueur* 2 *sur l'état est* $\frac{1}{2}\delta_{(\frac{3}{4},\frac{1}{4})} + \frac{1}{2}\delta_{(\frac{1}{4},\frac{3}{4})}$ *dans les deux cas. La croyance du joueur* 1 *sur l'état est* $\frac{1}{3}\delta_{(1,0)} + \frac{1}{3}\delta_{(\frac{1}{2},\frac{1}{2})} + \frac{1}{3}\delta_{(0,1)}$ *dans les deux cas.*

Pourtant dans le jeu répété suivant, à la Aumann et Maschler, v_1 *est différente selon que la probabilité initiale soit* π *ou* π'*. Soit* $\Gamma = (K, I, J, C, D, g)$ *le jeu où* $I = \{T, B\}$*,* $J = \{L, R\}$ *et* g *est donnée par*

$$\begin{array}{ccc} & L & R \\ T & 0 & 1 \\ B & 1 & 2 \end{array} \qquad \begin{array}{ccc} & L & R \\ T & 1 & \frac{1}{2} \\ B & 1 & 0 \end{array}$$
$$k_1 \qquad\qquad k_2$$

Il est optimal pour le joueur 1 *de jouer* T *s'il reçoit le signal* s_3 *(et sa croyance est* $(0, 1)$*) car* T *est une action dominante. S'il reçoit* s_1 *ou* s_2 *il est optimal de jouer l'action* B *qui est dominante. En particulier cela ne dépend pas de la stratégie du joueur* 2*. Maintenant, on cherche une meilleure réponse du joueur* 2 *à cette stratégie. On vérifie que dans les deux cas,* π *et* π'*, la meilleure réponse pour le joueur* 2 *est de jouer* L *s'il reçoit* u_1 *et de jouer* R *s'il reçoit* u_2*. Les deux joueurs suivent les mêmes stratégies, néanmoins* $v(\pi) = \frac{7}{8}$ *et* $v(\pi') = \frac{11}{12}$*.*

b) Modèle

Nous allons donner une définition formelle d'un jeu avec un contrôleur plus informé en 3 hypothèses sur les croyances des différents joueurs. Étant donné une variable aléatoire U sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et $\mathcal{F}$ une sous tribu, on note $\mathcal{L}_{\mathbb{P}}(U|\mathcal{F})$ la loi conditionnelle

de U sachant $\mathcal{F}$, vu comme une variable aléatoire $\mathcal{F}$-mesurable. De plus, on note $\mathcal{L}_{\mathbb{P}}$ la loi de X.

Définition 1.5.23 *Soit $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$ et un couple de stratégies (σ, τ), on définit pour tout $t \in \mathbb{N}^*$*

$$x_t := \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1) \in \Delta(K),$$

$$z_t = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}\left(\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t|h_t^1)\Big|h_t^2\right) \in \Delta_f(\Delta(K)),$$

et

$$\eta_t = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(z_t) \in \Delta_f(\Delta_f(\Delta(K))).$$

Si les joueurs connaîssent les stratégies alors x_t est la croyance du joueur 1 sur l'état, z_t est la croyance du joueur 2 sur x_t et η_t est la loi de z_t. On dit qu'une stratégie du joueur 1 est réduite si elle ne dépend à l'étape t que des variables auxiliaires x_t et z_t ; elle est "Markovienne" en (x, z).

Exemple 1.5.24 *Ainsi considérons la probabilité initiale π de l'exemple précédent :*

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ \\ k_1 \end{array} \begin{array}{c} \begin{array}{cc} u_1 & u_2 \end{array} \\ \begin{pmatrix} \frac{8}{24} & 0 \\ \frac{1}{24} & \frac{3}{24} \\ 0 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ \\ k_2 \end{array} \begin{array}{c} \begin{array}{cc} u_1 & u_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ \frac{1}{24} & \frac{3}{24} \\ \frac{2}{24} & \frac{6}{24} \end{pmatrix} \end{array}.$$

Si le joueur 1 recoit le signal (s_1, u_1) sa croyance sur l'état est $x_1 = (1, 0)$. S'il reçoit (s_2, u_1) ou (s_2, u_2), sa croyance est $(\frac{1}{2}, \frac{1}{2})$. Enfin s'il recoit (s_3, u_1) ou (s_3, u_2), sa croyance est $(0, 1)$.

A l'ordre supérieur si le joueur 2 recoit le signal u_1 sa croyance sur le signal du joueur 1 est $\frac{8}{24}\delta_{s_1} + \frac{2}{24}\delta_{s_2} + \frac{2}{24}\delta_{s_3}$, d'où

$$z_1(u_1) = \frac{8}{24}\delta_{(1,0)} + \frac{2}{24}\delta_{(1/2,1/2)} + \frac{2}{24}\delta_{(0,1)}.$$

S'il recoit le signal u_2, on obtient

$$z_1(u_2) = \frac{6}{24}\delta_{(1/2,1/2)} + \frac{6}{24}\delta_{(0,1)}.$$

Finalement on a donc en écrivant la loi de y_1 :

$$\eta_1 = \frac{1}{2}\delta_{\frac{8}{24}\delta_{(1,0)}+\frac{2}{24}\delta_{(1/2,1/2)}+\frac{2}{24}\delta_{(0,1)}} + \frac{1}{2}\delta_{\frac{6}{24}\delta_{(1/2,1/2)}+\frac{6}{24}\delta_{(0,1)}}.$$

Les hypothèses sont les suivantes.

Hypothèse $(A1)$: Le joueur 1 a une information plus précise sur l'état que le joueur 2. Elle est automatiquement vérifiée, par exemple, si le joueur 1 observe l'état ou bien si le joueur 1 a connaissance des actions et des signaux du joueur 2.

$$(A1)\ \forall t \in \mathbb{N}^*, \forall \sigma \in \Sigma, \forall \tau \in \mathcal{T}, \forall h_t^1,\ h_t^2 \qquad \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1, h_t^2) = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1).$$

Hypothèse $(A2)$: Le joueur 1 peut calculer la croyance de joueur 2 sur sa propre croyance sans connaître la stratégie du joueur 2. Cela implique en particulier qu'il peut calculer la croyance du joueur 2 sur l'état. Cette hypothèse se décompose en deux sous hypothèses.

(A2a) La probabilité initiale π est telle qu'il existe une application $f_\pi^1 : \mathbb{N} \to \Delta(\Delta(K))$ avec $z_1 = f_\pi^1(c')$ π-presque surement.

(A2b) Pour toute stratégie réduite σ_1 et pour tout π vérifiant (A2a), il existe une suite d'applications $(f^t_{\pi,\sigma_1})_{n\in\mathbb{N}}$ telle que pour tout $n \geq 1$,
$$f^t_{\pi,\sigma_1} : H_1^n \to \Delta(\Delta(K)),$$
et pour tout τ, $z_t = f^t_{\pi,\sigma_1}(h_1^t)$ $\mathbb{P}_{\pi,\sigma,\tau}$ -presque surement.

Cette hypothèse est plus forte que la mesurabilité de z_t par rapport à l'histoire du joueur 1 car elle suppose l'existence d'une fonction indépendante de la stratégie du joueur 2.

On note $\Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$ les probabilités initiales qui vérifient $(A1)$ et $(A2)$. Lorsque ces deux hypothèses sont vérifiées, on montre que le jeu stochastique auxiliaire peut être défini sur l'espace $Z = \Delta_f(\Delta(K))$ des croyances du joueur 2 sur les croyances du joueur 1. Cet espace permet donc d'exprimer les modèles suivants de la littérature : les POMDPs, les jeux avec un joueur parfaitement informé et les jeux répétés avec information symétrique. En fait, comme on l'a vu dans cette introduction, pour ces cas particuliers cet espace est trop grand et on peut exprimer la formule de récurrence sur des espaces plus petits. Notons que le joueur 2 peut avoir une information privée tant qu'elle ne concerne pas l'état, on est donc très proches des cas étudiés par Mertens [Mer87], où un joueur apprend toute l'information observée par l'autre joueur.

Hypothèse $(A3)$: On suppose que le joueur 2 n'a pas d'influence sur l'évolution de l'état dans ce jeu auxiliaire.

(A3) Pour tout $\pi \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$, si le joueur 1 suit une stratégie réduite alors η_2 ne dépend pas de la stratégie du joueur 2.

c) Résultat

Dans le cas où les trois hypothèses sont vérifiées, on obtient un jeu stochastique auxiliaire sur un espace d'états compact mais contrôlé par un joueur.

Théorème 1.5.25 *Soit $\Gamma = (K, I, J, C, D, q, g)$ un jeu répété qui vérifie $(A1)$,$(A2)$ et $(A3)$, i.e. avec un contrôleur plus informé, alors pour toute probabilité $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$ compatible avec $(A1)$ et $(A2)$, le jeu $\Gamma(\pi)$ admet une valeur uniforme.*

La preuve se décompose en trois parties.

D'abord, on montre que pour chaque évaluation les fonctions valeurs v_θ ne dépendent que de la projection de la probabilité initiale π, notée $\Phi(\pi)$, sur l'espace $\Delta_f(Z)$. D'autre part si on note $\widetilde{v}_\theta$, la fonction induite sur $\Delta_f(Z)$, elle est linéaire et 1-Lipschitz pour la distance de Kantorovitch-Rubinstein.

On définit ensuite un jeu stochastique auxiliaire Ψ sur l'espace Z. La définition est similaire à la définition du MDP auxiliaire associé à un POMDP, on considère la probabilité sur les histoires de longueur 2 comme une probabilité initiale et on la projette sur $\Delta_f(Z)$. Soit Ψ le jeu stochastique défini par

- l'espace d'état $Z = \Delta_f(\Delta(K))$,
- l'espace d'action du joueur 1, $A = \{f : \Delta(K) \to \Delta(I), \text{measurable}\}$,
- l'espace d'action du joueur 2, $B = \Delta(J)$,
- la fonction de paiement $\widetilde{g} : Z \times A \times B \to [0, 1]$ définie sur Z par

$$\widetilde{g}(z, a, b) = \sum_{p \in \text{supp}(z)} \sum_{(i,j) \in I \times J} b(j)a(p, i)g(p, i, j)z(p),$$

  où $\text{supp}(z)$ est le support de z.
- la fonction de transition $\widetilde{q} : Z \times A \times B \to \Delta_f(Z)$ définie par $\widetilde{q}(z, a, b) = \Phi(Q(z, a, b))$, où $Q(z, a, b) \in \Delta_f((K) \times (\Delta(K) \times C) \times (D))$ est la loi jointe induite sur $(k_2, (p, c_1), (d_1))$ dans le jeu réduit où p est tiré selon la loi z, puis le joueur 1 joue $a(p)$ et le joueur 2 joue b.
  Les ensembles C, D, K et $\text{supp}(z)$ sont finis, on peut donc considérer Q comme une loi à supports finis sur $K \times \mathbb{N} \times \mathbb{N}$ et sa projection par Φ sur $\Delta_f(Z)$.

La valeur de ce jeu sous l'évaluation θ est égale à la valeur réduite $\widetilde{v}$. Ce jeu ne vérifie pas les hypothèses de Renault [Ren12b] car un des ensembles d'actions n'est pas compact. Néanmoins cette hypothèse est nécessaire uniquement pour appliquer un théorème de Sion, or il existe des versions du théorème de Sion où un des espaces n'est pas compact. La preuve de Renault reste vraie avec ce changement et on obtient une version modifiée de son résultat sur les jeux stochastiques avec un espace d'états compact [Ren12b] qui s'applique à Ψ. Le jeu stochastique auxiliaire Ψ admet une valeur uniforme et elle est caractérisée par des fonctions auxiliaires.

Les deux joueurs peuvent garantir cette valeur dans le jeu répété. Le joueur 1 peut copier sa stratégie optimale du jeu stochastique auxiliaire dans le jeu répété. Le joueur 2 ne peut pas calculer l'état du jeu stochastique auxiliaire sans connaître la stratégie du joueur 1. Il ne peut

donc pas copier une stratégie optimale du jeu auxiliaire dans le jeu répété. Par contre comme il n'influence pas la transition, il peut garantir la valeur en jouant par blocs dans le jeu répété.

1.5.4 Stratégies à paiement constant

On étudie dans le dernier chapitre la relation entre la convergence uniforme des valeurs des jeux avec n étapes et le comportement asymptotique des stratégies. Ce sujet a fait l'objet d'un article rédigé en collaboration avec Sylvain Sorin et Guillaume Vigeral. On s'intéresse à une notion plus forte que l'approche asymptotique qui s'intéresse aux comportements asymptotiques des valeurs v_n et v_λ sans se préoccuper des stratégies et plus faible que l'approche uniforme des chapitres précédents qui étudie l'existence d'une stratégie bonne pour tout jeu suffisament long. La convergence uniforme des fonctions v_n n'implique pas l'existence de la valeur uniforme. Par contre, si on considère un MDP, elle implique l'existence d'une suite de stratégies σ_n telle que pour tout $n \geq 1$, σ_n est ϵ-optimale dans le jeu de longeur n et le paiement sur une fraction des n étapes converge aussi vers la limite, lorsque n tend vers l'infini. Sans perte de généralité, on se place dans le cadre de la programmation dynamique, quitte à formuler le problème sur l'espace des probabilités. Dans le problème $\Gamma = (Z, F, r)$, où F est une correspondance de Z dans Z, à chaque étape le décideur choisit un nouvel état z_{t+1} dans $F(z_t)$ et reçoit le paiement $r(z_{t+1})$. Étant donné un état z, une suite admissible est une suite d'états telle que $z_1 \in F(z)$ et pour tout $t \geq 1$, z_{t+1} est dans $F(z_t)$.

Définition 1.5.26 *Le jeu vérifie la propriété* $\mathbf{P}$ *s'il existe* $w : Z \to \mathbb{R}$ *telle que : pour tout* $\varepsilon > 0$*, il existe* n_0*, tel que pour tout* $n \geq n_0$*, pour tout état* z *et pour toute suite admissible d'états* $(z_t)_{t\in\mathbb{N}^*}$ *en* z*,* ε*-optimale dans* $\Gamma_n(z)$ *et pour tout* $l \in [0, 1]$ *:*

$$-3\varepsilon \leq \frac{1}{n}\left(\sum_{t=1}^{[l\,n]} r(z_t) - l\, w(z)\right) \leq 3\varepsilon. \tag{1.3}$$

où $[l\,n]$ *est la partie entière de* $l\,n$*.*

Alors le jeu satisfait $\mathbf{P}$, si et seulement si, la convergence est uniforme.

Théorème 1.5.27 *Si les valeurs des jeux avec* n *étapes,* v_n *convergent uniformément sur l'espace d'états alors le jeu vérifie* $\mathbf{P}$ *et* $w = \lim v_n$*.*

Ainsi la convergence uniforme des valeurs des jeux avec n étapes implique l'existence de stratégies où le paiement devient constant le long de la trajectoire. Le même résultat reste vrai si l'on considère les fonctions v_λ.

Définition 1.5.28 *Le jeu vérifie la propriété* $\mathbf{P}'$ *s'il existe* $w : Z \to \mathbb{R}$ *telle que : pour tout* $\varepsilon > 0$*, il existe* λ_0*, tel que pour tout* $\lambda \leq \lambda_0$*, pour tout état* z *et pour toute suite admissible*

d'état $\{z_t\}$ *ε-optimale pour* $G_\lambda(z)$ *et pour tout* $t \in [0,1]$ *:*

$$-3\varepsilon \leq \sum_{t=1}^{n(l;\lambda)} \lambda(1-\lambda)^{t-1} r_t - l\, w(z) \leq 3\varepsilon. \tag{1.4}$$

où $n(l;\lambda) = \inf\{p \in ; \sum_{t=1}^{p} \lambda(1-\lambda)^{t-1} \geq l\}$*. L'étape* $n(l;\lambda)$ *correspond à la fraction* l *de la durée totale du problème* G_λ*.*

Théorème 1.5.29 *Si les valeurs des jeux avec* n *étapes,* v_n *convergent uniformément sur l'espace d'états alors le jeu vérifie* **P'** *et* $w = \lim v_\lambda$*.*

Si la suite converge uniformément alors il existe une suite des stratégies telle que le paiement, sur une suite consécutive d'étapes représentant une fraction l du poids total du jeu, converge vers $l\, v$, où v est la valeur limite.

Lorsque l'on considère un jeu stochastique à deux joueurs, on montre que le résultat n'est pas vrai si on ne fait pas des hypothèses sur les stratégies jouées par les deux joueurs.

On doit considérer des trajectoires où les deux joueurs jouent optimal dans le jeu de longueur n. Dans le Big Match, si le joueur 1 joue de manière optimal et le joueur 2 joue l'action L, qui peut entrainer l'absorption dans l'état de paiement 1 mais donne un paiement courant de 0 sinon, alors le paiement pendant la première moitié du jeu est en moyenne 1/4 et non 1/2.

Mais même si les deux joueurs jouent de manière optimale, le résultat reste faux. On montre un exemple où v_n converge uniformément vers une limite v tel qu'il existe un état initial k_1, pour tout $n \in \mathbb{N}$ un couple (σ_n, τ_n) de stratégies 0-optimales dans le jeu de longueur n et un réel $l \in [0,1]$ tel que

$$\frac{1}{n} I\!E_{k_1,\sigma_n,\tau_n} \left(\sum_{t=1}^{[l\,n]} r(k_t, i_t, j_t) \right)$$

ne converge pas vers $lv(k_1)$. Ainsi les paiements moyens ne convergent pas. L'exemple est construit autour des deux idées suivantes : pour chaque $n \geq 1$, il existe au moins un couple de stratégies qui garantit la valeur dès l'étape 2, ce qui assure de la convergence uniforme, et il existe un couple de stratégies qui garantit la valeur de manière non uniforme : un bon paiement pendant $n/2$ étapes puis un paiement de -1 pour le reste du jeu. Le long de ce couple de stratégies, le paiement moyen ne converge pas.

Chapitre 2

Existence of long-term values in MDPs and Repeated Games

Résumé : Soit K un ensemble fini, on note $X = \Delta(K)$ l'ensemble des probabilités sur K et $Z = \Delta_f(X)$ l'ensemble des probabilités Boréliennes sur X à support fini. Afin d'étudier un processus de décision Markovien partiellement observable (POMDP) sur K, on introduit un processus de décision Markovien sur X. On définit une nouvelle distance d_* sur Z telle que les transitions de ce MDP soient 1-Lipschitz de $(X, \|.\|_1)$ to (Z, d_*). La première partie du chapitre est consacrée à la définition de d_* et aux démonstrations de certaines de ses propriétés. En particulier d_* satisfait une équation de dualité comme la distance de Kantorovitch-Rubinstein et peut être caractérisée par les désintégrations. Dans la seconde partie, on caractérise la valeur limite dans plusieurs problèmes de MDP où l'espace d'état est compact et la transition "non expansive". De plus on prouve qu'il existe des notions de valeur plus fortes que d'habitude où on ne se limite plus à considérer les limites de Cesàro mais toutes les évaluations $\theta \in \Delta(\mathbb{N}^*)$ lorsque leur irrégularité $I(\theta) = \sum_{t\geq 1} |\theta_{t+1} - \theta_t|$ est suffisamment petite. À l'aide de la distance d_*, on vérifie que ces résultats s'appliquent aux POMDPs et aux jeux répétés avec un contrôleur informé.

Ce chapitre est extrait d'un article écrit en collaboration avec Jérôme Renault.

Abstract: Given a finite set K, we denote by $X = \Delta(K)$ the set of probabilities on K and by $Z = \Delta_f(X)$ the set of Borel probabilities on X with finite support. Studying a Partial Observation MDP on K leads to a MDP with full information on X. We introduce a new metric d_* on Z such that the transitions of this MDP become 1-Lipschitz from $(X, \|.\|_1)$ to (Z, d_*). In the first part of the article, we define and prove several properties of the metric d_*. Especially, d_* satisfies a Kantorovich-Rubinstein type duality formula and can be characterized by using disintegrations. In the second part, we characterize the limit values in several classes of "compact non expansive" Markov Decision Processes. In particular, we use the metric d_* to characterize the limit value in Partial Observation MDPs with finitely many states and in Repeated Games with an informed controller with finite sets of states and actions. In each case we prove the existence of a generalized notion of long-term value where we consider not only the Cesàro mean when the number of stages is large enough, but any evaluation function $\theta \in \Delta(\mathbb{N}^*)$ when the irregularity $I(\theta) = \sum_{t\geq 1} |\theta_{t+1} - \theta_t|$ is small enough.

This chapter is extracted from an article in collaboration with Jérôme Renault.

2.1 Introduction

The classic model of Markov Decision Processes with finitely many states, particular class of the model of Stochastic Games introduced by Shapley [Sha53], was explicitly introduced by Bellman [Bel57] in the 1950s and has been extensively studied since then. When the set of actions is also finite, Blackwell [Bla62] proved the existence of a strategy which is optimal for all discount factors close to 0.This model was generalized later to MDPs with Partial Observations (POMDP), (for references see Arapostathis *et al.* [ABFG$^+$93]). The decision maker observes neither the state nor his payoff. Instead at each stage, he receives a signal which depends on the previous state and his previous action. In order to solve this problem a classic approach is to go back to the classic model of MDPs by introducing an auxiliary problem with full observation and Borel state space: the space of belief on the state as shown in Astrom, K.J. [Ast65], Sawaragi and Yoshikawa [SY70] and Rhenius [Rhe74]. For optimality criteria like the Cesàro mean and the Abel mean, these two problems are equivalent and the question of the existence of the limit value is the same. Then given some sufficient conditions of ergodicity, one can search for a solution of the Average Cost Optimality Criterion in order to find "the" value of the MDP, for example as in Borkar [Bor00] [Bor07]. An introduction to the ACOE in the framework of MDP and the reduction of POMDP can be found in Hernández-Lerma [HL89]. From another point of view, if we know that the limit value exists, the ACOE may be used as a characterization of the value. For finite MDP, for example, Denardo and Fox [DF68] proved that the limit value is the solution of a linear programming problem deduced from the ACOE. Moreover by standard linear programming results, it is also equal to the solution of a dual problem from which Hordjik and Kallenberg [HK79] deduced an optimal strategy. This dual problem focuses on the maximal payoff that the decision maker can guarantee on invariant measures. This approach was extended to different criteria (see Kallenberg [Kal94]) and to a convex analytic approach by Borkar (for references see Borkar [Bor02]) in order to study problems with a countable state space and a compact action space.

Given an initial POMDP on a finite space K, we will follow the usual approach and introduce a MDP on $X = \Delta(K)$ but instead of assuming some ergodicity on the process we will use the structure of $\Delta(K)$ and a new metric on $Z = \Delta_f(\Delta(K))$. We extend and relax the MDP on $\Delta_f(X)$ with a uniformly continuous affine payoff function and non-expansive affine transitions. The structure of Z was already used in Rosenberg, *et al.* [RSV02] and in Renault [Ren11]. Under our new metric, we highlight a stronger property since the transitions became 1-Lipschitz on Z and Z is still precompact. We use this property to focus on general evaluations. Given a probability distribution θ on positive integers, we evaluate a sequence of payoffs $g = (g_t)_{t\geq 1}$ by $\gamma_\theta(g) = \sum_t \theta_t g_t$. In a MDP or a POMDP, the θ-value is then defined as the maximum expected payoff that the player can guarantee with this evaluation. Most of the literature focuses on the n-stage game where we consider the Cesàro mean of length n, and on the λ discounted games, where we consider the Abel mean with parameter λ. The first type of results focuses on the limit when n converges to $+\infty$ and when λ converges to 0 or the relation between them. When there

is no player, the relation between them is directly linked to a Hardy-Littlewood theorem (see Filar and Sznajder, [SF92]). One of the limit exists if and only if the other exists and whenever they exist they are equal. Lehrer and Sorin [LS92] proved that this result extends to the case where there is one player provided we ask for uniform convergence. The other approach focuses on the existence of a good strategy in any long game or for any discount factor close to 0. We say that the MDP has a uniform value. For MDP with finitely many states, Blackwell's result [Bla62] solved both problems. In POMDPs, Rosenberg, *et al.* [RSV02] proved the existence of the uniform value when the sets of states, actions and signals are finite, and Renault [Ren11] removed the finiteness assumption on signals and actions.

Concerning stochastic games, Mertens and Neyman [MN81] proved the existence of the uniform value when the set of states and the set of actions are finite. The model also generalizes to partial information but the existence of possible private information implies a more complex structure on the auxiliary state space. Mertens and Zamir [MZ85] and Mertens [Mer87] introduced the universal belief space which synthesizes all the information for both players in a general repeated game: their beliefs about the state, their beliefs about the beliefs of the other player, etc... So far, the results always concern some subclasses of games where we can explicitly write the auxiliary game in a "small" tractable set. A lot of work has been done on games with one fully informed player and one player with partial information, introduced by Aumann and Maschler (see reference from [AMS95]). A state is chosen at stage 0 and remains fixed for the rest of the game. Renault [Ren06] extended the analysis to a general underlying Markov chain on the state space (see also Neyman, [Ney08]). Rosenberg *et al.* [RSV04] and Renault [Ren12b] proved the existence of the uniform value when the informed player can additionally control the evolution of the state variable.

The first section is dedicated to the description of the (pseudo)-distance d_* on $\Delta(X)$ in the general framework when X is a compact subset of a normed vector space. We provide different definitions and show that they all define this pseudo-distance. Then we focus on the case where X is a simplex. We prove that d_* is a real metric and prove a "Kantorovich-Rubinstein like " duality formula for probabilities with finite support on X. We give new definitions and a characterization by the disintegration mapping. The second section focuses on Gambling Houses and standard Markov Decision Processes. We first introduce the definitions of general limit value and general uniform value. Then we give sufficient conditions for the existence of the general uniform value and a characterization in several "compact" cases of Gambling Houses and Markov Decision Processes, including the finite state case. We study the limit value as a linear function of the initial probability so there are similarities with the convex analytic approach, but we are able to avoid any assumption on the set of actions. Moreover the MDPs that we are considering may not have 0-optimal strategies as shown in Renault [Ren11]. Finally we apply these results to prove the existence of the general uniform value in finite state POMDPs and repeated games with an informed controller.

2.2 A distance for belief spaces

2.2.1 A pseudo-distance for probabilities on a compact subset of a normed vector space

We fix a compact subset X of a real normed vector space V. We denote by $E = \mathcal{C}(X)$ the set of continuous functions from X to the reals, and by E_1 the set of 1-Lipschitz functions in E. We denote by $\Delta(X)$ the set of Borel probability measures on X, and for each x in X we write δ_x for the Dirac probability measure on x. It is well known that $\Delta(X)$ is a compact set for the weak-* topology, and this topology can be metrizable by the (Wasserstein) Kantorovich-Rubinstein distance:

$$\forall u, v \in \Delta(X), \;\; d_{KR}(u,v) = \sup_{f \in E_1} \; u(f) - v(f).$$

We will introduce a pseudo-distance on $\Delta(X)$, which is not greater than d_{KR} and in some cases also metrizes the weak-* topology. We start with several definitions, which will turn out to be equivalent. Let u and v be in $\Delta(X)$.

Definition 2.2.1

$$d_1(u,v) = \sup_{f \in D_1} \; u(f) - v(f),$$

$$\textit{where} \;\; D_1 = \{f \in E, \forall x, y \in X, \forall a, b \geq 0, \; af(x) - bf(y) \leq \|ax - by\|\}.$$

Note that any linear functional in V' with norm 1 induces an element of D_1. d_1 is a pseudo-distance on $\Delta(X)$, and $d_1(u,v) = \sup_{f \in D_1} \; |u(f) - v(f)|$, since if f is in D_1, $-f$ is also in D_1. We also have $D_1 \subset E_1$, so that $d_1(u,v) \leq d_{KR}(u,v)$ and the supremum in the definition of $d_1(u,v)$ is achieved.

Given x and y in X, there exists a linear functional f in V' with norm 1 such that $f(y-x) = \|y-x\|$. Then the restriction of f to X is in D_1 and $d_1(\delta_x, \delta_y) \geq \|x-y\|$. One can easily deduce that $d_1(\delta_x, \delta_y) = \|x-y\|$ for x and y in X.

Example 2.2.2 Consider the particular case where $X = [0,1]$ endowed with the usual norm. Then all f in D_1 are linear. As a consequence, $d_1(u,v) = 0$ for $u = 1/2\,\delta_0 + 1/2\,\delta_1$ and $v = \delta_{1/2}$. We do not have the separation property and d_1 is not a distance in this case.

Let us modify the example. X now is the set of probability distributions over 2 elements, viewed as $X = \{(x, 1-x), x \in [0,1]\}$. We use the norm $\|.\|_1$ to measure the distance between $(x, 1-x)$ and $(y, 1-y)$, so that $V = \mathbb{R}^2$ is endowed with $\|(x_1, x_2) - (y_1, y_2)\| = |x_1 - y_1| + |x_2 - y_2|$. Consider f in E such that $f((x, 1-x)) = x(1-x)$ for all x. f now belongs to D_1, and $d_1(u,v) \geq 1/4 > 0$ for $u = 1/2\,\delta_0 + 1/2\,\delta_1$ and $v = \delta_{1/2}$. One can show that $(\Delta(X), d_1)$ is a compact metric space in this case (see proposition 2.2.15 later), and for applications in this chapter d_1 will be a particularly useful distance whenever X is a simplex $\Delta(K)$ endowed with $\|x-y\| = \sum_{k \in K} |x^k - y^k|$.

Furthermore it is known that the Kantorovitch Rubinstein metric on $\Delta(X)$ only depends on the restriction of the norm $\|.\|$ on the set X. Especially if for all $x, x' \in X$ such that $x \neq x'$, $\|x - x'\| = 2$, then for all $u, v \in \Delta(X)$, $d_{KR}(u, v) = \|u - v\|_1$. This is not the case when considering the metric d_1. Two norms on V giving the same metric on X may leads to different pseudo-metrics on $\Delta(X)$. We consider in the next example different norms on the Euclidean space $\mathbb{R}^K$.

Example 2.2.3 We consider $V = \mathbb{R}^K$, $X = \{e_1, .., e_K\}$ the set of canonical vectors of V and a norm such that for all $k \neq k'$, $\|e_k - e_{k'}\| = 2$. We know that d_1 is smaller than the Kantorovitch-Rubinstein metric, so for all $u \in \Delta(X)$ and $v \in \Delta(X)$, we have $d_1(u, v) \leq \|u - v\|_1$.

We first consider the particular case of the norm defined by $\|x - y\| = 2^{1-\frac{1}{p}}\|x - y\|_p$ where $\|x - y\|_p = \left(\sum_{k=1}^K |x_k - y_k|^p\right)^{1/p}$ is the usual L^p-norm on $\mathbb{R}^K$, with p a fixed positive integer. Given $u, v \in \Delta(X)$, the function f defined by

$$\forall k \in K \;\; f(k) = \begin{cases} 1 & \text{if } u(k) \geq v(k) \\ -1 & \text{otherwise,} \end{cases}$$

satisfies $u(f) - v(f) = \sum_{k \in K} |u(k) - v(k)| = \|u - v\|_1$. Moreover for all $a \geq 0$, $b \geq 0$ and $k, k' \in K$ such that $k \neq k'$, we have

$$af(k) - bf(k') \leq a + b \leq \frac{2}{2^{1/p}} \left(a^p + b^p\right)^{1/p} = \|ae_k - be_{k'}\|,$$

and $af(k) - bf(k) \leq |a - b| \leq |a - b|\frac{2}{2^{1/p}}$. Therefore f is in D_1 and $d_1(u, v) = \|u - v\|_1$, independently [1] of p.

Nevertheless the inequality $d_1(u, v) \leq \|u - v\|_1$ may be strict as in the following example. We consider the case $K = 3$ and given a vector $(x_1, x_2, x_3) \in \mathbb{R}^3$, we define the norm $\|(x_1, x_2, x_3)\| = \max(|x_1| + |x_2|, 2|x_3|)$, which satisfies $\|e_1 - e_2\| = \|e_2 - e_3\| = \|e_3 - e_2\| = 2$. Let f be a function in D_1, then we have among others the following constraints:

$$\begin{array}{rlll} & \forall a, b \geq 0 & af(e_3) - bf(e_1) & \leq \|(-b, 0, a)\| = \max(2a, b) \\ \text{and} & \forall a \geq 0 & af(e_2) & \leq \|(0, a, 0)\| = a. \end{array}$$

Let $u = (0, 1/2, 1/2)$, $v = (1, 0, 0)$ and $f \in D_1$, then

$$u(f) - v(f) = \frac{1}{2}f(e_2) + \frac{1}{2}f(e_3) - f(e_1) \leq \frac{1}{2} + \max(2/2, 1) = \frac{3}{2}.$$

By symmetry of D_1, we deduce that $d_1(u, v) \leq \frac{3}{2} < \|u - v\|_1$. In fact one can show that $d_1(u, v) = \frac{3}{2}$ by checking that the function defined by $f(e_1) = 0$, $f(e_2) = 1$ and $f(e_3) = 2$ is in D_1 and satisfies $u(f) - v(f) = \frac{3}{2}$.

1. Similarly, the same result holds for the case $p = +\infty$, i.e. where $\|x - y\| = 2\|x - y\|_\infty$.

We now give other expressions for the pseudo-distance d_1.

Definition 2.2.4

$$d_2(u,v) = \sup_{(f,g)\in D_2} u(f) + v(g),$$

$$\text{where } D_2 = \{(f,g) \in E \times E, \forall x,y \in X, \forall a,b \geq 0,\ af(x) + bg(y) \leq \|ax - by\|\}.$$

Definition 2.2.5

$$d_2^+(u,v) = \inf_{\varepsilon>0} d_2^\varepsilon(u,v), \text{ where } d_2^\varepsilon(u,v) = \sup_{(f,g)\in D_2^\varepsilon} u(f) + v(g)$$

$$\text{and } \forall \varepsilon > 0,\ \ D_2^\varepsilon = \{(f,g) \in E \times E, \forall x,y \in X, \forall a,b \in [0,1],\ af(x) + bg(y) \leq \varepsilon + \|ax - by\|\}.$$

Definition 2.2.6

$$d_3(u,v) = \inf_{\gamma\in\mathcal{M}_3(u,v)} \int_{X^2\times[0,1]^2} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu),$$

where $\mathcal{M}_3(u,v)$ is the set of finite positive measures on $X^2 \times [0,1]^2$ satisfying for each f in E:

$$\int_{(x,y,\lambda,\mu)\in X^2\times[0,1]^2} \lambda f(x) d\gamma(x,y,\lambda,\mu) = u(f),\ \text{and}\ \int_{(x,y,\lambda,\mu)\in X^2\times[0,1]^2} \mu f(y) d\gamma(x,y,\lambda,\mu) = v(f).$$

In the next subsection we will prove the following result.

Theorem 2.2.7 *For all u and v in $\Delta(X)$, $d_1(u,v) = d_2(u,v) = d_2^+(u,v) = d_3(u,v)$.*

2.2.2 A second expression of the metric

The proof is split into several parts.

Proposition 2.2.8 $d_1 = d_2 = d_2^+$.

It is plain that $d_1 \leq d_2 \leq d_2^+$, so all we have to prove is $d_2^+ \leq d_1$. We start with a lemma.

Lemma 2.2.9 *Fix $\varepsilon > 0$, and let f in E be such that: $\forall x \in X$, $\forall a \in [0,1]$, $af(x) \leq \varepsilon + a\|x\|$. Define $\hat{f}$ by:*

$$\forall y \in X,\ \hat{f}(y) = \inf_{a\in[0,1], b\in(0,1], x\in X} \frac{1}{b}\left(\varepsilon + \|ax - by\| - af(x)\right).$$

Then for each y in X, $-\|y\| \leq \hat{f}(y) \leq -f(y) + \varepsilon$. Moreover $\hat{f} \in E_1$, and:

$$\forall x \in X, \forall y \in X, \forall a \in [0,1], \forall b \in [0,1],\ a\hat{f}(x) - b\hat{f}(y) \leq a\varepsilon + \|by - ax\|.$$

Proof of lemma 2.2.9: By assumption on f, we have for all y in X, a in $[0,1]$, b in $(0,1]$, x in X: $\frac{1}{b}(\varepsilon + \|ax - by\| - af(x)) \geq \frac{1}{b}(-a\|x\| + \|ax - by\|) \geq -\|y\|$. In the definition of $\hat{f}(y)$, considering $a = b = 1$ and $x = y$ yields $\hat{f}(y) \leq -f(y) + \varepsilon$.

Fix x and y in X, a and b in $[0,1]$. We have:

$$\begin{aligned} a\hat{f}(x) - b\hat{f}(y) &= a \inf_{a',b',x'} \frac{1}{b'}(\varepsilon + \|a'x' - b'x\| - a'f(x')) \\ &\quad -b \inf_{a'',b'',x''} \frac{1}{b''}(\varepsilon + \|a''x'' - b''y\| - a''f(x'')). \end{aligned}$$

If $a = 0$, then the inequality $\hat{f}(y) \geq -\|y\|$ leads to $-b\hat{f}(y) \leq b\|y\|$. If $b = 0$, choose $a' = 0$, $b' = 1$ and $x' = x$ to get $a\hat{f}(x) \leq a\varepsilon + \|ax\|$.

If $ab > 0$, given $\eta > 0$, choose a'', b'', x'' η-optimal in the second infimum. We can define $x' = x''$, and choose $a' \in [0,1]$ and $b' \in (0,1]$ such that $\frac{a'}{b'} = \frac{b}{a}\frac{a''}{b''}$. We obtain:

$$\begin{aligned} a\hat{f}(x) - b\hat{f}(y) &\leq b\eta + (\frac{a}{b'} - \frac{b}{b''})\varepsilon + (\|\frac{a''}{b''}bx'' - ax\| - \|\frac{a''}{b''}bx'' - by\|) \\ &\leq b\eta + (\frac{a}{b'} - \frac{b}{b''})\varepsilon + \|ax - by\|. \end{aligned}$$

If $a = b > 0$, choose $a' = a''$ and $b' = b''$ to obtain: $\hat{f}(x) - \hat{f}(y) \leq \|x - y\|$ and therefore $\hat{f}$ is 1-Lipschitz.

Otherwise, we distinguish two cases. If $\frac{a}{b}b'' \leq 1$, we define $b' = \frac{a}{b}b''$ and $a' = a''$ and we get $a\hat{f}(x) - b\hat{f}(y) \leq b\eta + \|ax - by\|$. If $\frac{a}{b}b'' > 1$, we define $b' = 1$ and $a' = \frac{a''b}{b''a} \in [0,1]$ and obtain $a\hat{f}(x) - b\hat{f}(y) \leq b\eta + a\varepsilon + \|ax - by\|$. Thus for all $\eta > 0$, we have

$$a\hat{f}(x) - b\hat{f}(y) \leq b\eta + a\varepsilon + \|ax - by\|,$$

and therefore $a\hat{f}(x) - b\hat{f}(y) \leq a\varepsilon + \|ax - by\|$. □

Proof of proposition 2.2.8: Fix u and v in $\Delta(X)$, and consider $\varepsilon > 0$. For each (f,g) in D_2^ε, we have $-f + \varepsilon \geq \hat{f} \geq g$ and $(f, \hat{f})$ in D_2^ε. We also have $(\hat{f}, f) \in D_2^\varepsilon$ so iterating the construction, we get $(\hat{f}, \hat{\hat{f}}) \in D_2^\varepsilon$, and $-\hat{f} + \varepsilon \geq \hat{\hat{f}} \geq f$.

Now, $u(f) + v(g) \leq u(\hat{\hat{f}}) + v(\hat{f}) \leq -u(\hat{f}) + \varepsilon + v(\hat{f})$. Hence we have obtained:

$$d_2^\varepsilon(u,v) \leq \varepsilon + \sup_{f \in C_{\varepsilon(u,v)}} -u(f) + v(f),$$

where $C_{\varepsilon(u,v)}$ is the set of functions f in E_1 satisfying:

$$\forall x \in X, \forall y \in X, \forall a \in [0,1], \forall b \in [0,1],\ af(x) - bf(y) \leq a\varepsilon + \|ax - by\| \text{ and } f(y) \geq -\|y\|.$$

For each positive k, one can choose f_k in E_1 achieving the above supremum for $\varepsilon = 1/k$. Taking a limit point of $(f_k)_k$ yields a function f in D_1 such that: $-u(f)+v(f) \geq d_2^+(u,v)$. The function $f^* = -f$ is in D_1 and satisfies $u(f^*) - v(f^*) \geq d_2^+(u,v)$, and the proof of proposition 2.2.8 is complete. □

Proposition 2.2.10 $d_2^+ \geq d_3$.

Proof: The proof is based on (a corollary of) Hahn-Banach theorem. Define: $H = \mathcal{C}(X^2 \times [0,1]^2)$ and

$$L = \{\varphi \in H, \exists f, g \in \mathcal{C}(X) \; s.t. \; \forall x, y \in X, \forall \lambda, \mu \in [0,1], \varphi(x,y,\lambda,\mu) = \lambda f(x) + \mu g(y)\}.$$

H is endowed with the uniform norm and L is a linear subspace of H. Note that the unique constant mapping in L is 0. Fix u and v in $\Delta(X)$, and let r be the linear form on L defined by $r(\varphi) = u(f) + v(g)$, where $\varphi(x,y,\lambda,\mu) = \lambda f(x) + \mu g(y)$ for all x, y, λ, μ.

Fix now $\varepsilon > 0$, and put:

$$U_\varepsilon = \{\varphi \in H, \forall x, y \in X, \forall \lambda, \mu \in [0,1], \varphi(x,y,\lambda,\mu) \leq \|\lambda x - \mu y\| + \varepsilon\}.$$

We have:

$$\sup_{\varphi \in L \cap U_\varepsilon} r(\varphi) = d_2^\varepsilon(u,v).$$

U_ε is a convex subset of H which is radial at 0, in the sense that: $\forall \varphi \in H$, $\exists \delta > 0$ such that $t\varphi \in U_\varepsilon$ as soon as $|t| \leq \delta$. By a corollary of Hahn-Banach theorem (see theorem 6.2.11 p.202 in Dudley, 2002), r can be extended to a linear form on H such that:

$$\sup_{\varphi \in U_\varepsilon} r(\varphi) = d_2^\varepsilon(u,v).$$

Given $\varphi \in H$, we have $\varepsilon\varphi/\|\varphi\|_\infty \in U_\varepsilon$, which implies that $r(\varphi) \leq \|\varphi\|_\infty d_2^\varepsilon(u,v)/\varepsilon$, so that r belongs to H'. And if $\varphi \geq 0$, we have $t\varphi \in U_\varepsilon$ if $t \leq 0$, so that $r(\varphi) \geq d_2^\varepsilon(u,v)/t$ for all $t \leq 0$ and $r(\varphi) \geq 0$. By Riesz Theorem, r can be represented by a positive finite measure γ on $X^2 \times [0,1]^2$.

Given f in E, one can consider $\varphi_f \in L$ defined by $\varphi_f(x,y,\lambda,\mu) = \lambda f(x)$. $r(\varphi = f) = \gamma(\varphi_f)$ gives:

$$u(f) = \int_{(x,y,\lambda,\mu) \in X^2 \times [0,1]^2} \lambda f(x) d\gamma(x,y,\lambda,\mu),$$

and similarly

$$v(f) = \int_{(x,y,\lambda,\mu) \in X^2 \times [0,1]^2} \mu f(y) d\gamma(x,y,\lambda,\mu),$$

and we obtain that $\gamma \in \mathcal{M}_3(u,v)$.

Because $\gamma \geq 0$, $\sup_{\varphi \in U_\varepsilon} r(\varphi) = r(\varphi^*)$ where $\varphi^*(x,y,\lambda,\mu) = \|\lambda x - \mu y\| + \varepsilon$. We get $d_2^\varepsilon(u,v) = \int_{X^2\times[0,1]^2} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) + \varepsilon\gamma(X^2 \times [0,1]^2)$, so

$$d_2^\varepsilon(u,v) \geq \int_{X^2\times[0,1]^2} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) \geq d_3(u,v).$$

□

Lemma 2.2.11 $d_3 \geq d_2$.

Proof: Fix $(f,g) \in D_2$ and $\gamma \in \mathcal{M}_3(u,v)$.

$$\begin{aligned} u(f) + v(g) &= \int_{X^2\times[0,1]^2} \lambda f(x) d\gamma(x,y,\lambda,\mu) + \int_{X^2\times[0,1]^2} \mu g(y) d\gamma(x,y,\lambda,\mu) \\ &= \int_{X^2\times[0,1]^2} (\lambda f(x) + \mu g(y)) d\gamma(x,y,\lambda,\mu) \\ &\leq \int_{X^2\times[0,1]^2} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu). \end{aligned}$$ □

2.2.3 The case of probabilities over a simplex

We assume here that $X = \Delta(K)$, where K is a non empty finite set. We use $\|p\| = \sum_k |p^k|$ for every vector $p = (p^k)_{k\in K}$ in $\mathbb{R}^K$, and view X as the set of vectors in $\mathbb{R}_+^K$ with norm 1.

$$X = \{p = (p^k)_{k\in K} \in \mathbb{R}_+^K, \sum_{k\in K} p^k = 1\}.$$

Recall that for u and v in $\Delta(X)$, we have $d_1(u,v) = \sup_{f\in D_1} |u(f) - v(f)|$, where $D_1 = \{f \in E, \forall x,y \in X, \forall a,b \geq 0,\ af(x) - bf(y) \leq \|ax - by\|\}$.

We now introduce an alternative definition of d_1 using "non revealing game functions". These functions come from the theory of repeated games with incomplete information *à la* Aumann Maschler [AMS95], and the interest for the distance d_0 emerged several years ago while doing research on Markov decision processes with partial observation and repeated games with an informed controller (see Renault [Ren11] and [Ren12b]).

Given a collection of matrices $(G^k)_{k\in K}$ (all of the same finite size $I \times J$) indexed by K and with values in $[-1,1]$, we define the "non revealing function" f in $\mathcal{C}(X)$ by:

$$\begin{aligned} \forall p \in X, f(p) &= \mathrm{Val}\left(\sum_{k\in K} p^k G^k\right), \\ &= \max_{a\in\Delta(I)} \min_{b\in\Delta(J)} \sum_{i\in I, j\in J} a(i)b(j) \left(\sum_{k\in K} p^k G^k(i,j)\right), \\ &= \min_{b\in\Delta(J)} \max_{a\in\Delta(I)} \sum_{i\in I, j\in J} a(i)b(j) \left(\sum_{k\in K} p^k G^k(i,j)\right). \end{aligned}$$

$f(p)$ is the minmax value of the average matrix $\sum_k p^k G^k$. The set of all such non revealing functions f, where I, J and $(G^k)_{k\in K}$ vary, is denoted by D_0.

Clearly, all affine functions from X to $[-1,1]$ belong to D_0. It is known that the set of non revealing functions is dense in $\mathcal{C}(X)$. However, we only consider here non revealing functions defined by matrices with values in $[-1,1]$, and D_0 is *not* dense in the set of continuous functions from X to $[-1,1]$. As an example, consider the case where $K=\{1,2\}$ and f in E is piecewise-linear with $f(1,0)=f(0,1)=0$ and $f(1/2,1/2)=1$. If a function g in D_0 is such that $g(1/2,1/2)=1$, then necessarily the values of the two matrix games G^1 and G^2 are also equal to 1 since it is the maximum value. Therefore f is not in D_0. In fact f is 1-Lipschitz, however $2f(1/2,1/2)-f(1,0)=2>\|2(1/2,1/2)-(1,0)\|=1$, so it is not in D_1 which we will see later contains D_0 (see lemma 2.2.13).

Lemma 2.2.12 *If f, g belong to D_0 and $\lambda\in[0,1]$, then $-f$, $\sup\{f,g\}$, $\inf\{f,g\}$ and $\lambda f+(1-\lambda)g$ are in D_0. The linear span of D_0 is dense in $\mathcal{C}(X)$.*

Proof: The proof can be easily deduced from proposition 5.1. page 357 in MSZ, part B. For instance, let f and g in D_0 be respectively defined by the collections of matrices $(G^k)_{k\in K}$ with size $I_1\times J_1$ and $(H^k)_{k\in K}$ with size $I_2\times J_2$.

Defining for each k, i_1, j_1: $G'^k(i_1,j_1)=-G^k(j_1,i_1)$ yields a family of matrices $(G'^k)_k$ with size $J_1\times I_1$ inducing $-f$. So $-f\in D_0$.

To get that $\sup\{f,g\}$ belongs to D_0, one can assume w.l.o.g. that $I_1\cap I_2=J_1\cap J_2=\emptyset$. Set $I=I_1\cup I_2$ and $J=J_1\times J_2$. Define for each k the matrix game L^k in $\mathbb{R}^{I\times J}$ by $L^k(i,(j_1,j_2))=G^k(i,j_1)$ if $i\in I_1$, $L^k(i,(j_1,j_2))=H^k(i,j_2)$ if $i\in I_2$. Then for each p in X, we have $\mathrm{Val}(\sum_k p^k L^k)=\sup\{f(p),g(p)\}$, so that $\sup\{f,g\}\in D_0$.

Lemma 2.2.13 *The closure of D_0 is D_1.*

Proof: We first show that $D_0\subset D_1$. Let I and J be finite sets, and $(G^k)_{k\in K}$ be a collection of $I\times J$-matrices with values in $[-1,1]$. Consider p and q in X and a and b non negative. Then for all i and j:

$$\left|\sum_k p^k a G^k(i,j)-\sum_k q^k b G^k(i,j)\right|\le\sum_k|ap^k-bq^k|=\|ap-bq\|.$$

As a consequence,

$$\begin{aligned} a\mathrm{Val}\left(\sum_{k\in K}p^kG^k\right)-b\mathrm{Val}\left(\sum_{k\in K}q^kG^k\right) &= \mathrm{Val}\left(\sum_{k\in K}ap^kG^k\right)-\mathrm{Val}\left(\sum_{k\in K}bq^kG^k\right)\\ &\le \|ap-bq\| \end{aligned}$$

We now show that the closure of D_0 is D_1. Consider f in D_1, in particular we have $\|f\|_\infty\le 1$. Let p and q be distinct elements in X, and define Y as the linear span of p and q, and define φ from Y to $\mathbb{R}$ such that: $\varphi(\lambda p+\mu q)=\lambda f(p)+\mu f(q)$ for all reals λ and μ.

If $\lambda \geq 0$ and $\mu \geq 0$, we have $\varphi(\lambda p + \mu q) \leq \lambda + \mu = \|\lambda p + \mu q\|$. If $\lambda \geq 0$ and $\mu \leq 0$, we directly use the definition of D_1 to get: $\varphi(\lambda p + \mu q) \leq \|\lambda p + \mu q\|$. As a consequence, φ is a linear form with norm at most 1 on Y. By Hahn-Banach theorem, it can be extended to a linear mapping on $\mathbb{R}^K$ with the same norm, and we denote by g the restriction of this mapping to X. g is affine with $g(p) = \varphi(p) = f(p)$ and $g(q) = \varphi(q) = f(q)$. Moreover, for each r in X, we have $\|g(r)\| \leq \|r\| = 1$. As a consequence g belongs to D_0.

Because D_0 is stable under the sup and inf operations, we can use Stone-Weierstrass theorem (see for instance lemma A7.2 in Ash p.392 [Ash72]) to conclude that f belongs to the closure of D_0. □

Definition 2.2.14 *Given u and v in $\Delta(X)$, define:*

$$d_0(u, v) = \sup_{f \in D_0} u(f) - v(f)$$

Proposition 2.2.15 *d_0 is a distance on $\Delta(X)$ metrizing the weak-* topology. Moreover $d_0 = d_1 = d_2 = d_3$.*

Proof: $d_0 = d_1 = d_2 = d_3$ follows from lemma 2.2.13 and theorem 2.2.7. Because the linear span of D_0 is dense in $\mathcal{C}(X)$, we obtain the separation property and d_0 is a distance on $\Delta(X)$. Because $D_0 \subset D_1 \subset E_1$, we have $d_0 = d_1 \leq d_{KR}$. Since $(\Delta(X), d_{KR})$ is a compact metric space, the identity map $(\Delta(X), d_{KR})$ to $(\Delta(X), d_0)$ is bicontinuous, and we obtain that $(\Delta(X), d_0)$ is a compact metric space and d_0 and d_{KR} are equivalent. (see for instance proposition 2 page 138 Aubin [Aub77]). □

Remark: one can show that allowing for infinite sets I, J in the definition of D_0 (still assuming that all games $\sum_k p^k G^k$ have a value) would not change the value of d_0.

From now on, we just write $d_*(u, v)$ for the distance $d_0 = d_1 = d_2 = d_3$ on $\Delta(X)$. Elements of X can be viewed as elements of $\Delta(X)$ (using Dirac measures), and it is well known that for p, q in X, we have: $d_{KR}(\delta_p, \delta_q) = \|p - q\|$. We have the same result with d_*.

Lemma 2.2.16 *For p, q in X, we have $d_*(\delta_p, \delta_q) = \|p - q\|$.*

Proof: Define $K_1 = \{k \in K, p^k \geq q^k\}$, and $K_2 = K \backslash K_1$. Consider f affine on X such that $f(k) = +1$ if $k \in K_1$, and $f(k) = -1$ if $k \in K_2$. Then $f \in D_1$, and $d_*(\delta_p, \delta_q) \geq |f(p) - f(q)| = \|p - q\|$. The other inequality is clear. □

We now present a dual formulation for our distance, in the spirit of Kantorovich duality formula from optimal transport. For any u, v in $\Delta(X)$, we denote by $\Pi(u, v)$ the set of transference plans, or couplings, of u and v, that is the set of probability distributions over

$X \times X$ with first marginal u and second marginal v. Recall (see for instance Villani [Vil03], p.207):

$$d_{KR}(u,v) = \sup_{f \in E_1} |u(f) - v(f)| = \min_{\gamma \in \Pi(u,v)} \int_{(x,y) \in X \times X} \|x - y\| \, d\gamma(x,y)$$

We will concentrate on probabilities on X with finite support. We denote by $Z = \Delta_f(X)$ the set of such probabilities.

Definition 2.2.17 *Let u and v be in Z with respective supports U and V. We define $\mathcal{M}_4(u,v)$ as the set*

$$\left\{ (\alpha, \beta) \in ({\mathbb{R}_+}^{U \times V})^2, s.t. \forall x \in U, \forall y \in V, \sum_{y' \in V} \alpha(x,y') = u(x) \text{ and } \sum_{x' \in U} \beta(x',y) = v(y) \right\}.$$

$$\textit{And } \; d_4(u,v) = \inf_{(\alpha,\beta) \in \mathcal{M}_4(u,v)} \sum_{(x,y) \in U \times V} \|x\alpha(x,y) - y\beta(x,y)\|$$

Notice that diagonal elements in $\mathcal{M}_4(u,v)$, i.e. measures α such that $(\alpha,\alpha) \in \mathcal{M}_4(u,v)$, coincide with elements of $\Pi(u,v)$. $\mathcal{M}_4(u,v)$ is a polytope in the Euclidean space $(\mathbb{R}^{U \times V})^2$, so the infimum in the definition of $d_4(u,v)$ is achieved.

Theorem 2.2.18 *(Duality formula) Let u and v be in Z with respective supports U and V.*

$$d_*(u,v) = \sup_{f \in D_1} |u(f) - v(f)| = \min_{(\alpha,\beta) \in \mathcal{M}_4(u,v)} \sum_{(x,y) \in U \times V} \|x\alpha(x,y) - y\beta(x,y)\|$$

where $D_1 = \{f \in E, \forall x,y \in X, \forall a,b \geq 0, \; af(x) - bf(y) \leq \|ax - by\|\}$,
and $\mathcal{M}_4(u,v) = \Big\{ (\alpha,\beta) \in {\mathbb{R}_+}^{U \times V} \times {\mathbb{R}_+}^{U \times V}, s.t. \; \forall (x,y) \in U \times V,$
$\sum_{y' \in V} \alpha(x,y') = u(x) \text{ and } \sum_{x' \in U} \beta(x',y) = v(y) \Big\}$.

The proof is postponed to the next subsection. We conclude this part by a simple but fundamental property of the distance d_*.

Definition 2.2.19 *We define the posterior mapping $\psi_{\mathbb{N}}$ from $\Delta_f(K \times \mathbb{N})$ to $\Delta_f(X)$ by:*

$$\psi_{\mathbb{N}}(\pi) = \sum_{c' \in \mathbb{N}} \pi(c') \delta_{p(c')}$$

where for each c', $\pi(c') = \sum_k \pi(k,c')$ and $p(c') = (p^k(c'))_{k \in K} \in X$ is the posterior on K given c' (defined arbitrarily if $\pi(s) = 0$) : for each k in K, $p^k(c') = \frac{\pi(k,c')}{\pi(c')}$.

$\psi_{\mathbb{N}}(\pi)$ is a probability with finite support over X. Intuitively, think of a joint variable (k,c') being selected according to π, and an agent just observes c'. His knowledge on K is then represented by $p(c')$. And $\psi_{\mathbb{N}}(\pi)$ represents the ex-ante information that the agent will know about the variable k. $\Delta(K \times \mathbb{N})$ is endowed with the $\|.\|_1$ norm. One can show that

$\psi_{\mathbb{N}}$ is continuous whenever X is endowed with the weak-* topology. Intuitively, $\psi_{\mathbb{N}}(\pi)$ has less information than π, because the agent does not care about c' itself but just on the information about k given by c'. So one may hope that the mapping $\psi_{\mathbb{N}}$ is 1-Lipschitz (non expansive) for a well chosen distance on $\Delta(X)$. This is not the case if one uses the Kantorovich-Rubinstein distance d_{KR}, as shown by the example below:

Example 2.2.20 *Consider the case where $K = \{k_1, k_2, k_3\}$ and the signal take only two values among integers $C = \{1, 2\}$. We denote by π and π' the following laws on $\Delta(K \times \mathbb{N})$:*

$$\begin{array}{c} \\ K \\ \\ \\ \pi \end{array} \begin{array}{c} C \\ \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix} \end{array} \quad \textit{and} \quad \begin{array}{c} C \\ \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \\ 0 & \frac{1}{4} \end{pmatrix} \\ \pi' \end{array}.$$

Their disintegrations are respectively $\psi_{\mathbb{N}}(\pi) = \frac{1}{2}\begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ *and* $\psi_{\mathbb{N}}(\pi') = \frac{1}{4}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \frac{3}{4}\begin{pmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$.
We define the test function $f : \Delta(K) \to [-1, 1]$ by

$$f\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{3} \quad , \quad f\begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix} = -\frac{1}{3},$$
$$f\begin{pmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{pmatrix} = 1 \quad \textit{and} \quad f\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{2}{3}.$$

We have $\|\pi - \pi'\| = \frac{1}{2}$ and since f is 1-Lipschitz, $d_{KR}(\psi_{\mathbb{N}}(\pi), \psi_{\mathbb{N}}(\pi')) \geq \psi_{\mathbb{N}}(\pi')(f) - \psi_{\mathbb{N}}(\pi)(f) = \frac{11}{12} - 0 > \frac{1}{2}$. The posterior mapping $\psi_{\mathbb{N}}$ is not 1-Lipschitz from $(\Delta(K \times \mathbb{N}), \|.\|_1)$ to $(\Delta(X), d_{KR})$
.

However, the next proposition shows that the distance d_* has the desirable property.

Proposition 2.2.21 *The mapping $\psi_{\mathbb{N}}$ is 1-Lipschitz from $(\Delta(K \times \mathbb{N}), \|.\|_1)$ to $(\Delta_f(X), d_*)$.*
Moreover, d_* is the largest distance on Z having this property: given u and v in Z, we have

$$d_*(u, v) = \inf\{\|\pi - \pi'\|_1, \ s.t.\ \pi, \pi' \in \Delta_f(K \times \mathbb{N}), \ \psi_{\mathbb{N}}(\pi) = u, \ \psi_{\mathbb{N}}(\pi') = v\}.$$

Proof: First fix π, π' in $\Delta_f(K \times \mathbb{N})$. Write $u = \psi_{\mathbb{N}}(\pi)$, $u' = \psi_{\mathbb{N}}(\pi')$ and C such that the

support of π and π' is included in $K \times C$. For any f in D_1, we have:

$$\begin{aligned} u(f) - u'(f) &= \sum_{c \in C} (\pi(c) f(p(c)) - \pi'(c) f(p'(c)) \\ &\leq \sum_{c \in C} \|\pi(c)p(c) - \pi'(c)p'(c)\| \\ &\leq \sum_{c \in C} \|(\pi(k,c))_k - (\pi'(k,c))_k\| \\ &\leq \sum_{c \in C} \sum_{k \in K} |\pi(k,c) - \pi'(k,c)| = \|\pi - \pi'\|_1. \end{aligned}$$

So $d_*(u, u') \leq \|\pi - \pi'\|_1$, and $\psi_{\mathbb{N}}$ is 1-Lipschitz.

Let now u and v be in Z. There exists $(\alpha, \beta) \in \mathcal{M}_4(u, v)$ such that

$$d_*(u, v) = \sum_{(x,y) \in U \times V} \|\alpha(x, y)x - \beta(x, y)y\|.$$

Choose an enumeration of $C = U \times V$ in order to embed this set in $\mathbb{N}$ and define $\pi, \pi' \in \Delta(K \times C)$ by $\pi(k, (x, y)) = x(k)\alpha(x, y)$ and $\pi'(k, (x, y)) = y(k)\beta(x, y)$. By definition of $\mathcal{M}_4(u, v)$, π and π' are probabilities and

$$\begin{aligned} \|\pi - \pi'\|_{1,K\times\mathbb{N}} &= \sum_{k \in K, (x,y) \in U \times V} |x(k)\alpha(x, y) - y(k)\beta(x, y)| \\ &= \sum_{(x,y) \in U \times V} \|\alpha(x, y)x - \beta(x, y)y\|. \end{aligned}$$

□

2.2.4 Proof of the duality formula

Let u and v be in $\Delta(X)$, and denote by U and V the respective supports of u and v. We write $S = X^2 \times [0, 1]^2$, and we start with a lemma, where no finiteness assumption on U or V is needed.

Lemma 2.2.22 *For each $\gamma \in \mathcal{M}_3(u, v)$, we have:*

$$\int_{X^2 \times [0,1]^2} \|\lambda x - \mu y\| d\gamma(x, y, \lambda, \mu) = 2 + \int_{U \times V \times [0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu)\, d\gamma(x, y, \lambda, \mu).$$

Proof: Write $A(\gamma) = \int_S \|\lambda x - \mu y\| d\gamma(x, y, \lambda, \mu)$. By definition of $\mathcal{M}_3(u, v)$, we have:

$$\int_S \lambda \mathbf{1}_{x \notin U} d\gamma = 0, \text{ and } \int_S \mu \mathbf{1}_{y \notin V} d\gamma = 0.$$

So that $\lambda \mathbf{1}_{x\notin U} = \mu \mathbf{1}_{y\notin V} = 0$ γ. a.e. We can write:

$$A(\gamma) = \int_S \mathbf{1}_{x\in U, y\in V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) + \int_S \mathbf{1}_{x\in U, y\notin V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu)$$

$$+ \int_S \mathbf{1}_{x\notin U, y\in V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) + \int_S \mathbf{1}_{x\notin U, y\notin V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu)$$

$$= \int_S \mathbf{1}_{x\in U, y\in V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) + \int_S \mathbf{1}_{x\in U, y\notin V} \lambda d\gamma(x,y,\lambda,\mu) + \int_S \mathbf{1}_{x\notin U, y\in V} \mu d\gamma(x,y,\lambda,\mu) + 0.$$

We also have by definition of $\mathcal{M}_3(u,v)$ that $1 = \int_S \mathbf{1}_{x\in U} \lambda d\gamma$, so that:

$$1 = \int_S \mathbf{1}_{x\in U, y\in V} \lambda d\gamma + \int_S \mathbf{1}_{x\in U, y\notin V} \lambda d\gamma.$$

And similarly $1 = \int_S \mathbf{1}_{x\in U, y\in V} \mu d\gamma + \int_S \mathbf{1}_{x\notin U, y\in V} \mu d\gamma$. We obtain:

$$A(\gamma) = 2 + \int_S \mathbf{1}_{x\in U, y\in V} \|\lambda x - \mu y\| d\gamma(x,y,\lambda,\mu) - \int_S \mathbf{1}_{x\in U, y\in V} \lambda d\gamma - \int_S \mathbf{1}_{x\in U, y\in V} \mu d\gamma.$$

□

We assume in the sequel that U and V are finite, and define $d_5(u,v)$ as follows:

Definition 2.2.23 *Define*

$$\mathcal{M}_5(u,v) = \Big\{(\alpha,\beta) = (\alpha(x,y),\beta(x,y))_{(x,y)\in U\times V} \in (\mathbb{R}^{U\times V})^2, s.t. \forall x \in U, \forall y \in V,$$

$$\alpha(x,y) \geq 0, \beta(x,y) \geq 0, \sum_{y'\in V} \alpha(x,y') \leq u(x) \text{ and } \sum_{x'\in U} \beta(x',y) \leq v(y)\Big\}.$$

$$\textit{And } d_5(u,v) = \inf_{(\alpha,\beta)\in\mathcal{M}_5(u,v)} 2 + \sum_{(x,y)\in U\times V} \left(\|x\alpha(x,y) - y\beta(x,y)\| - \alpha(x,y) - \beta(x,y)\right).$$

$\mathcal{M}_5(u,v)$ is a polytope in the Euclidean space $(\mathbb{R}^{U\times V})^2$, so the infimum in the definition of $d_5(u,v)$ is achieved.

Lemma 2.2.24 $d_3(u,v) \geq d_5(u,v)$.

Proof: Let γ be in $\mathcal{M}_3(u,v)$. Fix for a while (x,y) in $U\times V$, and assume that $\gamma(x,y) > 0$. We define $\gamma(.|x,y)$ the conditional probability on $[0,1]^2$ given (x,y) by: for all $\varphi \in C([0,1]^2)$,

$$\int_{[0,1]^2} \varphi(\lambda,\mu) d\gamma(\lambda,\mu|x,y) = \frac{1}{\gamma(x,y)} \int_{(x',y',\lambda,\mu)\in S} \mathbf{1}_{x'=x,y'=y} \varphi(\lambda,\mu) d\gamma(x',y',\lambda,\mu).$$

So that

$$\gamma(x,y) \int_{[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) d\gamma(\lambda,\mu|x,y) = \int_{(\lambda,\mu)\in[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) d\gamma(x,y,\lambda,\mu).$$

The mapping $\Psi : (\lambda, \mu) \mapsto \|\lambda x - \mu y\| - \lambda - \mu$ is convex so by Jensen's inequality we get:

$$\int_{(\lambda,\mu)\in[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) d\gamma(\lambda, \mu|x, y) \geq$$

$$\|x \int_{(\lambda,\mu)\in[0,1]^2} \lambda d\gamma(\lambda, \mu|x, y) - y \int_{(\lambda,\mu)\in[0,1]^2} \mu d\gamma(\lambda, \mu|x, y)\|$$

$$- \int_{(\lambda,\mu)\in[0,1]^2} \lambda d\gamma(\lambda, \mu|x, y) - \int_{(\lambda,\mu)\in[0,1]^2} \mu d\gamma(\lambda, \mu|x, y).$$

We write:

$$P(x, y) = \int_{(\lambda,\mu)\in[0,1]^2} \lambda d\gamma(\lambda, \mu|x, y) \text{ and } Q(x, y) = \int_{(\lambda,\mu)\in[0,1]^2} \mu d\gamma(\lambda, \mu|x, y),$$

so that

$$\int_{(\lambda,\mu)\in[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) d\gamma(\lambda, \mu|x, y) \geq \|xP(x, y) - yQ(x, y)\| - P(x, y) - Q(x, y).$$

Now, by lemma 2.2.22

$$\begin{aligned} A(\gamma) &= 2 + \sum_{x\in U, y\in V} \int_{(\lambda,\mu)\in[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) \, d\gamma(x, y, \lambda, \mu) \\ &= 2 + \sum_{x\in U, y\in V, \gamma(x,y)>0} \int_{(\lambda,\mu)\in[0,1]^2} (\|\lambda x - \mu y\| - \lambda - \mu) \, d\gamma(x, y, \lambda, \mu) \\ &\geq 2 + \sum_{x\in U, y\in V, \gamma(x,y)>0} \gamma(x, y) \left(\|xP(x, y) - yQ(x, y)\| - P(x, y) - Q(x, y)\right). \end{aligned}$$

For (x, y) in $U \times V$, define $\alpha(x, y) = \gamma(x, y)P(x, y) \geq 0$ *and* $\beta(x, y) = \gamma(x, y)Q(x, y) \geq 0$ (with $\alpha(x, y) = \beta(x, y) = 0$ if $\gamma(x, y) = 0$). We get:

$$A(\gamma) \geq 2 + \sum_{x\in U, y\in V} \left(\|x\alpha(x, y) - y\beta(x, y)\| - \alpha(x, y) - \beta(x, y)\right).$$

And we have, for each x in U:

$$\begin{aligned} \sum_{y\in V} \alpha(x, y) &= \sum_{y\in V, \gamma(x,y)>0} \int_{(\lambda,\mu)\in[0,1]^2} \lambda d\gamma(x, y, \lambda, \mu) \\ &\leq \int_{(y,\lambda,\mu)\in X\times[0,1]^2} \lambda d\gamma(x, y, \lambda, \mu) = u(x). \end{aligned}$$

where the last equality comes from the definition of $\mathcal{M}_3(u, v)$. Similarly, for each y in V we can show that $\sum_{x\in U} \beta(x, y) \leq v(y)$, and lemma 2.2.24 is proved. □

Lemma 2.2.25 $d_5(u, v) \geq d_4(u, v)$.

Proof: Consider (α^*, β^*) achieving the minimum in the definition of $d_5(u,v)$. Assume that there exists x^* such that $\sum_{y\in V}\alpha(x^*,y) < u(x^*)$. For any x in X and z in $\mathbb{R}_+^K$, one can check that the mapping $l : (\alpha \mapsto \|x\alpha - z\| - \alpha)$ is nonincreasing from $\mathbb{R}_+$ to $\mathbb{R}$ (as the sum of the mappings $l_k : (\alpha \mapsto |\alpha x^k - z^k| - \alpha x^k)$, each l^k being non increasing in α). As a consequence, one can choose any y^* in V and increase $\alpha(x^*,y^*)$ in order to saturate the constraint without increasing the objective. So we can assume without loss of generality that $\sum_{y\in V}\alpha(x^*,y) = u(x^*)$ for all x^* and similarly $\sum_{x\in U}\beta(x,y^*) = v(y^*)$ for all y^*.

Consequently,

$$\begin{aligned} d_5(u,v) &= 2 + \sum_{(x,y)\in U\times V} (\|x\alpha^*(x,y) - y\beta^*(x,y)\| - \alpha^*(x,y) - \beta^*(x,y)) \\ &= \sum_{(x,y)\in U\times V} \|x\alpha^*(x,y) - y\beta^*(x,y)\| \geq d_4(u,v). \end{aligned}$$

Lemma 2.2.26 $d_4(u,v) \geq d_2(u,v)$.

Proof: Fix $(f,g) \in D_2$ and $(\alpha,\beta) \in \mathcal{M}_4(u,v)$.

$$\begin{aligned} u(f) + v(g) &= \sum_{x\in U} f(x)u(x) + \sum_{y\in Y} g(y)v(y) \\ &= \sum_{(x,y)\in U\times V} f(x)\alpha(x,y) + g(y)\beta(x,y) \\ &\leq \sum_{(x,y)\in U\times V} \|\alpha(x,y)x - \beta(x,y)y\| \leq d_4(u,v). \end{aligned}$$

We have shown that $d_3(u,v) \geq d_5(u,v) \geq d_4(u,v) \geq d_2(u,v) = d_3(u,v) = d_1(u,v)$. This ends the proof of theorem 2.2.18.

2.3 Long-term values for compact non expansive Markov Decision Processes

In this section we consider Markov Decision Processes, or Controlled Markov Chains, with bounded payoffs and transitions with finite support. We will consider two closely related models of MDP and prove in each case the existence and a characterization for a general notion of long-term value. The first model deals with MDP without any explicit action set (hence, payoffs only depend on the current state), such MDP will be called *gambling houses* using the terminology of gambling theory (see Maitra and Sudderth [MS96]). We will assume in this setup that the set of states X is metric compact and that the transitions are non expansive with respect to the KR-distance on $\Delta(X)$. Since we only use the KR-distance here, the theorem for the first model, namely theorem 2.3.9, does not use the distance for belief spaces studied in section 2.2. The second model is the standard model of Markov Decision Processes with states, actions,

transitions and payoffs, and we will assume that the state space X is a compact subset of a simplex $\Delta(K)$. We will need for this second case an assumption of non expansiveness for the transitions which is closely related to the distance d_* introduced in section 2.2, see theorem 2.3.19 later. The applications in sections 2.4.1 and 2.4.2 will be based on the second model.

2.3.1 Long-term values for Gambling Houses

In this section we consider Markov Decision Processes of the following form. There is a non empty set of states X, a transition given by a multi-valued mapping $F : X \rightrightarrows \Delta_f(X)$ with non empty values, and a payoff (or reward) function $r : X \to [0,1]$. The idea is that given an initial state x_0 in X, a decision-maker (or player) can choose a probability with finite support u_1 in $F(x_0)$, then x_1 is selected according to u_1 and there is a payoff $r(x_1)$. Then the player has to select u_2 in $F(x_1)$, x_2 is selected according to u_1 and the player receives the payoff $r(x_2)$, etc... Note that there is no explicit action set here, and that the transitions take values in $\Delta_f(X)$ and hence all have finite support.

We say that $\Gamma = (X, F, r)$ is a Gambling House. We assimilate the elements in X with their Dirac measures in $\Delta(X)$, and in case the values of F only consist of Dirac measures on X, we view F as a correspondence from X to X and say that Γ is a *deterministic* Gambling House (or a Dynamic Programming problem). In general we write $Z = \Delta_f(X)$, and an element in Z is written $u = \sum_{x\in X} u(x)\delta_x$. The set of stages is $\mathbb{N}^* = \{1, ..., t, ...\}$, and a probability distribution over stages is called an evaluation. Given an evaluation $\theta = (\theta_t)_{t\geq 1}$ and an initial stage x_0 in X, the θ-problem $\Gamma_\theta(x_0)$ is the problem induced by a decision-maker starting from x_0 and maximizing the expectation of $\sum_{t\geq 1} \theta_t r(x_t)$.

Formally, we first linearly extend r and F to $\Delta_f(X)$ by defining for each $u = \sum_{x\in X} u(x)\delta_x$ in Z, the payoff $r(u) = \sum_{x\in X} r(x)u(x)$ and the transition $F(u) = \{\sum_{x\in X} u(x)f(x), s.t.\ f : X \to Z$ and $f(x) \in F(x) \forall x \in X\}$. We also define the *mixed extension* of F as the correspondence from Z to itself which associates to every $u = \sum_{x\in X} u(x)\delta_x$ in $\Delta_f(X)$ the image:

$$\hat{F}(u) = \left\{ \sum_{x\in X} u(x)f(x),\ s.t.\ f : X \to Z \text{ and } f(x) \in \text{conv}F(x)\ \forall x \in X \right\}.$$

The graph of $\hat{F}$ is the convex hull of the graph of F. Moreover $\hat{F}$ is an affine correspondence, as shown by the lemma below.

Lemma 2.3.1 *$\forall u, u' \in Z$, $\forall \alpha \in [0,1]$, $\hat{F}(\alpha u + (1-\alpha)u') = \alpha\hat{F}(u) + (1-\alpha)\hat{F}(u')$.*

Proof: The $\subset$ part is clear. To see the reverse inclusion, let

$$v = \alpha \sum_{x\in X} u(x)f(x) + (1-\alpha) \sum_{x\in X} u'(x)f'(x)$$

be in $\alpha\hat{F}(u) + (1-\alpha)\hat{F}(u')$, with transparent notations. Define

$$h(x) = \frac{\alpha u(x) f(x) + (1-\alpha)u'(x)f'(x)}{\alpha u(x) + (1-\alpha)u'(x)},$$

for each x such that the denominator is positive. Then $h(x) \in \text{conv}F(x)$, and

$$v = \sum_{x \in X} (\alpha u(x) + (1-\alpha)u'(x))h(x) \in \hat{F}(\alpha u + (1-\alpha)u').$$

Definition 2.3.2 *A pure play, or deterministic play, at x_0 is a sequence $\sigma = (u_1, ..., u_t, ...) \in Z^\infty$ such that $u_1 \in F(x_0)$ and $u_{t+1} \in F(u_t)$ for each $t \geq 1$. A play, or mixed play, at x_0 is a sequence $\sigma = (u_1, ..., u_t, ...) \in Z^\infty$ such that $u_1 \in \text{conv}F(x_0)$ and $u_{t+1} \in \hat{F}(u_t)$ for each $t \geq 1$. We denote by $\Sigma(x_0)$ the set of mixed plays at x_0.*

A pure play is a particular case of a mixed play. Mixed plays corresponds to situations where the decision-maker can select, at every stage t and state x_{t-1}, *randomly* the law u_t of the new state. A mixed play at x_0 naturally induces a probability distribution over the set $(X \times \Delta_f(X))^\infty$ of sequences $(x_0, u_0, x_1, u_1, ...)$, where X and Z are endowed with the discrete σ-algebra and $(X \times \Delta_f(X))^\infty$ is endowed with the product σ-algebra.

Definition 2.3.3 *Given an evaluation θ, the θ-payoff of a play $\sigma = (u_1, ..., u_t, ...)$ is defined as: $\gamma_\theta(\sigma) = \sum_{t \geq 1} \theta_t r(u_t)$, and the θ-value at x_0 is:*

$$v_\theta(x_0) = \sup_{\sigma \in \Sigma(x_0)} \gamma_\theta(\sigma).$$

It is easy to see that the supremum in the definition of v_θ can be taken over the set of pure plays at x_0. We have the following recursive formula. For each evaluation $\theta = (\theta_t)_{t \geq 1}$ such that $\theta_1 < 1$, we denote by θ^+ the "shifted" evaluation $(\frac{\theta_{t+1}}{1-\theta_1})_{t \geq 1}$. We extend linearly v_θ to Z, so that the recursive formula can be written:

$$\forall \theta \in \Delta(\mathbb{N}^*), \forall x \in X,\ v_\theta(x) = \sup_{u \in \text{conv}F(x)} \left(\theta_1 r(u) + (1-\theta_1) v_{\theta^+}(u)\right).$$

And by linearity the supremum can be taken over $F(x)$. It is also easy to see that for all evaluation θ and initial state x, we have the inequality:

$$|v_\theta(x) - \sup_{u \in F(x)} v_\theta(u)| \leq \theta_1 + \sum_{t \geq 2} |\theta_t - \theta_{t-1}|. \tag{2.1}$$

In this chapter, we are interested in the limit behavior when the decision-maker has a regular evaluation. Given an evaluation θ, we define :

$$I(\theta) = \sum_{t \geq 1} |\theta_{t+1} - \theta_t|$$

The decision-maker is considered to have a regular evaluation whenever $I(\theta)$ is small, so $I(\theta)$ may be seen as the irregularity of θ (see Sorin, [Sor02] p. 105 and Renault [Ren12a]). When $\theta = (\theta_t)_{t\geq 1}$ is non increasing, then $I(\theta)$ is just θ_1 and coincides with a measure of the impatience. A classic example is when $\theta = \sum_{t=1}^{n} \frac{1}{n}\delta_t$, the value v_θ is just denoted v_n and the evaluation corresponds to the average payoff from stage 1 to stage n. In this case $I(\theta) = 1/n \longrightarrow_{n\to\infty} 0$. We also have $I(\theta) = 2/n$ if $\theta = \sum_{t=m}^{m+n-1} \frac{1}{n}\delta_t$ for some $m \geq 2$. Another example is the case of discounted payoffs, when $\theta = (\lambda(1-\lambda)^{t-1})_{t\geq 1}$ for some discount factor $\lambda \in (0,1]$, and in this case $I(\theta) = \lambda \longrightarrow_{\lambda\to 0} 0$.

Definition 2.3.4 *The Gambling House $\Gamma = (X, F, r)$ has a general limit value v^* if (v_θ) uniformly converges to v^* when $I(\theta)$ goes to zero, i.e.:*

$$\forall \varepsilon > 0, \exists \alpha > 0, \forall \theta \in \Delta(\mathbb{N}^*), \ \ (\ I(\theta) \leq \alpha \implies (\forall x \in X, |v_\theta(x) - v^*(x)| \leq \varepsilon)\ ).$$

The existence of the general limit value implies in particular that $(v_n)_n$ and $(v_\lambda)_\lambda$ converge to the same limit when n goes to $+\infty$ and λ goes to 0. This is coherent with the result of Lehrer and Sorin [LS92], which states that the uniform convergence of $(v_n)_n$ and $(v_\lambda)_\lambda$ are equivalent.

In the definition of the general limit value, we require all value functions to be close to v^* when the regularity is high, but the plays used may depend on the precise expression of θ. In the following definition, we require the same play to be simultaneously optimal for all θ regular enough.

Definition 2.3.5 *The Gambling House $\Gamma = (X, F, r)$ has a general uniform value if it has a general limit value v^* and moreover for each $\varepsilon > 0$ one can find $\alpha > 0$ and for each initial state x a mixed play $\sigma(x)$ at x satisfying:*

$$\forall \theta, \ \ (I(\theta) \leq \alpha \implies (\forall x \in X, \gamma_\theta(\sigma(x)) \geq v^*(x) - \varepsilon)\ ).$$

Up to now, the literature in repeated games has focused on the evaluations $\theta = \sum_{t=1}^{n} \frac{1}{n}\delta_t$ and $\theta = (\lambda(1-\lambda)^{t-1})_{t\geq 1}$. The standard (Cesàro)-uniform value can be defined by restricting the evaluations to be Cesàro means: for each $\varepsilon > 0$ one can find n_0 and for each initial state x a mixed play $\sigma(x)$ at x satisfying: $\forall n \geq n_0, \forall x \in X, \gamma_n(\sigma(x)) \geq v^*(x) - \varepsilon$. Recently, Renault [Ren11]) considered deterministic Gambling Houses and characterized the uniform convergence of the value functions $(v_n)_n$. He also proved the existence of the standard Cesàro-uniform value under some assumptions, including the case where the set of states X is metric precompact, the transitions are non expansive and the payoff function is uniformly continuous. As a corollary, he proved the existence of the uniform value in Partial Observation Markov Decision Processes with finite set of states (after each stage the decision-maker just observes a stochastic signal more or less correlated to the new state).

We now present our main theorem for Gambling Houses. Equation (2.1) implies that the general limit value v^* necessarily has to satisfy some rigidity property. The function v^* (or more precisely its linear extension to Z) can only be an "excessive function" in the terminology of potential theory (Choquet [Cho56]) and gambling houses (Dubins and Savage [DS65], Maitra and Sudderth [MS96]).

Definition 2.3.6 *An affine function w defined on Z (or $\Delta(X)$) is said to be excessive if for all x in X, $w(x) \geq \sup_{u \in F(x)} w(u)$.*

Example 2.3.7 Let us consider the splitting transition given by K a finite set, $X = \Delta(K)$ and $\forall x \in X, F(x) = \{u \in \Delta(X),\ \sum_{p \in X} u(p)p = x\}$. Then the function w from $Z = \Delta(X)$ to $[0,1]$ is excessive if and only if the restriction of w to X is concave. Moreover given $u, u' \in \Delta(X)$, $u' \in \hat{F}(u)$ if and only u' is the sweeping of u as defined by Choquet [Cho56]: for all continuous concave functions f from X to $[0,1]$, $u'(f) \leq u(f)$.

Assume now that X is a compact metric space and r is continuous. r is naturally extended to an affine continuous function on $\Delta(X)$ by $r(u) = \int_{p \in X} r(p) du(p)$ for all Borel probabilities on X. In the following definition, we consider the closure of the graph of $\hat{F}$ within the (compact) set $\Delta(X \times X)$.

Definition 2.3.8 *An element u in $\Delta(X)$ is said to be an invariant measure of the Gambling House $\Gamma = (X, F, r)$ if $(u,u) \in \mathrm{cl}(Graph\ \hat{F})$. The set of invariant measures of Γ is denoted by R, so that:*

$$R = \{u \in \Delta(X), (u,u) \in \mathrm{cl}(Graph\ \hat{F})\}.$$

R is a convex compact subset of $\Delta(X)$. Recall that for u and u' in $\Delta(X)$, the Kantorovich-Rubinstein distance between u and u' is denoted by $d_{KR}(u,u') = \sup_{f \in E_1} |u(f) - u'f)|$.

Theorem 2.3.9 *Consider a Gambling House $\Gamma = (X, F, r)$ such that X is a compact metric space, r is continuous and F is non expansive with respect to the KR distance:*

$$\forall x \in X, \forall x' \in X, \forall u \in F(x), \exists u' \in F(x') s.t.\ d_{KR}(u,u') \leq d(x,x').$$

Then the Gambling House has a general uniform value v^* characterized by:

$$\begin{aligned} \forall x \in X,\ v^*(x) \ &= \inf \ \Big\{ w(x), w : \Delta(X) \to [0,1] \text{ affine } C^0 \ s.t. \\ &\qquad (1)\ \forall y \in X, w(y) \geq \sup_{u \in F(y)} w(u) \text{ and } (2) \forall u \in R, w(u) \geq r(u) \Big\}. \end{aligned}$$

That is, v^* is the smallest continuous affine function on X which is 1) excessive and 2) above the running payoff r on invariant measures.

Notice that:

1) when $\Gamma = (X, F, r)$ is deterministic, the hypotheses are satisfied as soon as X is metric compact for some metric d, r is continuous and F is non expansive for d.

2) when X is finite, one can use the distance $d(x, x') = 2$ for all $x \neq x'$ in X, so that for u and u' in $\Delta(X)$, $d_{KR}(u, u') = \|u - u'\|_1 = \sum_{x \in X} |u(x) - u'(x)|$, and the hypotheses are automatically satisfied. We will prove later a more general result for a model of MDP with finite state space, allowing for explicit actions influencing transitions and payoffs (see corollary 2.3.20).

Remark 2.3.10 The formula also holds when there is no decision maker, i.e. when F is single-valued, and there are some similarities with the Von Neumann ergodic theorem [VN32]. Let Z be a Hilbert space and Q be a linear isometry on Z, this theorem states that for all $z \in Z$, the sequence $z_n = \frac{1}{n} \sum_{t=1}^{n} Q^t(z)$ converges to the projection z^* of z on the set R of fixed points of Q. Using the linearity and the non expansiveness leads to a characterization by the set of fixed points. In particular, having in mind linear payoff functions of the form $(z \mapsto < l, z >)$, we have that the projection z^* of z on R is characterized by:

$$\forall l \in Z, < l, z^* > = < l^*, z > = \inf\{< l', z >, l' \in R \text{ and } < l', u > \geq < l, u > \forall u \in R\}.$$

Example 2.3.11 We consider here a basic periodic sequence of 0 and 1. Let $X = \{0, 1\}$ and for all $x \in X$, $F(x) = \{1 - x\}$ and $r(x) = x$. There is a unique invariant measure $u = 1/2\delta_0 + 1/2\delta_1$, and the general uniform value exists and satisfies $v^*(x) = \frac{1}{2}$ for all states x. Notice that considering evaluations $\theta = (\theta_t)_t$ such that θ_t is small for each t without requiring $I(\theta)$ small, would not necessarily lead to v^*. Consider for instance $\theta^n = \sum_{t=1}^{n} \frac{1}{n} \delta_{2t}$ for each n, we have $v_{\theta^n}(x) = x$ for all x in X.

Example 2.3.12 The state space is the unit circle, let $X = \{x \in \mathbb{C}, |x| = 1\}$ and $F(e^{i\alpha}) = e^{i(\alpha+1)}$ for all real α. If we denote by μ the uniform distribution (Haar probability measure) on the circle, the mapping F is μ-ergodic and μ is F-invariant. By Birkhoff's theorem [Bir31], we know that the time average converges to the space average μ-almost surely. Here μ is the unique invariant measure, and we obtain that the general uniform value is the constant:

$$\forall x \in X,\ v^*(x) = \frac{1}{2\pi} \int_0^{2\pi} r(e^{i\alpha}) d\alpha.$$

Notice that the value $v_\theta(x)$ converges to $v^*(x)$ for all x in X, and not only for μ-almost all x in X.

Example 2.3.13 Let $\Gamma = (X, F, r)$ be a gambling house satisfying the hypotheses of the theorem 2.3.9 such that for all $x \in X$, $\delta_x \in F(x)$. Therefore the set R is equal to $\Delta(X)$. In the terminology of Gambling Theory (see Maitra Sudderth, [MS96]), Γ is called a *leavable* gambling house since at each stage the player can stay at the current state. The limit value v^* is here characterized by:

$$v^* = \inf\{v : X \to [0, 1]\ C^0, v \textit{ is excessive and } v \geq r\}.$$

In the above formula, v excessive means: $\forall x \in X, v(x) \geq \sup_{u \in F(x)} I\!E_u(v)$. This is a variant of the *fundamental theorem of gambling theory* (see section 3.1 in Maitra Sudderth [MS96]).

Example 2.3.14 The following deterministic Gambling House, which is an extension of example 1.4.4. in Sorin [Sor02] and of example 5.2 of Renault [Ren11], shows that the assumptions of theorem 2.3.9 allow for many speeds of convergence to the limit value v^*. Here $l > 1$ is a fixed parameter, X is the simplex $\{x = (p^a, p^b, p^c) \in \mathbb{R}^3_+, p^a + p^b + p^c = 1\}$ and the initial state is $x_0 = (1, 0, 0)$. The payoff is $r(p^a, p^b, p^c) = p^b - p^c$, and the transition is defined by: $F(p^a, p^b, p^c) = \{((1 - \alpha - \alpha^l)p^a, p^b + \alpha p^a, p^c + \alpha^l p^a), \alpha \in [0, 1/2]\}$.

The probabilistic interpretation is the following: there are 3 points a, b and c, and the initial point is a. The payoff is 0 at a, it is +1 at b, and -1 at c. At point a, the decision maker has to choose $\alpha \in [0, 1/2]$: then b is reached with probability α, c is reached with probability α^l, and the play stays in a with the remaining probability $1 - \alpha - \alpha^l$. When b (resp. c) is reached, the play stays at b (resp. c) forever. So the decision maker starting at point a wants to reach b and to avoid c. By playing at each stage $\alpha > 0$ small enough, he can get as close to b as he wants.

Back to our deterministic setup, we use norm $\|.\|_1$ and obtain that X is compact, F is non expansive and r is continuous, so that theorem 2.3.9 applies. The limit value is given by $v^*(p^a, p^b, p^c) = p^a + p^b$, and if we denote by x_λ the value $v_\lambda(x_0)$, we have for all $\lambda \in (0, 1]$: $x_\lambda = \phi(x_\lambda)$, where for all $x \in \mathbb{R}$,

$$\phi(x) = \max_{\alpha \in [0,1/2]} (1 - \lambda)(1 - \alpha - \alpha^l)x + \alpha.$$

Since $x_\lambda \in (0, 1)$, the first order condition gives $(1 - \lambda)x_\lambda(-1 - l\alpha^{l-1}) + 1 = 0$ and we can obtain:

$$x_\lambda = \frac{1}{(1 - \lambda)} \left(l \left(\frac{\lambda}{(1 - \lambda)(l - 1)} \right)^{\frac{l-1}{l}} + 1 \right)^{-1}.$$

Finally we can compute an equivalent of x_λ as λ goes to 0. We have

$$\left(\frac{\lambda}{(1 - \lambda)(l - 1)} \right)^{\frac{l-1}{l}} = (\frac{1}{l - 1})^{\frac{l-1}{l}} \lambda^{\frac{l-1}{l}} (1 + o(\lambda^{\frac{l-1}{l}}))$$

so that

$$v_\lambda(x_0) = (1 - \lambda) \frac{1}{l \left((\frac{1}{l-1})^{\frac{l-1}{l}} \lambda^{\frac{l-1}{l}} + o(\lambda^{\frac{2l-2}{l}}) \right) + 1}$$

$$v_\lambda(x_0) = 1 - C\lambda^{\frac{l-1}{l}} + o(\lambda^{\frac{l-1}{l}}) \text{ with } C = \frac{l}{(l - 1)^{\frac{(l-1)}{l}}}.$$

2.3.2 Long-term values for standard MDPs

A standard Markov Decision Process Ψ is given by a non empty set of states X, a non empty set of actions A, a mapping $q : X \times A \to \Delta_f(X)$ and a payoff function $g : X \times A \to [0, 1]$. At

each stage, the player learns the current state x and chooses an action a. He then receives the payoff $g(k,a)$, a new state is drawn accordingly to $q(k,a)$ and the game proceeds to the next stage.

Definition 2.3.15 *A pure, or deterministic, strategy is a sequence of mappings $\sigma = (\sigma_t)_{t\geq 1}$ where $\sigma_t : (X \times A)^{t-1} \to A$ for each t. A strategy (or behavioral strategy) is a sequence of mappings $\sigma = (\sigma_t)_{t\geq 1}$ where $\sigma_t : (X \times A)^{t-1} \to \Delta_f(A)$ for each t. We denote by Σ the set of strategies.*

A pure strategy is a particular case of strategy. An initial state x_1 in X and a strategy σ naturally induce a probability distribution with finite support over the set of finite histories $(X \times A)^n$ for all n, which can be uniquely extended to a probability over the set $(X \times A)^\infty$ of infinite histories.

Definition 2.3.16 *Given an evaluation θ and an initial state x_1 in X, the θ-payoff of a strategy σ at x_1 is defined as $\gamma_\theta(x_1,\sigma) = I\!E_{x_1,\sigma}\left(\sum_{t\geq 1}\theta_t g(x_t,a_t)\right)$, and the θ-value at x_1 is:*

$$v_\theta(x_1) = \sup_{\sigma\in\Sigma}\gamma_\theta(x_1,\sigma).$$

As for gambling houses, it is easy to see that the supremum can be taken over the smaller set of pure strategies, and one can derive a recursive formula linking the value functions. General limit and uniform values are defined as in the previous subsection 2.3.1.

Definition 2.3.17 *Let $\Psi = (X, A, q, g)$ be a standard MDP.*

Ψ has a general limit value v^* if (v_θ) uniformly converges to v^* when $I(\theta)$ goes to zero, i.e. for each $\varepsilon > 0$ one can find $\alpha > 0$ such that:

$$\forall\theta,\ \left(\ I(\theta)\leq\alpha \implies (\forall x\in X, |v_\theta(x) - v^*(x)|\leq\varepsilon)\ \right).$$

Ψ has a general uniform value if it has a general limit value v^* and moreover for each $\varepsilon > 0$ one can find $\alpha > 0$ and a behavior strategy $\sigma(x)$ for each initial state x satisfying:

$$\forall\theta,\ \left(I(\theta)\leq\alpha \implies (\forall x\in X, \gamma_\theta(x,\sigma(x))\geq v^*(x)-\varepsilon)\ \right).$$

We now present a notion of invariance for the MDP Ψ. The next definition will be similar to definition 2.3.8, however one needs to be slightly more sophisticated here to incorporate the payoff component. Assume now that X is a compact metric space, and define for each (u,y) in $\Delta_f(X)\times[0,1]$,

$$\hat{F}(u,y) = \left\{\left(\sum_{x\in X}u(x)q(x,a(x)), \sum_{x\in X}u(x)g(x,a(x))\right),\ \textit{where } a : X\to\Delta_f(A)\right\}.$$

where $q(x,.)$ and $g(x,.)$ have been linearly extended for all x. We have defined a correspondence $\hat{F}$ from $\Delta_f(X)\times[0,1]$ to itself. It is easy to see that $\hat{F}$ always is an affine correspondence (see

lemma 2.3.26 later). In the following definition we consider the closure of the graph of $\hat{F}$ within the compact set $(\Delta(X) \times [0,1])^2$, with the weak topology.

Definition 2.3.18 *An element (u, y) in $\Delta(X) \times [0,1]$ is said to be an invariant couple for the MDP Ψ if $((u,y),(u,y)) \in cl(Graph(\hat{F}))$. The set of invariant couples of Ψ is denoted by RR.*

Our main result for standard MDPs is the following, where X is assumed to be a compact subset of a simplex $\Delta(K)$. Recall that $D_1 = \{f \in \mathcal{C}(\Delta(K)), \forall x, y \in \Delta(K), \forall a, b \geq 0,\ af(x) - bf(y) \leq \|ax - by\|_1\}$, and any f in D_1 is linearly extended to $\Delta(\Delta(K))$.

Theorem 2.3.19 *Let $\Psi = (X, A, q, g)$ be a standard MDP where X is a compact subset of a simplex $\Delta(K)$, such that:*

$$\forall x \in X, \forall y \in X, \forall a \in A, \forall f \in D_1, \forall \alpha \geq 0, \forall \beta \geq 0,$$

$$|\alpha f(q(x,a)) - \beta f(q(y,a))| \leq \|\alpha x - \beta y\|_1 \text{ and } |\alpha g(x,a) - \beta g(y,a)| \leq \|\alpha x - \beta y\|_1.$$

then Ψ has a general uniform value v^* characterized by: for all x in X,

$$\begin{aligned} v^*(x) \quad = \inf \quad & \Big\{ w(x), w : \Delta(X) \to [0,1] \text{ affine } C^0 \text{ s.t.} \\ & (1)\ \forall x' \in X, w(x') \geq \sup_{a \in A} w(q(x',a)) \text{ and } (2)\ \forall (u,y) \in RR, w(u) \geq y \Big\}. \end{aligned}$$

The proof of theorem 2.3.19 will be in section 2.3.4. An immediate corollary is when the state space is finite.

Corollary 2.3.20 *Consider a standard MDP (K, A, q, g) with a finite set of states K. Then it has a general uniform value v^*, and for each state k:*

$$\begin{aligned} v^*(k) \quad = \inf \quad & \Big\{ w(k), w : \Delta(K) \to [0,1] \text{ affine } \ s.t. \\ & (1)\ \forall k' \in K, w(k') \geq \sup_{a \in A} w(q(k',a)) \text{ and } (2) \forall (p,y) \in RR, w(p) \geq y \Big\}. \end{aligned}$$

with $RR = \{(p,y) \in \Delta(K) \times [0,1], ((p,y),(p,y)) \in \text{cl}(conv(Graph(F)))\}$ and $F(k,y) = \{(q(k,a), g(k,a)), a \in A\}$.

Proof: K is viewed as a subset of the simplex $\Delta(K)$, endowed with the L^1-norm. Fix k, k' in K, a in A, $\alpha \geq 0$ and $\beta \geq 0$. We have

$$\|\alpha k - \beta k'\| = \begin{cases} |\alpha - \beta| \text{ if } k = k', \\ \alpha + \beta \text{ otherwise.} \end{cases}$$

First,

$$|\alpha g(k,a) - \beta g(k',a)| \leq \begin{cases} |\alpha - \beta| g(k,a) \text{ if } k = k' \\ \alpha + \beta \text{ otherwise,} \end{cases}$$

so in all cases $|\alpha g(k,a) - \beta g(k',a)| \leq \|\alpha k - \beta k'\|$. Secondly, consider $f \in D_1$. f takes values in $[-1,1]$, so similarly we have: $|\alpha f(q(k,a)) - \beta f(q(k',a))| \leq \|\alpha k - \beta k'\|$. So we can apply theorem 2.3.19, and the graph of $\hat{F}$ is the convex hull of the graph of F. □

Remark 2.3.21 When the set of actions is finite, we are in the setting of Blackwell [Bla62] and the value is characterized by the Average Cost Optimality Equation. In fact in this setting, our characterization leads to a dual formulation of a result of Denardo and Fox [DF68]. Denardo and Fox [DF68] showed that the value v^* is the smallest (pointwise) excessive function for which there exists a vector $h \in \mathbb{R}^K$ such that (v^*, h) is superharmonic in the sense of Hordjik and Kallenberg [HK79] , i.e.

$$\forall k \in K,\ a \in A\ v^*(k) + h(k) \geq g(k,a) + \sum_{k'} q(k,a)(k')h(k'). \tag{2.2}$$

Given a function w the existence of a vector h such that (w,h) is superharmonic is a linear programming problem with $K \times A$ inequalities. By Farkas' lemma it has a solution if and only if a dual problem has no solution, and the dual programming problem is to find a solution $\pi \in \mathbb{R}^{K \times A}$ of the following system:

$$\begin{array}{rrl} \forall k \in K & \sum_{a' \in A} \pi(k,a') & = \sum_{k' \in K, a' \in A} \pi(k',a') q(k',a')(k) \\ \forall (k,a) \in K \times A & \pi(k,a) & \geq 0 \\ \forall k \in K & \sum_{a' \in A} \pi(k,a) g(k,a') & > v(k). \end{array}$$

If we denote by $p \in \mathbb{R}^K$ the marginal of π on K and define for all k such that $p^k > 0$, $\sigma(k) = \frac{\pi(k,a)}{p^k}$ and set $\sigma(k)$ to any probability otherwise, then σ is a strategy in the MDP. Moreover p is invariant under σ and the stage payoff y is greater than $v(p)$, thus the couple (p,y) is in RR and the condition (2) in corollary 2.3.20 is not satisfied. Reciprocally since the action state is compact, given $(p,y) \in RR$, there exists a strategy σ such that p is invariant under σ and the payoff is y. Therefore if the condition (2) is not true then there exists $h \in \mathbb{R}^k$ such that (w,h) is superharmonic. Note that Denardo and Fox state a dual of the minimization problem and obtain an explicit dual maximization problem whose solution is the value. Hordjik and Kallenberg exhibit from the solutions of this dual problem an optimal strategy.

2.3.3 Proof of the existence in Gambling Houses

In this section we consider a compact metric space (X,d), and we use the Kantorovich-Rubinstein distance $d = d_{KR}$ on $\Delta(X)$. We write $Z = \Delta_f(X)$, $\overline{Z} = \Delta(X)$. We start with a lemma.

Lemma 2.3.22 *Let $F : X \rightrightarrows \Delta_f(X)$ be non expansive for d_{KR}. Then the mixed extension of F is 1-Lipschitz from $\Delta_f(X)$ to $\Delta_f(X)$ for d_{KR}.*

Proof of lemma 2.3.22. We first show that the mapping $(p \mapsto \mathrm{conv}F(p))$ is non expansive from X to Z. Indeed, consider p and p' in X, and $u = \sum_{i \in I} \alpha_i u_i$, with I finite, $\alpha_i \geq 0$,

$u_i \in F(p)$ for each i, and $\sum_{i\in I} \alpha_i = 1$. By assumption for each i one can find u'_i in $F(p')$ such that $d_{KR}(u_i, u'_i) \leq d(p, p')$. Define $u' = \sum_{i\in I} \alpha_i u'_i$ in $\mathrm{conv}F(p')$. We have:

$$\begin{aligned} d_{KR}(u, u') &= \sup_{f\in E_1} \left(\sum_i \alpha_i u_i(f) - \sum_i \alpha_i u'_i(f) \right), \\ &= \sup_{f\in E_1} \sum_{i\in I} \alpha_i (u_i(f) - u'_i(f)), \\ &\leq \sum_{i\in I} \alpha_i \; d_{KR}(u_i, u'_i), \\ &\leq d(p, p'). \end{aligned}$$

We now prove that $\hat{F}$ is 1-Lipschitz from Z to Z. Let u_1, u_2 be in Z and $v_1 = \sum_{p\in X} u_1(p) f_1(p)$, where $f_1(p) \in \mathrm{conv}F(p)$ for each p. By the Kantorovich duality formula, there exists a coupling $\chi = (\chi(p, q))_{(p,q)\in X\times X}$ in $\Delta_f(X \times X)$ with first marginal u_1 and second marginal u_2 satisfying:

$$d_{KR}(u_1, u_2) = \sum_{(p,q)\in X\times X} \chi(p, q) d(p, q).$$

For each p, q in X by the first part of this proof there exists $f^p(q) \in \mathrm{conv}F(q)$ such that $d_{KR}(f^p(q), f_1(p)) \leq d(p, q)$. We define:

$$f_2(q) = \sum_{p\in X} \frac{\chi(p, q)}{u_2(q)} f^p(q) \in \mathrm{conv}F(q), \text{ and } v_2 = \sum_{q\in X} u_2(q) f_2(q) \in \hat{F}(u_2).$$

We now conclude.

$$\begin{aligned} d_{KR}(v_1, v_2) &= d_{KR} \left(\sum_{p\in X} u_1(p) f_1(p), \sum_{q\in X} u_2(q) f_2(q) \right) \\ &= d_{KR} \left(\sum_{p,q} \chi(p, q) f_1(p), \sum_{q,p} \chi(p, q) f^p(q) \right) \\ &\leq \sum_{p,q} \chi(p, q) d_{KR}(f_1(p), f^p(q)) \\ &\leq \sum_{p,q} \chi(p, q) d(p, q) = d_{KR}(u_1, u_2). \end{aligned}$$

The mixed extension of F is 1-Lipschitz. □

We now consider a Gambling House $\Gamma = (X, F, r)$ and assume the hypotheses of theorem 2.3.9 are satisfied. We will work [2] with the deterministic Gambling House $\hat{\Gamma} = (\Delta_f(X), \hat{F}, r)$.

2. A variant of the proof would be to consider the Gambling House on $\Delta(X)$ where the transition correspondence is defined so that its graph is the closure of the graph of $\hat{F}$. Part 1) of lemma 2.3.23 shows this correspondence is also non expansive.

Recall that r is extended to an affine and continuous mapping on $\Delta(X)$ whereas $\hat{F}$ is an affine non expansive correspondence from Z to Z.

For p in X, the pure plays in $\hat{\Gamma}$ at the initial state δ_p coincide with the mixed plays in Γ at the initial state p. As a consequence, the θ-value for Γ at p coincides with the θ-value for $\hat{\Gamma}$ at δ_p, which is written $v_\theta(p) = v_\theta(\delta_p)$. Because $\hat{F}$ and r are affine on Z, the θ-value for $\hat{\Gamma}$, as a function defined on Z, is the affine extension of the original v_θ defined on X. So we have a unique value function v_θ which is defined on Z and is affine. Because $\hat{F}$ is 1-Lipschitz and r is uniformly continuous, all the value functions v_θ have the same modulus of continuity as r, so $(v_\theta)_\theta$ is an equicontinuous family of mappings from Z to $[0,1]$. Consequently, we extend v_θ to an affine mapping on $\overline{Z}$ with the same modulous of continuity, and the family $(v_\theta)_\theta$ now is an equicontinuous [3] family of mappings from $\overline{Z}$ to $[0,1]$.

We define R and v^* as in the statements of theorem 2.3.9, so that for all x in X,

$$\begin{aligned} v^*(x) \;\; = \inf \;\; & \Big\{ w(x), w : \overline{Z} \to [0,1] \text{ affine } C^0 \text{ } s.t. \\ & (1)\ \forall y \in X, w(y) \geq \sup_{u \in F(y)} w(u) \text{ and } (2) \forall u \in R, w(u) \geq r(u) \Big\}. \end{aligned}$$

We start with a technical lemma using the non-expansiveness of $\hat{F}$.

Lemma 2.3.23 *1) Given (u,u') in $\mathrm{cl}(Graph(\hat{F}))$, v in Z and $\varepsilon > 0$, there exists $v' \in \hat{F}(v)$ such that $d(u',v') \leq d(u,v) + \varepsilon$.*

2) Given a sequence $(z_t)_{t\geq 0}$ of elements of $\overline{Z}$ such that $(z_t, z_{t+1}) \in \mathrm{cl}(Graph(\hat{F}))$ for all $t \geq 1$, for each ε one can find a sequence $(z'_t)_{t\geq 0}$ of elements of Z such that $(z'_t)_{t\geq 1}$ is a play at z'_0, and $d(z_t, z'_t) \leq \varepsilon$ for each $t \geq 0$.

Proof of lemma 2.3.23: 1) For all $\varepsilon > 0$ there exists $(z,z') \in Graph(\hat{F})$ such that $d(z,u) \leq \varepsilon$ and $d(z',u') \leq \varepsilon$. Because $\hat{F}$ is non expansive, one can find v' in $\hat{F}(v)$ such that $d(z',v') \leq d(z,v)$. Consequently, $d(v',u') \leq d(v',z') + d(z',u') \leq d(z,v) + \varepsilon \leq d(u,v) + 2\varepsilon$.

2) It is first easy to construct (z'_0, z'_1) in the graph of $\hat{F}$ such that $d(z'_0, z_0) \leq \varepsilon$ and $d(z'_1, z_1) \leq \varepsilon$. $(z_1, z_2) \in \mathrm{cl}(Graph(\hat{F}))$ so by 1) one can find (z'_2) in $\hat{F}(z'_1)$ such that $d(z_2, z'_2) \leq d(z_1, z'_1) + \varepsilon^2 \leq \varepsilon + \varepsilon^2$. Iterating, we construct a play $(z'_t)_{t\geq 1}$ at z'_0 such that $d(z_t, z'_t) \leq \varepsilon + \varepsilon^2 + ... + \varepsilon^t$ for each t.

Proposition 2.3.24 *Γ has a general limit value given by v^*.*

Proof of proposition 2.3.24: By Ascoli's theorem, it is enough to show that any limit point of $(v_\theta)_\theta$ (for the uniform convergence) coincides with v^*. We thus assume that $(v_{\theta^k})_k$ uniformly converges to v on $\overline{Z}$ when k goes to ∞, for a family of evaluations satisfying:

$$\sum_{t\geq 1} |\theta^k_{t+1} - \theta^k_t| \longrightarrow_{k\to\infty} 0.$$

3. Z being precompact, this is enough to obtain the existence of a general limit value, see Renault 2012b. Here we will moreover obtain a characterization of this value and the existence of the general uniform value.

And we need to show that $v = v^*$.

A) We first show that $v \geq v^*$.

It is plain that v can be extended to an affine function on $\overline{Z}$ and has the same modulus of continuity of r. Because $\sum_{t\geq 1} |\theta^k_{t+1} - \theta^k_t| \longrightarrow_{k\to\infty} 0$, we have by equation (2.1) of section 2.3.1 that: $\forall y \in X, v(y) = \sup_{u\in F(y)} v(u)$.

Let now u be in R. By lemma 2.3.23 for each ε one can find u_0 in Z and a play $(u_1, u_2, ..., u_t, ...)$ such that $u_t \in \hat{F}(u_{t-1})$ and $d(u, u_t) \leq \varepsilon$ for all $t \geq 0$. Because r is uniformly continuous, we get $v(u) \geq r(u)$.

By definition of v^* as an infimum, we obtain: $v^* \leq v$.

B) We show that $v^* \geq v$. Let w be a continous affine mapping from $\overline{Z}$ to $[0, 1]$ satisfying (1) and (2) of the definition of v^*. It is enough to show that $w(x) \geq v(x)$ for each x in X. Fix x in X and $\varepsilon > 0$.

For each k, let $\sigma^k = (u^k_1, ..., u^k_t, ...) \in Z^\infty$ be a play at δ_x for $\hat{\Gamma}$ which is almost optimal for the θ^k-value, in the sense that $\sum_{t\geq 1} \theta^k_t r(u^k_t) \geq v_{\theta^k}(x) - \varepsilon$. Define:

$$u(k) = \sum_{t=1}^{\infty} \theta^k_t u^k_t \in \overline{Z}, \text{ and } u'(k) = \sum_{t=1}^{\infty} \theta^k_t u^k_{t+1} \in \overline{Z}.$$

$u(k)$ and $u'(k)$ are well-defined limits of normal convergent series in the Banach space $\mathcal{C}(X)'$. Because $\hat{F}$ is affine, its graph is a convex set and $(u(k), u'(k)) \in \mathrm{cl}(Graph(\hat{F}))$ for each k.

Moreover, we have $d(u(k), u'(k)) \leq \mathrm{diam}(X)(\theta^k_1 + \sum_{t=2}^{\infty} |\theta^k_t - \theta^k_{t-1}|)$, where $\mathrm{diam}(X)$ is the diameter of X. Consequently, $\sum_{t\geq 1} |\theta^k_{t+1} - \theta^k_t| \longrightarrow_{k\to\infty} 0$ implies $d(u(k), u'(k)) \longrightarrow_{k\to\infty} 0$. Considering a limit point of the sequence $(u(k), u'(k))_k$, we obtain some u in R. By assumption on w, $w(u) \geq r(u)$. Moreover, for each k we have $r(u(k)) = \sum_{t\geq 1} \theta^k_t r(u^k_t) \geq v_{\theta^k}(x) - \varepsilon$, so $r(u) \geq v(x) - \varepsilon$.

Because w is excessive, we obtain that for each k the sequence $(w(u^k_t))_t$ is non increasing, so $w(u(k)) = \sum_{t\geq 1} \theta^k_t w(u^k_t) \leq w(x)$. So we obtain:

$$w(x) \geq w(u) \geq r(u) \geq v(x) - \varepsilon.$$

This is true for all ε, so $w \geq v$. □

Proposition 2.3.25 *Γ has a general uniform value.*

Proof of proposition 2.3.25: First we can extend the notion of mixed play to Z. A mixed play at $u_0 \in Z$, is a sequence $\sigma = (u_1, ..., u_t, ...) \in Z^\infty$ such that $u_{t+1} \in \hat{F}(u_t)$ for each $t \geq 0$, and we denote by $\Sigma(u_0)$ the set of mixed play at u_0. Given t, T in $\mathbb{N}$, $n \in \mathbb{N}^*$ and $u_0 \in Z$, we define for each mixed play $\sigma = (u_t)_{t\geq 1} \in \Sigma(u_0)$ the auxiliary payoff:

$$\gamma_{t,n}(\sigma) = \frac{1}{n} \sum_{l=t+1}^{t+n} r(u_l), \text{ and } \beta_{T,n}(\sigma) = \inf_{t\in\{0,...,T\}} \gamma_{t,n}(\sigma).$$

And we also define the auxiliary value function: for all u in Z,

$$h_{T,n}(u_0) = \sup_{\sigma\in\Sigma(u_0)} \beta_{T,n}(\sigma).$$

Clearly, $\beta_{T,n}(\sigma) \leq \gamma_{0,n}(\sigma)$ and $h_{T,n}(u_0) \leq v_n(u_0)$. We can write:

$$\begin{aligned} h_{T,n}(u_0) &= \sup_{\sigma\in\Sigma(u_0)} \inf_{\theta\in\Delta(\{0,...,T\})} \frac{1}{n}\sum_{t=0}^{T}\theta_t \sum_{l=t+1}^{t+n} r(u_l) \\ &= \sup_{\sigma\in\Sigma(u_0)} \inf_{\theta\in\Delta(\{0,...,T\})} \sum_{l=1}^{T+n} \beta_l(\theta,n) r(u_l). \end{aligned}$$

where for each l in $1, ..., T+n$,

$$\beta_l(\theta,n) = \frac{1}{n}\sum_{t=Max\{0,l-n\}}^{\text{Min}\{T,l-1\}} \theta_t.$$

By construction, $\hat{F}$ is affine, so $\Sigma(u_0)$ is a convex subset of Z^∞. $\Delta(\{0,...,T\})$ is convex compact and the payoff $\sum_{l=1}^{T+n}\beta_l(\theta,n)r(u_l)$ is affine both in θ and in σ. We can apply a standard minmax theorem to get:

$$h_{T,n}(u_0) = \inf_{\theta\in\Delta(\{0,...,T\})} \sup_{\sigma\in\Sigma(u_0)} \sum_{l=1}^{T+n} \beta_l(\theta,n) r(u_l).$$

We write $\theta_t = 0$ for $t > T$ and for each $l \geq 0$: $\beta_l(n,\theta) = \frac{1}{n}(\theta_0 + ... + \theta_{l-1})$ if $l \leq n$, $\beta_l(\theta,n) = \frac{1}{n}(\theta_{l-n} + ... + \theta_{l-1})$ if $n+1 \leq l \leq n+T$, $\beta_l(n,\theta) = 0$ if $l > n+T$. The evaluation $\beta(\theta,n)$ is a particular probability on stages and $h_{T,n}(u_0) = \inf_{\theta\in\Delta(\{0,...,T\})} v_{\beta(\theta,n)}(u_0)$. It is easy to bound the irregularity of $\beta(\theta,n)$:

$$\sum_{l\geq 0} |\beta_{l+1}(\theta,n) - \beta_l(\theta,n)| = \sum_{l=0}^{n-1}\frac{\theta_l}{n} + \sum_{l\geq n}\frac{1}{n}|\theta_l - \theta_{l-n}| \leq \frac{3}{n} \longrightarrow_{n\to\infty} 0.$$

The irregularity of $\beta(\theta,n)$ goes to zero as n goes to infinity, uniformly in θ. So we can use the previous proposition 2.3.24 to get:

$$\forall\varepsilon > 0, \exists n_0, \forall n \geq n_0, \forall\theta\in\Delta(\mathbb{N}), \forall u_0 \in Z,\ |v_{\beta(\theta,n)}(u_0) - v^*(u_0)| \leq \varepsilon.$$

This implies that $h_{\infty,n}(u_0) :=_{def} \inf_{\theta\in\Delta(\mathbb{N})} v_{\beta(\theta,n)}(u_0) = \inf_{T\geq 0} h_{T,n}(u_0)$ converges to $v^*(u_0)$ when $n \to \infty$, and the convergence is uniform over Z. Consequently, if we fix $\varepsilon > 0$ there exists n_0 such that for all u_0 in Z, for all $T \geq 0$, there exists a play $\sigma^T = (u_t^T)_{t\geq 1}$ in $\Sigma(u_0)$ such that the average payoff is good on every interval of n_0 stages starting before $T+1$: for all $t = 0, ..., T$, $\gamma_{t,n_0}(\sigma^T) \geq v^*(u_0) - \varepsilon$.

We fix u_0 in Z and consider, for each T, the play $\sigma^T = (u_t^T)_{t\geq 1}$ in $\Sigma(u)$ as above. By a diagonal argument we can construct for each $t \geq 1$ a limit point z_t in $\overline{Z}$ of the sequence

$(u_t^T)_{T\geq 0}$ such that for each t we have $(z_t, z_{t+1}) \in \mathrm{cl}(Graph(\hat{F}))$, with $z_0 = u_0$. For each $m \geq 0$, we have $\frac{1}{n_0}\sum_{t=m+1}^{m+n_0} r(u_t^T) \geq v^*(u_0) - \varepsilon$ for T large enough, so at the limit we get: $\frac{1}{n_0}\sum_{t=m+1}^{m+n_0} r(z_t) \geq v^*(u_0) - \varepsilon$.

r being uniformly continuous, there exists η such that $|r(z) - r(z')| \leq \varepsilon$ as soon as $d(z, z') \leq \eta$. By lemma 2.3.23, one can find a $\sigma' = (z'_1, ..., z'_t, ...)$ at $\Sigma(z_0)$ such that for each t, $d(z_t, z'_t) \leq \eta$. We obtain that for each $m \geq 0$, $\frac{1}{n_0}\sum_{t=m+1}^{m+n_0} r(z'_t) \geq v^*(u) - 2\varepsilon$.

Consequently we have proved: $\forall \varepsilon > 0$, there exists n_0 such that for each initial state x in X, there exists a mixed play $\sigma' = (z'_t)_t$ at x such that: $\forall m \geq 0$, $\frac{1}{n_0}\sum_{t=m+1}^{m+n_0} r(z'_t) \geq v^*(x) - 2\varepsilon$. Let $\theta \in \Delta(\mathbb{N}^*)$ be an evaluation, it is now easy to conclude. First if $v^*(x) - 2\epsilon < 0$, then any play is 2ϵ-optimal. Otherwise, for each $j \geq 1$, denote by $\overline{\theta_j}$ the maximum of θ on the block $B^j = \{(j-1)n_0 + 1, ..., jn_0\}$. For all $t \in B^j$, we have:

$$\overline{\theta_j} \geq \theta_t \geq \overline{\theta_j} - \sum_{t' \in \{(j-1)n_0+1, ... jn_0 - 1\}} |\theta_{t'+1} - \theta_{t'}|.$$

As a consequence, for all j we have:

$$\begin{aligned}\sum_{t=(j-1)n_0+1}^{jn_0} \theta_t r(z'_t) &\geq \overline{\theta_j} \sum_{t=(j-1)n_0+1}^{jn_0} r(z'_t) \;-\; n_0 \sum_{t' \in \{(j-1)n_0+1, ..., jn_0-1\}} |\theta_{t'+1} - \theta_{t'}| \\ &\geq \sum_{t=(j-1)n_0+1}^{jn_0} \theta_t (v^*(x) - 2\varepsilon) \;-\; n_0 \sum_{t' \in \{(j-1)n_0+1, ..., jn_0-1\}} |\theta_{t'+1} - \theta_{t'}|\end{aligned}$$

and by summing over j, we get: $\gamma_\theta(x, \sigma') \geq v^*(x) - 2\epsilon - n_0 I(\theta) \geq v^*(x) - 3\epsilon$ as soon as $I(\theta)$ is small enough. □

2.3.4 Proof of the existence in MDPs

Assume that X is a compact subset of a simplex $\Delta(K)$, and let $\Psi = (X, A, q, g)$ be a standard MDP such that: $\forall x \in X, \forall y \in X, \forall a \in A, \forall f \in D_1, \forall \alpha \geq 0, \forall \beta \geq 0$,

$$|\alpha f(q(x,a)) - \beta f(q(y,a))| \leq \|\alpha x - \beta y\|_1 \text{ and } |\alpha g(x,a) - \beta g(y,a)| \leq \|\alpha x - \beta y\|_1.$$

We write $Z = \Delta_f(X) \times [0,1]$, and $\overline{Z} = \Delta(X) \times [0,1]$. We will use the metric $d_* = d_0 = d_1 = d_2 = d_3$ on $\Delta(\Delta(K))$ introduced in section 2.2.3 and its restriction to $\Delta(X)$, so that $\overline{Z}$ is a compact metric space. For all $(u,y), (u',y') \in \Delta_f(X) \times [0,1]$, we put $d((u,y),(u',y')) = \max(d_*(u,u'), |y - y'|)$ so that (Z,d) is a precompact metric space. Recall we have defined the correspondence $\hat{F}$ from Z to itself such that for all (u,y) in Z,

$$\hat{F}(u,y) = \{(Q(u,\sigma), G(u,\sigma)) \; s.t. \; \sigma : X \to \Delta_f(A)\},$$

with the notations $Q(u,\sigma) = \sum_{x\in X} u(x)q(x,\sigma(x))$ and $G(u,\sigma) = \sum_{x\in X} u(x)g(x,\sigma(x))$. And we simply define the payoff function r from Z to $[0,1]$ by $r(u,y) = y$ for all (u,y) in Z. We start with a crucial lemma, which shows the importance of the duality formula of theorem 2.2.18.

Lemma 2.3.26 *$\hat{F}$ is an affine and non expansive correspondence from Z to itself.*

Proof of lemma 2.3.26. We first show that: $\forall u,u' \in \Delta_f(X)$, $\forall \alpha \in [0,1]$, $\forall y,y' \in [0,1]$, $\hat{F}(\alpha u + (1-\alpha)u', \alpha y + (1-\alpha)y') = \alpha\hat{F}(u,y) + (1-\alpha)\hat{F}(u',y')$. First the transition does not depend on the second coordinate so we can forget it for the rest of the proof. The $\subset$ part is clear. To see the reverse inclusion, consider $\sigma : X \to \Delta_f(A)$, $\sigma' : X \to \Delta_f(A)$ and $v = \alpha\sum_{x\in X} u(x)q(x,\sigma(x)) + (1-\alpha)\sum_{x\in X} u'(x)q(x,\sigma'(x))$ in $\alpha\hat{F}(u) + (1-\alpha)\hat{F}(u')$. Define

$$\sigma^*(x) = \frac{\alpha u(x)\sigma(x) + (1-\alpha)u'(x)\sigma'(x)}{\alpha u(x) + (1-\alpha)u'(x)},$$

for each x such that the denominator is positive. Then $v = \sum_{x\in X}(\alpha u + (1-\alpha)u'(x))q(x,\sigma^*(x))$, and $\hat{F}$ is affine.

We now prove that $\hat{F}$ is non expansive. Let $z = (u,y)$ and $z' = (u',y')$ be in Z. We have $d((u,y),(u',y')) \geq d_*(u,u')$ and denote by U and U' the respective supports of u and u'. By the duality formula of theorem 2.2.18, there exists $\alpha = (\alpha(p,p'))_{(p,p')\in U\times U'}$ and $\beta = (\beta(p,p'))_{(p,p')\in U\times U'}$ with non-negative coordinates satisfying: $\sum_{p'\in U'}\alpha(p,p') = u(p)$ for all $p \in U$, $\sum_{p\in U}\beta(p,p') = u'(p')$ for all $p' \in U'$, and

$$d_*(u,u') = \sum_{(p,p')\in U\times U'} \|p\,\alpha(p,p') - p'\,\beta(p,p')\|_1.$$

Consider now $v = Q(u,\sigma) = \sum_{p\in U} u(p)q(p,\sigma(p))$ for some $\sigma : X \to \Delta_f(A)$. We define for all p' in U':

$$\sigma'(p') = \sum_{p\in U} \frac{\beta(p,p')}{u'(p')}\sigma(p),$$

and $v' = Q(u',\sigma') = \sum_{p'\in U'} u'(p')q(p',\sigma'(p'))$. Then $v' \in \hat{F}(u',y')$, and for each test function φ in D_1 we have:

$$\begin{aligned}
|\varphi(v) - \varphi(v')| &= |\sum_{p,p'} \alpha(p,p')\varphi(q(p,\sigma(p))) - \beta(p,p')\varphi(q(p',\sigma(p)))| \\
&= |\sum_{p,p',a} \alpha(p,p')\sigma(p)(a)\varphi(q(p,a)) - \beta(p,p')\sigma(p)(a)\varphi(q(p',a))| \\
&\leq \sum_{p,p'} \|\alpha(p,p')p - \beta(p,p')p'\|_1 = d_*(u,u'),
\end{aligned}$$

and therefore $d_*(v,v') \leq d_*(u,u')$. In addition we have a similar result on the payoff,

$$\begin{aligned}|G(u,\sigma) - G(u',\sigma')| &= |\sum_{p,p'} \alpha(p,p')g(p,\sigma(p)) - \beta(p,p')g(p',\sigma(p))| \\ &\leq \sum_{p,p'} \|\alpha(p,p')p - \beta(p,p')p'\|_1 \\ &\leq d_*(u,u').\end{aligned}$$

Thus we have $d((Q(u,\sigma),R(u,\sigma)),(Q(u',\sigma'),R(u',\sigma'))) \leq d_*(u,u') \leq d(z,z')$. □

Recall that the set of invariant couples of the MDP Ψ is:

$$RR = \{(u,y) \in \overline{Z}, ((u,y),(u,y)) \in cl(Graph(\hat{F}))\},$$

and the function $v^* : X \longrightarrow \mathbb{R}$ is defined by:

$$\begin{aligned}v^*(x) \quad = \inf \quad &\Big\{w(x), w : \Delta(X) \to [0,1] \text{ affine } C^0 \ s.t. \\ &(1)\ \forall y \in X, w(y) \geq \sup_{a\in A} w(q(y,a)) \text{ and } (2)\ \forall (u,y) \in RR, w(u) \geq y \Big\}.\end{aligned}$$

We now consider the deterministic Gambling House $\hat{\Gamma} = (Z, \hat{F}, r)$. Z is precompact metric, $\hat{F}$ is affine non expansive and r is obviously affine and uniformly continuous. Given an evaluation θ, the value of $\hat{\Gamma}_\theta$ at $z_0 = (u,y)$ is denoted by $\hat{v}_\theta(u,y) = \hat{v}_\theta(u)$ and does not depend on y. The recursive formula of section 2.3.1 yields:

$$\begin{aligned}\forall (u,y) \in Z, \quad \hat{v}_\theta(u) \quad &= \quad \sup_{(u',y') \in \hat{F}(u)} \theta_1 y' + (1-\theta_1)\hat{v}_{\theta^+}(u') \\ &= \quad \sup_{\sigma \in X \to \Delta_f(A)} \left(\theta_1 G(u,\sigma) + (1-\theta_1)\hat{v}_{\theta^+}(Q(u,\sigma))\right).\end{aligned}$$

Because $\hat{F}$ and r are affine, $\hat{v}_\theta$ is affine in u and the supremum in the above expression can be taken over the function from X to A. Because $\hat{F}$ is non expansive and r is 1-Lipschitz, each $\hat{v}_\theta$ is 1-Lipschitz.

We denote by v_θ the θ-value of the MDP Ψ and linearly extend it to $\Delta_f(X)$. It turns out that the recursive formula satisfied by v_θ is similar to the above recursive formula for $\hat{v}_\theta$, so that $v_\theta(u) = \hat{v}_\theta(u,y)$ for all u in $\Delta_f(X)$ and y in $[0,1]$. As a consequence, the existence of the general limit value in both problems $\hat{\Gamma}$ and Ψ is equivalent. Moreover, a deterministic play in $\hat{\Gamma}$ induces a strategy in Ψ, so that the existence of a general uniform value in $\hat{\Gamma}$ will imply the existence of the general uniform value in Ψ (note that deterministic and mixed plays in $\hat{\Gamma}$ are equivalent since $\hat{F}$ has convex values).

It is thus sufficient to show that $\hat{\Gamma}$ has a general uniform value given by v^*, and we can mimic the end of the proof of theorem 2.3.9. Lemma 2.3.23 applies word for word. Finally, one can proceed almost exactly as in propositions 2.3.24 and 2.3.25 to show that $\hat{\Gamma}$, hence Ψ, has a

general uniform value given by v^*.

2.4 Applications to partial observation and games

2.4.1 MDPs with partial observation and finitely many states

We now consider a more general model of MDP with actions where after each stage, the decision maker does not perfectly observe the state. A MDP with partial observation, or POMDP, $\Gamma = (K, A, S, q, g)$ is given by a finite set of states K, a non empty set of actions A and a non empty set of signals S. The transition q now goes from $K \times A$ to $\Delta_f(S \times K)$ (by assumption the support of the signals at each state is finite) and the payoff function g still goes from $K \times A$ to $[0, 1]$. Given an initial probability p on K, the POMDP $\Gamma(p)$ is played as following. An initial state k_1 in K is selected according to p and is not told to the decision maker. At every stage $t \geq 1$ he selects an action $a_t \in A$. He has a (unobserved) payoff $g(k_t, a_t)$ and a pair (k_{t+1}, s_t) is drawn according to $q(k_t, a_t)$. The player learns s_t, and the play proceeds to stage $t + 1$ with the new state k_{t+1}. A behavioral strategy is now a sequence $(\sigma_t)_{t\geq 1}$ of applications with for each t, $\sigma_t : (A \times S)^{t-1} \to \Delta_f(A)$. As usual, an initial probability on K and a behavior strategy σ induce a probability distribution over $(K \times A \times S)^\infty$ and we can define the θ-values and the notions of general limit and uniform values accordingly.

Theorem 2.4.1 *A POMDP with finitely many states has a general uniform value, i.e. there exists $v^* : \Delta(K) \to \mathbb{R}$ with the following property: for each $\varepsilon > 0$ one can find $\alpha > 0$ and for each initial probability p a behavior strategy $\sigma(p)$ such that for each evaluation θ with $I(\theta) \leq \alpha$,*

$$\forall p \in \Delta(K), |v_\theta(p) - v^*(p)| \leq \varepsilon \text{ and } \gamma_\theta(\sigma(p)) \geq v^*(p) - \varepsilon.$$

Proof: We introduce Ψ an auxiliary MDP on $X = \Delta(K)$ with the same set of actions A and the following payoff and transition functions:

- $\widetilde{g} : X \times A \longrightarrow [0, 1]$ such that $\widetilde{g}(p, a) = \sum_{k\in K} p^k g(k, a)$ for all p in X and $a \in A$,
- $\widetilde{q} : X \times A \to \Delta_f(X)$ such that

$$\widetilde{q}(p, a) = \sum_{s\in S} \left(\sum_k p^k q(k, a)(s) \right) \delta_{\hat{q}(p,a|s)},$$

where $\hat{q}(p, a|s) \in \Delta(K)$ is the belief on the new state after playing a at p and observing the signal s:

$$\forall k' \in K, \hat{q}(p, a|s)(k') = \frac{q(p, a)(k', s)}{q(p, a)(s)} = \frac{\sum_k p^k q(k, a)(k', s)}{\sum_k p^k q(k, a)(s)}.$$

The POMDP $\Gamma(p_1)$ and the standard MDP $\Psi(p_1)$ have the same value for all θ-evaluations. And for each strategy σ in $\Psi(p_1)$, the player can guarantee the same payoff in the original game

$\Gamma(p_1)$ by mimicking the strategy σ. So if we prove that Ψ has a general uniform value it will imply that the POMDP Γ has a general uniform value.

To conclude the proof, we will simply apply theorem 2.3.19 to the MDP Ψ. We need to check the assumptions on the payoff and on the transition.

Consider any p, p' in X, $a \in A$, $\alpha \geq 0$ and $\beta \geq 0$. We have:

$$|\alpha\widetilde{g}(p,a) - \beta\widetilde{g}(p',a)| = |\sum_k (\alpha p^k - \beta p'^k) g(k,a)| \leq \|\alpha p - \beta p'\|_1$$

Moreover for any $f \in D_1$, we have:

$$\begin{aligned}
|\alpha\widetilde{q}(p,a)(f) - \beta\widetilde{q}(p',a)(f)| &= |\sum_{s\in S} (\alpha q(p,a)(s) f(\hat{q}(p,a,s)) - \beta q(p',a)(s) f(\hat{q}(p',a,s)))| \\
&\leq \sum_s \|\alpha q(p,a)(.,s) - \beta q(p',a)(.,s)\|_1 \\
&\leq \sum_{s,k,k'} |\alpha p^{k'} q(k',a)(k,s) - \beta p'^{k'} q(k',a)(k,s)| \\
&\leq \sum_{s,k,k'} q(k',a)(k,s) |\alpha p^{k'} - \beta p'^{k'}| = \|\alpha p - \beta p'\|_1.
\end{aligned}$$

where the first inequality comes from the definition of D_1.

By theorem 2.3.19, the MDP Ψ has a general uniform value and we deduce that the POMDP Γ has a general uniform value. □

Example 2.4.2 Let $\Gamma = (K, A, S, q, g, p)$ be a POMDP where $K = \{k_1, k_2\}$, $A = \{a, b\}$, $S = \{s\}$ and $p = \delta_{k_1}$. The initial state is k_1 and since there is only one signal, the decision maker will obtain no additional information on the state. We say that he is in the dark. The payoff is given by $g(0,a) = g(0,b) = g(1,b) = 0$ and $g(1,a) = 1$, and the transition by $q(1,a) = q(1,b) = \delta_{1,s}$, $q(0,a) = \delta_{0,s}$ and $q(0,b) = \frac{1}{2}\delta_{0,s} + \frac{1}{2}\delta_{1,s}$. On one hand if the decision maker plays a then the state stays the same and he receives a payoff of 1 if and only if the state is 1, on the other hand if he plays b then he receives a payoff of 0 but the probability to be in state 1 increases.

We define the function $\widetilde{g}$ from $X = \Delta(K)$ to $[0,1]$ by $\widetilde{g}((p,1-p),a) = 1-p$ and $\widetilde{g}((p,1-p),b) = 0$ for all $p \in [0,1]$, and the function $\widetilde{q}$ from X to $\Delta_f(X)$ by

$$\widetilde{q}((p,1-p),a) = \delta_{(p,1-p)} \text{ and } \widetilde{q}((p,1-p),b) = \delta_{(p/2,1-p/2)}.$$

Then the standard MDP $\Psi = (\Delta(K), A, \widetilde{g}, \widetilde{q})$ is the MDP associated in the previous proof to Γ. This MDP is deterministic since the decision maker is in the dark.

In this example, the existence of a general uniform value is immediate. If we fix $n \in \mathbb{N}$, the strategy $\sigma = b^n a^\infty$ which plays n times b and then a for the rest of the game, guarantees a stage payoff of $(1 - \frac{1}{2^n})$ from stage $n+1$ on, so the game has a general uniform value equal to 1. Finally if we consider the discounted evaluations, one can show that the speed of convergence

of v_λ is slower than λ :

$$v_\lambda(p_1) = 1 - \frac{\ln(\lambda)}{\ln(2)}\lambda + O(\lambda).$$

All the spaces are finite but the partial observation implies that the speed of convergence is slower than λ contrary to the perfect observation case where it is well known that the convergence is in $O(\lambda)$.

Remark 2.4.3 It is unknown if the uniform value exists in pure strategies, i.e. if the behavior strategies $\sigma(p)$ of theorem 2.4.1 can be chosen with values in A. This was already an open problem for the Cesàro-uniform value (see Rosenberg *et al.* [RSV02] and Renault [Ren11] for different proofs requiring the use of behavioral strategies). In our proof, there are two related places where the use of lotteries on actions is important. First in the proof of the convergence of the function $h_{T,n}$ (within the proof of theorem 2.3.9), we used Sion's theorem in order to inverse a supremum and an infimum so we need the convexity of the set of strategies. Secondly when we prove that the extended transition is 1-Lipschitz (see lemma 2.3.26), the coupling between the two distributions u and u' introduces some randomization.

2.4.2 Zero-sum repeated games with an informed controller

We finally consider zero-sum repeated games with an informed controller. We start with a general model $\Gamma = (K, I, J, C, D, q, g)$ of a zero-sum repeated game, where we have 5 non empty finite sets: a set of states K, two sets of actions I and J and two sets of signals C and D, and we also have a transition mapping q from $K \times I \times J$ to $\Delta(K \times C \times D)$ and a payoff function g from $K \times I \times J$ to $[0, 1]$. Given an initial probability π on $\Delta_f(K \times C \times D)$, the game $\Gamma(\pi) = \Gamma(K, I, J, C, D, q, g, \pi)$ is played as follows: at stage 1, a triple (k_1, c', d') is drawn according to π, player 1 learns c' and player 2 learns d'. Then simultaneously player 1 chooses an action i_1 in I and player 2 chooses an action j_1 in J. Player 1 gets a (unobserved) payoff $g(k_1, i_1, j_1)$ and player 2 the opposite. Then a new triple (k_2, c_1, d_1) is drawn accordingly to $q(k_1, i_1, j_1)$. Player 1 observes c_1, player 2 observes d_1 and the game proceeds to the next stage, etc...

A (behavioral) strategy for player 1 is a sequence $\sigma = (\sigma_t)_{t\geq 1}$ where for each $t \geq 1$, σ_t is a mapping from $C \times (I \times C)^{t-1}$ to $\Delta(I)$. Similarly a strategy for player 2 is a sequence of mappings $\tau = (\tau_t)_{t\geq 1}$ where for each $t \geq 1$,τ_t is a mapping from $D \times (J \times D)^{t-1}$ to $\Delta(J)$. We denote respectively by Σ and $\mathcal{T}$ the set of strategies of player 1 and player 2. An initial distribution π and a couple of strategies (σ, τ) defines for each t a probability on the possible histories up to stage t. And by Kolmogorov extension theorem, it can be uniquely extended to a probability on the set of infinite histories $(K \times C \times D \times I \times J)^{+\infty}$.

Given θ an evaluation function, we define the θ-payoff of (σ, τ) in $\Gamma(\pi)$ as the expectation

under $I\!P_{\pi,\sigma,\tau}$ of the payoff function,

$$\gamma_\theta(\pi,\sigma,\tau) = I\!E_{\pi,\sigma,\tau}\left(\sum_{t\geq 1}\theta_t\ g(k_t,i_t,j_t)\right).$$

By Sion's theorem the game $\gamma_\theta(\pi)$ has a value:

$$v_\theta(\pi) = \max_{\sigma\in\Sigma}\min_{\tau\in\mathcal{T}}\gamma_\theta(\pi,\sigma,\tau) = \min_{\tau\in\mathcal{T}}\max_{\sigma\in\Sigma}\gamma_\theta(\pi,\sigma,\tau),$$

and we can define the general limit value as in the MDP framework. Note that we do not ask the convergence to be uniform for all π in $\Delta(K\times C\times D)$, because we will later make some assumptions on the initial distribution.

Definition 2.4.4 *The repeated game $\Gamma(\pi) = (K,I,J,C,D,q,g,\pi)$ has a general limit value $v^*(\pi)$ if $v_\theta(\pi)$ converges to $v^*(\pi)$ when $I(\theta)$ goes to zero, i.e.:*

$$\forall\varepsilon>0,\exists\alpha>0,\forall\theta,\ \ (\ I(\theta)\leq\alpha\implies(|v_\theta(\pi)-v^*(\pi)|\leq\varepsilon)\ ).$$

And we can define a general uniform value by symmetrizing the definition for MDP.

Definition 2.4.5 *The repeated game $\Gamma(\pi)$ has a general uniform value if it has a general limit value v^* and for each $\varepsilon>0$ one can find $\alpha>0$ and a couple of strategies σ^* and τ^* such that for all evaluations θ with $I(\theta)\leq\alpha$:*

$$\forall\tau\in\mathcal{T},\gamma_\theta(\pi,\sigma^*,\tau)\geq v^*(\pi)-\varepsilon\ \text{ and }\ \forall\sigma\in\Sigma,\gamma_\theta(\pi,\sigma,\tau^*)\leq v^*(\pi)+\varepsilon.$$

We now focus on the case of a repeated game with an informed controller. We follow the definitions introduced in Renault [Ren12b]. The first one concerns the information of the first player. We assume that he is always fully informed of the state and of the signal of the second player:

Assumption 2.4.6 *There exist two mappings $\widetilde{k}: C\to K$ and $\widetilde{d}: C\to D$ such that, if E denotes $\{(k,c,d)\in K\times C\times D,\ \widetilde{k}(c)=k,\ \widetilde{d}(c)=d\}$, we have: $\forall(k,i,j)\in K\times I\times J$, $q(k,i,j)(E)=1$, and $\pi(E)=1$.*

Moreover we will assume that only player 1 has a meaningful influence on the transitions, in the following sense.

Assumption 2.4.7 *The marginal of the transition on $K\times D$ is not influenced by player 2's action. For k in K, i in I and j in J, we denote by $\bar{q}(k,i)$ the marginal of $q(k,i,j)$ on $K\times D$.*

The second player may influence the signal of the first player but he can not prevent him neither to learn the state nor to learn his own signal. Moreover he can not influence his own information, thus he has no influence on his beliefs about the state or about the beliefs of player 1 about his beliefs. A repeated game satisfying assumptions 2.4.6 and 2.4.7 is called a repeated game with an informed controller. It was proved in Renault [Ren12b] that for such games the Cesàro-uniform value exists and we will generalize it here to the general uniform value.

Example 2.4.8 We consider Γ a zero-sum repeated game with incomplete information as studied by Aumann and Maschler (see reference from 1995). It is defined by a finite family $(G^k)_{k\in K}$ of payoff matrices in $[0,1]^{I\times J}$ and $p \in \Delta(K)$ an initial probability. At the first stage, some state k is selected according to p and told to player 1 only. The second player knows the initial distribution p but not the realization. Then the matrix game G^k is repeated over and over. At each stage the players observe past actions but not their payoff. Formally it is a zero-sum repeated game $\Gamma = (K, I, J, C, D, q, g)$ as defined previously, with $C = K \cup (I \times J)$ and $D = \{d\} \cup I \times J$, and for all $(k, i, j) \in K \times I \times J$, $g(k, i, j) = G^k(i, j)$ and $q(k, i, j) = \delta_{k,(i,j),(i,j)}$. For all $p \in \Delta(K)$, we denote by $\Gamma(p)$ the game where the initial probability $\pi \in \Delta(K \times C \times D)$ is given by $\pi = \sum_{k\in K} p^k \delta_{k,k,d}$.

For each n, we denote by $v_n(p)$ the value of the n-stage game with initial probability p, where the payoff is the expected mean average of the n first stages. It is known that it satisfies the following dynamic programming formula:

$$v_n(p) = \sup_{a\in\Delta(I)^K} \left(\frac{1}{n}\widetilde{g}(p,a) + \frac{n-1}{n} \sum_{k\in K, i\in I} p^k a^k(i) v_{n-1}(\hat{q}(p,a|i)) \right).$$

where $p \in \Delta(K)$, $\widetilde{g}(p,a) = \min_j(\sum_k p^k G^k(a^k, j))$ and $\hat{q}(p,a|i)$ is the conditional belief on $\Delta(K)$ given p, a, i:

$$\hat{q}(p,a|i)(k') = \frac{\sum_k p^k a^k(i) q(k,i)(k')}{\sum_k p^k a^k(i)}.$$

Starting from a belief p about the state, if player 2 observes action i and knows that the distribution of actions of player 1 is a, then he updates his beliefs to $\hat{q}(p,a|i)$. Aumann and Maschler have proved that the limit value exists and is characterized by

$$v^* = \mathrm{cav} f^* = \inf\{v : \Delta(K) \to [0,1], v \text{ concave } v \geq f^*\},$$

where $f^*(p) = Val\left(\sum_k p^k G^k\right)$ for all $p \in \Delta(K)$. The function f^* is the value of the game, called the non-revealing game, where player 1 is forbidden to use his information.

Theorem 2.4.9 *A zero-sum repeated game with an informed controller has a general uniform value.*

Proof of theorem 2.4.9: Assume that $\Gamma(\pi) = (K, I, J, C, D, q, g, \pi)$ is a repeated game with

an informed controller. The proof will consist of 5 steps. First we introduce an auxiliary standard Markov Decision Process $\Psi(\hat{\pi})$ on the state space $X = \Delta(K)$. Then we show that for all evaluations θ, the repeated game $\Gamma(\pi)$ and the MDP $\Psi(\hat{\pi})$ have the same θ-value. In step 3 we check that the MDP satisfies the assumption of theorem 2.3.19 so it has a general limit value and a general uniform value v^*. As a consequence the repeated game has a general limit value $v^*(\pi)$. Then we prove that player 1 can use an ϵ-optimal strategy of the auxilliary MDP in order to guarantee $v^*(\pi) - \epsilon$ in the original game. Finally we prove that player 2 can play by blocks in the repeated game in order to guarantee $v^*(\pi) + \epsilon$. And we obtain that $v^*(\pi)$ can be guaranteed by both players in the repeated game, so it is the general uniform value of $\Gamma(\pi)$.

For every $\pi \in \Delta(K \times C \times D)$, we denote by $\overline{\pi}$ the marginal of π on $K \times D$ and we put $\hat{\pi} = \psi_D(\overline{\pi})$ where ψ_D is the disintegration with respect to D (seen as a subset of $\mathbb{N}$, recall proposition 2.2.21): for all $\mu \in \Delta(K \times D)$, $\psi_D(\mu) = \sum_{d \in D} \mu(d)\delta_{\mu(.|d)}$.

Step 1: We put $X = \Delta(K)$ and $A = \Delta(I)^K$ and for every p in X, a in A and b in $\Delta(J)$, we define:

$$\begin{aligned}
\widetilde{g}(p,a,b) &= \sum_{(k,i,j) \in K \times I \times J} p^k a^k(i) b(j) g(k,i,j) \in [0,1], \\
r(p,a) &= \inf_{b \in \Delta(J)} \widetilde{g}(p,a,b) = \inf_{j \in J} \widetilde{g}(p,a,j), \\
\overline{q}(p,a) &= \sum_{(k,i) \in K \times I} p^k a^k(i) \overline{q}(k,i) \in \Delta(K \times D), \\
\widetilde{q}(p,a) &= \psi_D(\overline{q}(p,a)) = \sum_{d \in D} \overline{q}(p,a)(d) \delta_{\hat{q}(p,a|d)} \in \Delta_f(X).
\end{aligned}$$

Here $\hat{q}(p,a|d) \in \Delta(K)$ is the belief of the second player on the new state after observing the signal d and knowing that player 1 has played a at p:

$$\forall k' \in K, \hat{q}(p,a|d)(k') = \frac{\overline{q}(p,a)(k',d)}{\overline{q}(p,a)(d)} = \frac{\sum_k p^k q(k,a(k))(k',d)}{\sum_k p^k q(k,a(k))(d)}.$$

We define the auxiliary MDP $\Psi = (X, A, \widetilde{q}, r)$, and denote the θ-value in the MDP by w_θ. The MDP with initial state $\hat{\pi}$ has strong links with the repeated game $\Gamma(\pi)$.

Step 2: By proposition 4.23, part b) in Renault [Ren12b], we have for all evaluations θ with finite support:

$$v_\theta(\pi) = w_\theta(\hat{\pi}).$$

The proof relies on the same recursive formula satisfied by v and w, and the equality can be easily extended to any evaluation θ.

$$\forall \theta \in \Delta(\mathbb{N}^*), \forall p \in X,\ v_\theta(p) = \sup_{a \in A} \inf_{b \in B} \left(\theta_1 r(p,a,b) + (1-\theta_1) v_{\theta^+}(\hat{q}(p,a)) \right).$$

where v_{θ^+} is naturally linearly extended to $\Delta_f(X)$. As a consequence if $\Psi(\hat{\pi})$ has a general limit value so does the repeated game $\Gamma(\pi)$.

Step 3: Let us check that Ψ satisfies the assumption of 2.3.19. Consider p, p' in X, a in A, and $\alpha \geq 0$ and $\beta \geq 0$. We have:

$$\begin{aligned} |\alpha r(p,a) - \beta r(p',a)| &\leq \sup_{b\in\Delta(J)} |\alpha \widetilde{g}(p,a,b) - \beta \widetilde{g}(p',a,b)| \\ &\leq \sup_{b\in\Delta(J)} |\sum_{k\in K} \alpha p^k g(k,a^k,b) - \beta p'^k g(k,a^k,b)| \\ &\leq \sup_{b\in\Delta(J)} \sum_{k\in K} |\alpha p^k - \beta p'^k| = \|\alpha p - \beta p'\|_1. \end{aligned}$$

Moreover, let $\varphi : \Delta(K) \longrightarrow \mathbb{R}$ be in D_1.

$$\begin{aligned} |\alpha\varphi(\hat{q}(p,a)) - \beta\varphi(\hat{q}(p',a))| &= \sum_{d\in D} (\alpha \bar{q}(p,a)(d)\varphi(\hat{q}(p,a|d)) - \beta\bar{q}(p',a)(d)\varphi(\hat{q}(p',a|d))) \\ &\leq \sum_{d\in D} \|\alpha\ \bar{q}(p,a)(d)\ \hat{q}(p,a|d) - \beta\ \bar{q}(p',a)(d)\ \hat{q}(p',a|d)\|_1 \\ &\leq \sum_{d\in D} \|\alpha\ (\bar{q}(p,a)(k',d))_{k'} - \beta\ (\bar{q}(p',a)(k',d))_{k'}\|_1 \\ &\leq \sum_{d\in D}\sum_{k\in K} \|\alpha p^k\ (\bar{q}(k,a)(k',d))_{k'} - \beta p'^k\ (\bar{q}(k,a)(k',d))_{k'}\|_1 \\ &\leq \sum_{d\in D}\sum_{k'\in K}\sum_{k\in K} \bar{q}(k,a)(k',d)|\alpha p^k - \beta p'^k| = \|\alpha p - \beta p'\|_1. \end{aligned}$$

So $\Psi = (X, A, \widetilde{q}, r)$ has a general limit value and a general uniform value that we denote by v^*. As a consequence, $\Gamma(\pi)$ has a general limit value $v^*(\pi)$.

Step 4: Given $\varepsilon > 0$, there exists $\alpha > 0$ and a strategy σ in the MDP $\Psi(\hat{\pi})$ such that the θ-payoff in the MDP is large: $\hat{\gamma}_\theta(\hat{\pi}, \sigma) \geq v^*(\pi) - \varepsilon$ whenever $I(\theta) \leq \alpha$. Moreover if we look at the end of the proof of theorem 2.3.19 we can choose σ to be induced by a deterministic play in the Gambling House $\hat{\Gamma}$ with state space $Z = \Delta_f(X) \times [0,1]$. As a consequence one can mimic σ to construct a strategy σ^* in the original repeated game $\Gamma(\pi)$ such that: $\forall \tau \in \mathcal{T}, \gamma_\theta(\pi, \sigma^*, \tau) \geq v^*(\pi) - \varepsilon$ whenever $I(\theta) \leq \alpha$.

Step 5: Finally we show that player 2 can also guarantee the value v^* in the repeated game Γ. Note that in the repeated game he can not compute the state variable in $\Delta(K)$ without knowing the strategy of player 1. Nevertheless he has no influence on the transition function so playing independently by large blocks will be sufficient for him in order to guarantee $v^*(\pi)$. We use the following characterization of the value proved in Renault [Ren12b]:

$$v^*(\pi) = \inf_{n\geq 0} \sup_{m\geq 1} v_{m,n}(\pi).$$

where $v_{m,n}$ is the value of the game with payoff function the Cesàro mean of the stage payoffs between stages $m+1$ and $m+n$. We proceed as in proposition 4.22 of Renault 2012a. Fix $n_0 \geq 1$, then we consider the strategy τ^* which for each $j \in \mathbb{N}$, plays optimally in the game with the evaluation the Cesàro mean for the payoffs on the block of stages $B^j = \{n_0(j-1)+1, ..., n_0 j\}$. Since player 2 does not influence the state it is well defined and this strategy guarantees $\sup_{t\geq 0} v_{t,n_0}(z)$ for the overall Cesàro mean.

Let $\theta \in \Delta(\mathbb{N}^*)$ and σ be a strategy of player 1. For each $j \geq 1$, denote by $\underline{\theta_j}$ the minimum of θ on the block $B^j = \{(j-1)n_0 + 1, ..., jn_0\}$. We have

$$\begin{aligned}\gamma_\theta(\pi,\sigma,\tau^*) &= \sum_{j=1}^{+\infty} I\!E_{\pi,\sigma,\tau^*}\left(\sum_{t=(j-1)n_0+1}^{jn_0} \theta_t\ g(k_t,a_t,b_t)\right)\\ &\leq \sum_{j=1}^{+\infty} n_0\ \underline{\theta}_j \sup_{t\geq 0} v_{t,n_0}(\pi) + n_0 \sum_{t=1}^{+\infty} |\theta_{t+1} - \theta_t|\\ &\leq \sup_{t\geq 0} v_{t,n_0}(\pi) + n_0 I(\theta).\end{aligned}$$

Given ϵ, there exists n_0 such that $\sup_{t\geq 0} v_{t,n_0}(\pi) \leq v^*(\pi) + \epsilon$. Fix $\alpha = \frac{\epsilon}{n_0}$ and τ^* defined as before then for all θ such that $I(\theta) \leq \alpha$, we have

$$\sup_{\sigma\in\Sigma} \gamma_\theta(\pi,\sigma,\tau^*) \leq v^*(\pi) + 2\epsilon,$$

and this concludes the proof of theorem 2.4.9. □

Example 2.4.10 The computation of the value in two-player repeated games with incomplete information is a difficult problem as shown in the next example introduced in Renault [Ren06] and partially solved by Hörner *et al.* [HRSV10]. The value exists by a theorem in Renault [Ren06] but the value has been computed only for some values of the parameters. The set of states is $K = \{k_1, k_2\}$, the set of actions of player 1 is $I = \{T, B\}$, the set of actions of player 2 is $J = \{L, R\}$, and the payoff is given by

$$\begin{array}{c} \\ T \\ B \\ \\ k_1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} L & R \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \\ T \\ B \\ \\ k_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} L & R \end{array} \\ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad .$$

The evolution of the state does not depend on the actions: at each stage the state stays the same with probability p and changes to the other state with probability $1-p$. At each stage, both players observe the past actions played but only player 1 is informed of the current state (with previous notation $C = K \times I \times J$ and $D = I \times J$). For each $p \in [0,1]$, it defines a repeated game Γ^p. In the case $p = 1$, the matrix is fixed for all the game thus it is a repeated game with incomplete information on one side *à la* Aumann Maschler [AMS95]. For all other positive values of p, the process is ergodic so the limit value is constant, and it is sufficient to

study the case $p \in [1/2, 1)$ by symmetry of the problem. Hörner *et al.* [HRSV10] proved that if $p \in [1/2, 2/3)$, then the value is $v_p = \frac{p}{4p-1}$. If $p \geq 2/3$, we do not know the value except for p^*, the solution of $9p^3 - 12p^2 + 6p - 1 = 0$, where one has $v_p = \frac{p^*}{1-3p^*+6(p^*)^2}$.

Acknowledgements. The authors gratefully acknowledge the support of the Agence Nationale de la Recherche, under grant ANR JEUDY, ANR-10-BLAN 0112, as well as the PEPS project Interactions INS2I “Propriétés des Jeux Stochastiques de Parité à Somme Nulle avec Signaux".

Chapter 3

Commutative stochastic games

Résumé : La transition d'un jeu stochastique fini est dite commutative si l'ordre dans lequel les actions sont jouées n'est pas important. De nombreux problèmes peuvent être reformulés dans ce cadre comme les jeux absorbants. On montre deux résultats : le premier concernant les stratégies du décideur dans les MDPs et le second concernant les jeux stochastiques avec deux joueurs. Lorsqu'il n'y a qu'un seul joueur et les transitions sont déterministes, on montre que l'existence de la valeur uniforme en stratégies pures implique l'existence de stratégies 0-optimales. Lorsqu'il y a deux joueurs, on étudie une classe particulière de jeux stochastiques sur $\mathbb{R}^m$ avec un nombre fini d'actions et où les transitions sont déterministes et 1-Lipschitz pour la norme 1. On montre que ces jeux ont une valeur uniforme en classant les états initiaux selon leur nombre d'actions "cycliques" et en utilisant le résultat de Mertens et Neyman [MN81] sur les jeux stochastiques finis. La même preuve s'étend au cas à somme non-nulle si on utilise le résultat de Vieille [Vie00a][Vie00b]. Enfin ces résultats s'appliquent aux jeux répétés généraux où les joueurs observent les actions, mais pas l'état et où la transition commute, en introduisant un jeu stochastique auxiliaire commutatif et déterministe.

Abstract: We are interested in stochastic games with finite sets of actions where the transitions commute. Many problems satisfy this assumption or can be reformulated in order to satisfy it such as absorbing games. We prove two results: one on the strategies of the decision maker in a MDP and the other on the existence of the uniform value in stochastic games. When there is only one player and the transition mapping is deterministic, we show that the existence of a uniform value in pure strategies implies the existence of 0-optimal strategies. In the framework of two-player stochastic games, we study one class of stochastic games on $\mathbb{R}^m$ with a finite number of actions, where the transitions are deterministic and 1-Lipschitz for the L_1-norm. We prove that these games have a uniform value by induction on the number of "cyclic" actions and by using the theorem of Mertens and Neyman [MN81] on stochastic games with a finite number of states and actions. The same proof extends to the non zero-sum case if we use the result of Vieille [Vie00a][Vie00b]. Moreover, both theorems apply to finite general repeated games, where the players observe past actions, but do not observe the state and where the transition commute, by using their auxiliary stochastic game that is commutative and deterministic.

3.1 Introduction

We are interested in two-player zero-sum stochastic games where the transition commutes. Stochastic games were introduced by Shapley in 1953 [Sha53] in order to study repeated interactions between several players. At each stage, the players take independently some decisions, which lead to an instant payoff and a random lottery on a new state, they observe past decisions and the new state, and the game proceeds to next stage. We focus on stochastic games where the transition is commutative. Given a sequence of decisions, the order of the decisions is irrelevant to determine the state: playing the action profile $a1$ followed by the action profile $a2$ whatever is the state after one stage leads to the same distribution of state as playing first the action profile $a2$ and then $a1$. A Markov chain for example can be interpreted as a stochastic game where the players have a unique action and the commutation assumption is automatically fulfilled. The exploitation of a mineral resource such as oil or gold is an example of an economic problem fitting this assumption: it is enough to remember how much of the resource has been exploited in the past to define the remaining quantity. Another example is a competition between firms, which have to sell some stocks. If we consider the vector of stocks of all firms as the state and the quantities sold by each firm at each stage as the actions, the new state depends on the past decisions but not on their order. Nevertheless the stage payoff depends on the state and on the actions, thus two profiles of actions may lead to different payoffs depending on the order they are played. In game theory, several problems satisfy this assumption, for example the study of repeated games with incomplete information on one side as in Aumann and Maschler (Example 3.2.3, for references see [AMS95]) introduces an auxiliary commutative stochastic game.

When there is only one player, the problem is called a Markov Decision Process (MDP) and the player is called the decision maker. At each stage, the decision maker observes the state and takes a decision. He then receives a stage payoff and a new state is randomly chosen depending on the state and the decision taken. His aim is to maximize some criterium of the payoff. For example, for each $n \in \mathbb{N}$, we denote by v_n the maximal expected average payoff that the decision maker can guarantee in the n-stage game. We will focus on the notion of uniform value and the existence of robust strategies: an MDP has a uniform value if for all $\varepsilon > 0$ there exists a strategy, which guarantees the inferior limit of v_n in all sufficiently long games. Informally, the decision maker can play optimally without knowing exactly the length of the game. Blackwell [Bla62] proved that when the set of states and the set of actions are finite, there exists a uniform value and even a pure stationary strategy that is optimal for another family of criteria: every discount factor close to 1. Dynkin [DJ79] and Renault [Ren11] described sufficient conditions for the existence of the uniform value when the set of states is compact. In general, the decision maker may have to make small irreversible mistakes in order to guarantee the value. In Theorem 3.3.1, we prove that whenever a commutative MDP has a uniform value in pure strategies and a deterministic transition, the decision maker has a

strategy that guarantees exactly the value. Moreover under topological assumptions similar to Renault [Ren11], we build this strategy without randomization.

The model of MDPs was later generalized to MDPs with partial observation (POMDP). In finite MDPs with partial observation, at the end of every stage, the decision maker do not observe the current state and the current payoff but receives a random signal. A usual approach in order to study MDPs with partial observation is to introduce an auxiliary problem with full observation and Borel state space: the space of belief on the state as in Sawaragi and Yoshikawa [SY70] and Rhenius [Rhe74]. Rosenberg, Solan, and Vieille [RSV02] proved that POMDP with a finite set of states, a finite set of actions and a finite set of signals have a uniform value. Moreover for all $\varepsilon > 0$, there exist one strategy ε-optimal both for all discount factor close to 1 and for all sufficiently long finite horizon games. The existence of the uniform value was extended by Renault [Ren11] to any space of actions and signals, provided, at each stage, the decision maker chooses a random distribution with finite support and only a finite number of signals has a positive probability. If the decision maker has no information on the state and the transition is commutative, then the transition in the auxiliary MDP is commutative and deterministic. Therefore we can apply Theorem 3.3.1 to the auxiliary MDP and then deduce the existence of strategies, which are making no irreversible mistake, in the POMDP.

For two-player zero-sum stochastic games, we also have a notion of uniform value. The stochastic game has a uniform value if the players can respectively guarantee $\liminf v_n$ and $\limsup v_n$ in all sufficiently long games. The existence of the uniform value in the finite case was proven by Mertens and Neyman [MN81] with algebraic tools. The general model on a compact set of states is still open and only some special cases have been solved. For example, Renault [Ren12b] shows the existence of the uniform value for a two-player game on a compact subset of a normed vector space where one player controls the transition. In this paper, we are interested in a model where the set of states is a compact subset of $\mathbb{R}^m$, the sets of actions are finite and each transition is a deterministic function non-expansive for the norm $\|.\|_1$. In Theorem 3.3.4, we prove that, under these assumptions, the stochastic game has a uniform value.

Similarly to MDP, stochastic games have been generalized to model where the players are not perfectly informed about the states or the actions played. At each stage, the players receives only a signal. In the following , we will call this model repeated games. They contain both stochastic games where players observe the state and the actions and repeated games with incomplete information on one side. In the model of repeated games *à la* Aumann and Maschler, the state is fixed at the initial stage and does not change during the game but the beliefs of the player change depending on the signals that they are observing. Among all the structure of signals, we will focus on symmetric signals: at each stage, the players learn past actions and receives the same public signal. In repeated games *à la* Aumann and Maschler, Kohlberg and Zamir [KZ74] and Forges [For82] proved the existence of the uniform value for symmetric signalling structure. Neyman and Sorin extended their results to the non zero-sum

case [NS98] and Geitner [Gei02] to a game where at the initial stage, instead of a matrix, a stochastic game is chosen among a family of stochastic games. For the general model where the state is not fixed and can change during the game, the question is still open. We address the particular case of "state-blind" repeated games where there is no public signal: the players only learn the past couple of actions played and such that the transition is commutative. In this case, we can define an auxiliary stochastic game on the common belief of the players, which satisfies the assumption of Theorem 3.3.4. Therefore, this auxiliary game has a uniform value and the existence of the uniform value implies the existence of the uniform value in the original "state-blind" repeated game.

In Section 2 we introduce the formal definition of commutation, the model of stochastic games and the model of "state-blind" repeated games. In Section 3, we state the results. Section 4 is dedicated to several results around the Markov Decision Process framework and especially the proof of 3.3.1. In Section 5, we focus on the results in the framework of stochastic games and the proof of Theorem 3.3.4.

3.2 Model

If X is a non-empty set, we denote by $\Delta_f(X)$ the set of probabilities on X with finite support. When X is finite, we denote it by $\Delta(X)$ and by $\sharp X$ the cardinality of X. We will consider two types of games : stochastic games on a compact set X of states, denoted by $\Gamma = (X, I, J, q, g)$ and "state-blind" repeated games on a finite set K of states, denoted by $\Gamma^{sb} = (K, I, J, q, g)$. The sets of actions will always be finite.

3.2.1 Commutative stochastic games

A two-player zero-sum stochastic game $\Gamma = (X, I, J, q, g)$ is given by: a non-empty set of states X, two finite non-empty sets of actions I and J, a reward function $g : X \times I \times J \to [0, 1]$ and a transition function $q : X \times I \times J \to \Delta_f(X)$.

Given an initial probability distribution $z \in \Delta_f(X)$, the game $\Gamma(z)$ is played as follows. An initial state x_1 is drawn according to z and announced to the players. At each stage $t \geq 1$, player 1 and player 2 choose simultaneously an action, $i_t \in I$ and $j_t \in J$. Player 2 pays to Player 1 the amount $g(x_t, i_t, j_t)$ and a new state x_{t+1} is drawn according to the probability distribution $q(x_t, i_t, j_t)$. Then both players observe the couple of actions (i_t, j_t), the state x_{t+1} and the game proceeds to stage $t + 1$. When the initial distribution is a Dirac mass at $x \in X$, we denote by $\Gamma(x)$ the game $\Gamma(\delta_x)$.

Note that we restrict the transition to have value in the set of probabilities with finite support on X, so at each stage given a state and a couple of actions, the state at the next stage can

take only a finite number of values. Since there is a finite number of actions at each stage, given an initial probability $z \in \Delta_f(X)$ the set of states wich can be reached with positive probability is therefore countable.

For all $i \in I$ and $j \in J$, we extend $q(.,i,j)$ and $r(.,i,j)$ linearly on $\Delta_f(X)$ by

$$\forall z \in \Delta_f(X),\ \widetilde{q}(z,i,j) = \sum_{x \in X} z(x) q(x,i,j) \text{ and } \widetilde{g}(z,i,j) = \sum_{x \in X} z(x) g(x,i,j).$$

Definition 3.2.1 *The transition q commutes on X if for all $x \in X$, for all $i, i' \in I$ and $j, j' \in J$,*

$$\widetilde{q}(q(x,i,j),i',j') = \widetilde{q}(q(x,i',j'),i,j).$$

It means that the distribution over the state after two stages is equal if the couple of actions (i,j) is played before (i',j') or if (i,j) is played after (i',j'). The transition q is not supposed to be deterministic, so $\widetilde{q}(q(x,i',j'),i,j)$ is the law of a random variable x'' computed in two steps: x' is randomly chosen with law $q(x,i',j')$, then x'' is randomly chosen with law $q(x',i,j)$. The action at the second step is the same for all realizations of the first random variable. The commutation assumption is automatically fulfilled, for example, if no player can influence the transitions.

Example 3.2.2 Let X be the set of complex numbers of modulus 1 and f be a function from $I \times J$ to $\Delta_f([0,2\pi])$. We define the transition q from $X \times I \times J$ to $\Delta_f(X)$ by

$$q(x,i,j) = \sum_{\rho} f(i,j)(\rho) \delta_{xe^{i\rho}}.$$

If the state is x and the couple of actions (i,j) is played, then the new state is $x' = xe^{i\rho}$ with probability $f(i,j)(\rho)$. This transition is commutative by commutativity of the addition.

The next example comes from the theory of repeated games with incomplete information on one side (Aumann and Maschler [AMS95]).

Example 3.2.3 A repeated game with incomplete information on one side, Γ, is defined by a finite family of matrices $(G^k)_{k \in K}$, two finite sets of actions I and J, and an initial probability p. At stage 1, a matrix G^k is randomly chosen with law p and told to player 1 whereas player 2 only knows p. Then the matrix game G^k is repeated over and over. The players observe the actions played but not the payoff. In order to study this repeated games, one can introduce a stochastic game on the posterior beliefs of player 2 about the state. Let $\Psi = (X, A, B, \widetilde{q}, \widetilde{g})$ be the stochastic game such that $X = \Delta(K)$, $A = \Delta(I)^K$ and $B = \Delta(J)$, the payoff function is given by

$$\widetilde{g}(p,a,b) = \sum_{k \in K, i \in I, j \in J} p^k a^k(i) b(j) G^k(i,j),$$

and the transition by

$$\widetilde{q}(p,a,b) = \sum_{k\in K, i\in I} a^k(i)\delta_{\hat{p}(a|i)},$$

where $a(i) = \sum_{k\in K} p^k a^k(i)$ and $\hat{p}(a|i) = \left(\frac{p^k a^k(i)}{a(i)}\right)_{k\in K} \in \Delta(K)$.

Knowing the strategy played by player 1, player 2 updates his beliefs depending on the actions of player 1 observed. Note that the second player does not influence the transition so we can forget him.

Let us check that this auxiliary stochastic game is commutative. Let a and a' be two actions of player 1. If player 1 plays first a then a' and player 2 observe action i at the first step and then i', then player 2's belief after one step is p_2

$$\forall k \in K,\ p_2(k|i) = \frac{p^k a^k(i)}{\sum_{k\in K} p^k a^k(i)}$$

and after the second step, his belief p_3 is given by

$$\forall k \in K,\ p_3(k|i,i') = \frac{p_2(k|i)a'^k(i')}{\sum_{k\in K} p_2(k|i)a'^k(i')} = \frac{p^k a'^k(i')a^k(i)}{\sum_{k\in K} p^k a'^k(i')a^k(i)}.$$

If player 1 plays first a' then a, then player 2 observe i' at first stage and i at the second stage with the same probability as in the previous computation. Moreover he has the same belief. Thus the law of the state does not depend on the order and the transition $\widetilde{q}$ is commutative.

Remark 3.2.4 Note that if we consider an initial state x and a finite sequence of actions $(i_1, j_1, ..., i_n, j_n)$, the law of the state at stage $n+1$ does not depend on the order in which the couple of actions (i_t, j_t), $t = 1, ..., n$, are played. So we can represent a finite sequence of actions by a vector in $\mathbb{N}^{I\times J}$ counting how many times each couple of actions is played. Other assumptions have already been studied in the literature where the transition along a sequence of actions is only a function of a parameter in a smaller set. For example, a transition is State Independent (SIT) if it does not depend on the state. The law of the state at stage n is characterized only by the last couple of actions played. The law depends essentially on the order in which the actions are played. Thuijsman [TV92] proved in this framework the existence of stationary optimal strategies.

3.2.2 Evaluation of the payoffs in stochastic games

At stage t, the space of past histories for both players is $H_t = (X \times I \times J)^{t-1} \times X$ and we set $H_\infty = (X \times I \times J)^{+\infty}$ the space of infinite histories. Without additional assumption on X, H_t could be infinite but we will always restrict to probabilities with finite support on H_t. We consider the product topology on H_t and the Borel σ-algebra associated, which contains all countable sets and all complements of countable sets. Each $h_t \in H_t$ can be identified with a cylinder set of H_∞. We denote by $\mathcal{H}_t$ the algebra induced by H_t over H_∞ and by $\mathcal{H}_\infty$ the σ-

algebra spanned by all finite cylinders. A strategy for player 1 is a sequence $(\sigma_t)_{t\geq 1}$ of functions $\sigma_t : H_t \to \Delta(I)$. A strategy for player 2 is a sequence $\tau = (\tau_t)_{t\geq 1}$ of functions $\tau_t : H_t \to \Delta(J)$. We denote by Σ and $\mathcal{T}$ their respective sets of strategies. We do not need additional assumption of measurability since we will only consider probabilities with finite support. If a strategy is such that all images are Dirac measures, the strategy is said to be pure. Moreover if the transition is deterministic and both players use pure strategies, there is only one history with a positive probability. We call it the play and it can be uniquely defined either by the sequence of states visited or by the sequence of actions played. A strategy profile (σ, τ) and an initial probability z induce a probability on each finite cylinder, which can be extended as a unique probability distribution $I\!P_{z,\sigma,\tau}$ over the set of infinite histories $(H_\infty, \mathcal{H}_\infty)$. The set of actions is finite and q has values in laws with finite support, so the set of histories with positive probability under $I\!P_{z,\sigma,\tau}$ is countable.

We are going to use two types of evaluation in this chapter, the n-stage game and the expected average payoff between two stages m and n. For each positive n, the expected average payoff for player 1 up to n stages, induced by the strategy pair (σ, τ) and the initial distribution z, is given by

$$\gamma_n(z, \sigma, \tau) = I\!E_{z,\sigma,\tau}\left(\frac{1}{n}\sum_{t=1}^{n} g(x_t, i_t, j_t)\right).$$

The expected average payoff between two stages $m \leq n$ is given by

$$\gamma_{m/n}(z, \sigma, \tau) = I\!E_{z,\sigma,\tau}\left(\frac{1}{n-m+1}\sum_{t=m}^{n} g(x_t, i_t, j_t)\right).$$

To study the infinite game $\Gamma(z)$ we focus on the notion of uniform value and of ε-optimal strategies.

Definition 3.2.5 *Let v be a real number,*

- *player 1 can guarantee v in $\Gamma(z)$ if for all $\varepsilon > 0$, there exists a strategy $\sigma^* \in \Sigma$ of player 1, such that*
$$\liminf_n \inf_{\tau\in\mathcal{T}} \gamma_n(z, \sigma^*, \tau) \geq v - \varepsilon.$$
We say that such a strategy σ^* guarantees $v - \varepsilon$ in $\Gamma(z)$.
- *player 2 can guarantee v in $\Gamma(z)$ if for all $\varepsilon > 0$, there exists a strategy $\tau^* \in \mathcal{T}$ of player 2, such that*
$$\limsup_n \sup_{\sigma\in\Sigma} \gamma_n(z, \sigma, \tau^*) \leq v + \varepsilon.$$
We say that such a strategy τ^* guarantees $v + \varepsilon$ in $\Gamma(z)$.
- *If both players can guarantee v, v is called the uniform value of the game $\Gamma(z)$ and we denote it by $v^*(z)$.*

Whenever the uniform value exists, a strategy σ, which guarantees $v^*(z) - \varepsilon$ with $\varepsilon \geq 0$, is said to be ε-optimal. The strategy τ of player 2 is ε-optimal if it guarantees $v^*(z) + \varepsilon$.

3.2.3 The model of repeated games with "state-blind" players

A "state-blind" repeated game $\Gamma^{sb} = (K, I, J, q, g)$ is defined by the same objects as a stochastic game. The definition of commutativity is the same. The main difference is the way the game is played and formally the sets of strategies. We assume that at each stage, the players observe the actions played but not the state. We will restrict to a finite state space K.

Given an initial probability $p \in \Delta(K)$, the game $\Gamma^{sb}(p)$ is played as follows. An initial state k_1 is drawn according to p without being announced to the players. At each stage $t \geq 1$, player 1 and player 2 choose simultaneously an action, $i_t \in I$ and $j_t \in J$. Player 1 receives the (unobserved) payoff $g(k_t, i_t, j_t)$, player 2 receives the (unobserved) opposite $-g(k_t, i_t, j_t)$ and a new state k_{t+1} is drawn according to the probability distribution $q(k_t, i_t, j_t)$. Then both players observe only the couple of actions (i_t, j_t) and the game proceeds to stage $t + 1$.

At stage t, the space of past histories for both players becomes $H_t^{sb} = (I \times J)^{t-1}$ and they have a common history. A strategy in Γ^{sb}, for player 1 is a sequence $(\sigma_t)_{t\geq 1}$ of functions $\sigma_t : H_t^{sb} \to \Delta(I)$. A strategy for player 2 is a sequence $\tau = (\tau_t)_{t\geq 1}$ of functions $\tau_t : H_t^{sb} \to \Delta(J)$. We denote by Σ^{sb} and $\mathcal{T}^{sb}$ their respective sets of strategies. An initial distribution p and a couple of strategies $(\sigma, \tau) \in \Sigma^{sb} \times \mathcal{T}^{sb}$ give a unique probability on the infinite histories H^∞ with the σ-field $\mathcal{H}^\infty$. We can define the payoff as before and the notion of uniform value by restricting the players to play strategies in Σ^{sb} and $\mathcal{T}^{sb}$.

Definition 3.2.6 *Let v be a real number,*

- *player 1 can guarantee v in $\Gamma^{sb}(p)$ if for all $\varepsilon > 0$, there exists a strategy $\sigma^* \in \Sigma^{sb}$ of player 1, such that*
$$\liminf_n \inf_{\tau\in\mathcal{T}^{sb}} \gamma_n(p, \sigma^*, \tau) \geq v - \varepsilon.$$
We say that such a strategy σ^* guarantees $v - \varepsilon$ in $\Gamma^{sb}(p)$.
- *player 2 can guarantee v in $\Gamma^{sb}(p)$ if for all $\varepsilon > 0$, there exists a strategy $\tau^* \in \mathcal{T}^{sb}$ of player 2, such that*
$$\limsup_n \sup_{\sigma\in\Sigma^{sb}} \gamma_n(p, \sigma, \tau^*) \leq v + \varepsilon.$$
We say that such a strategy τ^* guarantees $v + \varepsilon$ in $\Gamma^{sb}(p)$.
- *If both players can guarantee v, v is called the uniform value of the game $\Gamma^{sb}(p)$ and we denote it by $v^{sb}(p)$.*

Remark 3.2.7 The sets Σ^{sb} and $\mathcal{T}^{sb}$ can be seen as respectively subsets of Σ and $\mathcal{T}$. There is no relation between $v^{sb}(p)$ and $v^*(p)$ since both players have restricted sets of strategies.

3.3 Results.

3.3.1 Commutative deterministic Markov Decision Processes and 0-optimal strategies.

An MDP is a stochastic process controlled by one decision maker who aims to maximize his payoff. Formally, with the previous notations, an MDP is a stochastic game where J is a singleton. In the following, $\Gamma = (X, I, q, g)$ will define an MDP. The first part of the theorem claims that a game, with a commutative deterministic transition, and, with a uniform value in pure strategies, has a 0-optimal strategy. As we will show in an example later, without the commutativity assumption, this result is false. However, the 0-optimal strategy is not pure since the decision maker has to use some randomizing device. In the second part of the theorem, we give sufficient topological conditions for the existence of the uniform value in pure strategies and for the existence of a pure 0-optimal strategy.

Theorem 3.3.1 *Let $\Gamma = (X, I, q, g)$ be an MDP such that I is finite, q is deterministic and commutative.*

1. *If for all $z \in \Delta_f(X)$, $\Gamma(z)$ has a uniform value in pure strategies then for all $z \in \Delta_f(X)$ there exists a 0-optimal strategy.*
2. *If X is a precompact metric space, q is 1-Lipschitz and g is uniformly continuous then for all $z \in \Delta_f(X)$, the MDP $\Gamma(z)$ has a uniform value and there exists a 0-optimal pure strategy.*

The first part of Theorem 3.3.1 is tight in the sense that a commutative deterministic MDP with a uniform value in pure strategies may have no 0-optimal pure strategy. An example is described at the beginning of Section 4. The topological assumptions of the second part were first introduced by Renault [Ren11] and imply the existence of the uniform value in pure strategies, thus also of a 0-optimal strategy by the first part of the theorem. Under these topological assumption, we prove the stronger result of the existence of a 0-optimal pure strategy. The decision maker can guarantee the payoff exactly without randomizing.

The assumption of precompactness and uniform continuity are natural whereas the assumption that q is 1-Lipschitz may seem too strong. It is a necessary assumption in the paper of Renault [Ren11]. When computing the uniform value, we iterate the transition an infinite number of times. This assumption implies that given two states x an x' and an infinite sequence of actions $(i_1, ..., i_t, ...)$, at each stage, the state on the play from x and the state on the play from x' are at less than $d(x, x')$. On the contrary if q is only 2-Lipschitz, we only know that at stage $t \geq 1$ the state on the play from x and the state on the play from x' are at less than $d(x, x')2^t$ which gives no constraint as soon as t is big enough. When q is not 1-Lipschitz the value may fail to exist as shown in Renault [Ren11]. Nevertheless, his example is not commutative. Maybe the additional assumption of commutativity can help us relaxing the property on q. In our proof, we use that q is non-expansive at two steps: first in

order to apply the result of Renault [Ren11] and then in order to concatenate strategies. It is open if one of these two steps could be done under the weaker assumption that q is uniformly continuous. The two open problems are: assume that the uniform value exists, X precompact, g uniformly continuous, and q is uniformly continuous deterministic and commutative, does there exist a 0-optimal strategy? Does an MDP with X precompact, g uniformly continuous, and q uniformly continuous deterministic and commutative, always have a uniform value?

We deduce from Theorem 3.3.1 the existence of a 0-optimal strategy for commutative MDPs with no information on the state. called MDPs in the dark in the literature. The auxiliary MDP associated to the POMDP is deterministic and commutative thus it satisfies the assumption of Theorem 3.3.1. The lemma proving that the existence of the uniform value in the MDP implies the existence of the uniform value in the POMDP will be proven in the next subsection in the more general framework of "state-blind" repeated games.

Corollary 3.3.2 *Let $\Gamma^{sb} = (K, I, q, g)$ be a commutative state-blind MDP with a finite state space K and a finite set of actions I. For all $p \in \Delta(K)$, $\Gamma^{sb}(p)$ has a uniform value and there exists a 0-optimal pure strategy.*

Rosenberg, Solan, and Vieille [RSV02] asked the question of the existence of a 0-optimal strategy in MDPs with signals. Our assumptions ensure that there exists such a strategy but the following example shows that it is not true without the commutativity assumption. Moreover it implies that there exist MDPs that cannot be transformed into a commutative MDP with finite sets of actions.

Example 3.3.3 We consider an MDP in the dark defined as follows. Let $X = \{\alpha, \beta, k_0, k_1\}$, and $I = \{T, B\}$. The payoff is 0 except in state k_1 where it is 1. The states k_0 and k_1 are absorbing and in the other states the transition function q is given by

$$\begin{aligned} q(\alpha, T) &= \frac{1}{2}\delta_\alpha + \frac{1}{2}\delta_\beta, \\ q(\beta, T) &= \delta_\beta, \\ q(\alpha, B) &= \delta_{k_0}, \\ q(\beta, B) &= \delta_{k_1}. \end{aligned}$$

This game is not commutative: if the initial state is α and the decision maker plays B then T, the state is k_0 with probability one, whereas if he plays first T then B, the state is k_0 with probability $1/2$ and k_1 with probability $1/2$.

This game has a uniform value in pure strategies but no 0-optimal strategies. An ε-optimal strategy in $\Gamma(\alpha)$ is to play the action T until the probability to be in β is more than $1 - \varepsilon$ then to play B. The uniform value starting from α is 1 but there exists no 0-optimal strategy.

In order to get a good payoff at some stage, the decision maker has to play B with positive probability and thus absorbed in state k_0 with some positive probability.

3.3.2 Existence of the uniform value in commutative determinstic stochastic games.

Concerning two-player stochastic games, the commutativity does not imply the existence of 0-optimal strategies. Indeed we will prove in proposition 3.5.1 that an absorbing game can be reformulated into a commutative stochastic game. Since there exist absorbing games with deterministic transitions without 0-optimal strategies, for example the Big Match, there exist deterministic commutative stochastic games without 0-optimal strategies. Instead we study the existence of the uniform value in one class of stochastic games on $\mathbb{R}^m$.

Theorem 3.3.4 *Let $\Gamma = (X, I, J, q, g)$ be a stochastic game such that X is a compact subset of $\mathbb{R}^m$, I and J are finite sets, q is commutative, deterministic and 1-Lipschitz for $\|.\|_1$, and g is continuous. Then for all $z \in \Delta_f(X)$, the stochastic game $\Gamma(z)$ has a uniform value.*

The state space is not finite but we assume that there exist some finite sets of actions fixed for all states. In the more general case where the players have to choose among finite sets $I(x)$ and $J(x)$, the commutativity assumption is not well defined. If we consider another norm on $\mathbb{R}^m$ such as the L_2-norm, the proof does not hold and the question is still open. For example, this theorem does not apply to Example 3.2.2 on the circle. If the unit ball has a finite number of extreme points, the proof can be adapted but the formal definition is postponed to the end of the chapter. Finally note that the most restrictive assumptions are on the transition.

As shown in the MDP framework the assumption, that q is 1-Lipschitz, is important for the existence of a uniform value and is used in the proof at two steps. The first time, we deduce that for all $(i, j) \in I \times J$, iterating infinitely often the couple of actions (i, j) leads to a limit cycle with a finite number of states. The second time we use this assumption in order to concatenate strategies.

We now study "state-blind" repeated games. Given a "state-blind" repeated game $\Gamma^{sb} = (K, I, J, q, g)$ such that q is commutative, we define the auxiliary stochastic game $\Psi = (X, I, J, \widetilde{q}, \widetilde{g})$ with $X = \Delta(K)$, $\widetilde{q}$ the linear extension of q and $\widetilde{g}$ the linear extension of g. A state in this new game is the common belief of the players over the state of $\Gamma^{sb}(K, I, J, q, g)$.

Since K is finite, X can be embedded in $\mathbb{R}^K$ and the transition $\widetilde{q}$ is deterministic, 1-Lipschitz for $\|.\|_1$ and commutative, whenever q is commutative. Furthermore $\widetilde{g}$ is continuous, so we can apply Theorem 3.3.4 to Ψ and for each initial probability $z \in \Delta_f(X)$, $\Psi(z)$ has a uniform value. We deduce that the state-blind repeated game Γ^{sb} has a uniform value by proving that the players can guarantee this value. In this set-up, it is easy since the set of strategies are almost the same in the two games: a player can use a strategy of the repeated game Γ in Ψ by looking only at the actions played and reciprocally a player can use a strategy of the stochastic

game Ψ in the repeated game Γ by completing the sequence of actions with the unique sequence of beliefs compatible.

Corollary 3.3.5 *Let $\Gamma^{sb} = (K, I, J, q, g)$ be a commutative state-blind repeated game with a finite set of states K and finite sets of actions I and J. For all $p \in \Delta(K)$, $\Gamma^{sb}(p)$ has a uniform value.*

Proof: The set of strategies in the game Γ^{sb} are respectively Σ^{sb} and $\mathcal{T}^{sb}$. We will denote in this proof the payoff in the n stage game by γ^{sb} and the value of the n-stage game by $v_n^{sb}(p)$ for all $n \geq 1$.

We denote by $\widetilde{H_t}$ the set of histories in Ψ of length t, $\widetilde{\Sigma}$ the set of strategies of player 1, and $\widetilde{\mathcal{T}}$ the set of strategies of player 2. Let $p \in \Delta(K)$, $\widetilde{\sigma} \in \widetilde{\Sigma}$ and $\widetilde{\tau} \in \widetilde{\mathcal{T}}$, the payoff in the n-stage game, starting from p and given that the players follow $\widetilde{\sigma}$ and $\widetilde{\tau}$, is denoted by $\widetilde{\gamma_n}(\delta_p, \widetilde{\sigma}, \widetilde{\tau})$ and the value by $w_n(p)$. The set X is compact, $\widetilde{g}$ is continuous and the transition $\widetilde{q}$ is commutative and deterministic, so we can apply Theorem 3.3.4 to Ψ. We denote by $w^*(p)$ the uniform value. Let us show that they are equal to their equivalent in Γ^{sb} by proving there exists functions in-between the two sets of strategies of player 1 in the two games and in-between the two sets of strategies of player 2.

We focus on the case of player 1 since the situation is symmetric for player 2. Let σ^{sb} be a strategy in Σ^{sb}, then it defines naturally a strategy $\widetilde{\sigma}$ in $\widetilde{\Sigma}$ by forgetting the states. If we denote by Π^t the projection from $\widetilde{H_t}$ on H_t^{sb} that keeps only the actions: for all $t \geq 1$, we define

$$\widetilde{\sigma}(\widetilde{h}_t) = \sigma^{sb}(\Pi^t(\widetilde{h}_t)).$$

Reciprocally for all $t \geq 1$, given a sequence of actions $h_t^{sb} = (i_1, j_1, ...i_t, j_t)$, the completion $\Xi^t(h^{sb})$ in $\widetilde{H_t}$ is the unique sequence such that for all $t \geq 1$, $q(p_t, i_t, j_t) = p_{t+1}$. Let $\widetilde{\sigma}$ be a strategy in $\widetilde{\Sigma}$, then we define the strategy σ^{sb} by completing the history: for all $t \geq 1$

$$\sigma^{sb}(h_t^{sb}) = \widetilde{\sigma}(\Xi^t(h_t^{sb})).$$

The same procedure gives two functions between the sets of strategies of player 2.

Given $\widetilde{\sigma} \in \widetilde{\Sigma}$ and $\tau^{sb} \in \mathcal{T}^{sb}$, set $\sigma^{sb} \in \Sigma^{sb}$ and $\widetilde{\tau} \in \widetilde{\mathcal{T}}$ as in the previous paragraph. By definition of $\widetilde{q}$, the state at stage t in Ψ under $I\!P_{\delta_p, \widetilde{\sigma}, \widetilde{\tau}}$ is equal to the law of the state in Γ^{sb} under $I\!P_{p, \sigma^{sb}, \tau^{sb}}$. Therefore for all $n \geq 1$, we have

$$\gamma_n^{sb}(p, \sigma^{sb}, \tau^{sb}) = \widetilde{\gamma_n}(\delta_p, \widetilde{\sigma}, \widetilde{\tau}).$$

Finally, let $\varepsilon > 0$, $\widetilde{\sigma}^*$ be an ε-optimal strategy in Ψ and an integer $N \geq 1$ such that for all $\widetilde{\tau} \in \widetilde{\mathcal{T}}$,

$$\widetilde{\gamma_n}(\delta_p, \widetilde{\sigma}^*, \widetilde{\tau}) \geq w^*(p) - \varepsilon,$$

then we denote by $\sigma^{sb,*}$ the corresponding strategy and for all $\tau^{sb} \in \mathcal{T}^{sb}$, we have

$$\begin{aligned}\gamma_n^{sb}(p, \sigma^{sb,*}, \tau^{sb}) &= \widetilde{\gamma_n}(\delta_p, \widetilde{\sigma}^*, \widetilde{\tau}) \\ &\geq w^*(p) - \varepsilon.\end{aligned}$$

The strategy σ^{sb} guarantees $w^*(p) - \varepsilon$ and therefore player 1 guarantees $w^*(p)$. By symmetry, player 2 guarantees $w^*(p)$ and the game $\Gamma^{sb}(p)$ has a uniform value equal to $w^*(p)$. □

Remark 3.3.6 We restrict in Corollary 3.3.5 to repeated games where the players observe past actions but not the state. The more general model, where the players observe past actions and have a public signal on the state, leads to the definition of an auxiliary stochastic game with a probabilistic transition. In this case, the commutativity assumption is not anymore an interesting assumption. Indeed in the definition, we consider that after one stage, the decision maker chooses the same action whatever is the intermediate state. It does not take into account the possibility for the decision maker to play differently depending on the signal he has observed. When the transition is deterministic, this problem does not appear since there is only one intermediate state.

Example 3.3.7 Let $K = \mathbb{Z}/m\mathbb{Z}$ and f be a function from $I \times J$ to $\Delta(K)$. We define the transition $q : K \times I \times J \to \Delta(K)$ by: given a state $k \in K$, if the players play (i, j) then for all $k' \in K$, the new state is $k + k'$ with probability $f(i, j)(k')$.

If the initial state is drawn with a distribution p, the new state, after the players have played (i, j), is given by the sum of two independent random variables of respective laws p and $f(i, j)$. The addition of independent random variables is a commutative and associative operation, therefore q commutes on K.

For example let $m = 3$, $I = \{T, B\}$, $J = \{L, R\}$ and the function f be given by

$$\begin{array}{cc} & \begin{array}{cc} L & \quad R \end{array} \\ \begin{array}{c} T \\ B \end{array} & \begin{pmatrix} \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2 & \delta_1 \\ \delta_1 & \delta_0 \end{pmatrix}. \end{array}$$

If the players play (T, L) then the new state is one of the two other states with equal probability. If the players play (B, R), then the state does not change. And otherwise the state goes from state k to state $k + 1$.

The transition of the auxiliary stochastic game is given by

$$\begin{aligned} \widetilde{q}((p^1, p^2, p^3), T, L) &= \left(\tfrac{p^2+p^3}{2}, \tfrac{p^1+p^3}{2}, \tfrac{p^1+p^2}{2}\right), \\ \widetilde{q}((p^1, p^2, p^3), B, R) &= (p^1, p^2, p^3), \\ \widetilde{q}((p^1, p^2, p^3), B, L) = \widetilde{q}((p^1, p^2, p^3), T, R) &= (p^3, p^1, p^2). \end{aligned}$$

3.4 Existence of 0-optimal strategies in commutative deterministic MDP.

This section is divided into four parts and we focus on Theorem 3.3.1. In the first part, we provide an example showing that the result of Theorem 3.3.1(1) of the existence of a 0-optimal strategy in a commutative deterministic MDP with a uniform value in pure strategies, can not be strengthen in the existence of a 0-optimal pure strategy. In this example, the decision maker has to randomize in order to play 0-optimally. In the second part, we show that the assumptions imply the existence for all $\varepsilon > 0$ of ε-optimal pure strategies such that the value is constant on the play. Along these strategies, the decision maker ensures that when balancing between current payoff and future states, he is not making irreversible mistakes. In the third part, we show the existence of 0-optimal strategies in commutative deterministic stochastic games with a uniform value in pure strategies and thus prove Theorem 3.3.1(2). By using randomization, the decision maker can progressively switch between ε-optimal strategies where the value is constant on the induced play, while ensuring that the expected mean payoff is not dropping. The fourth part is dedicated to the proof of Theorem 3.3.1(2) and show the existence of a pure 0-optimal strategy in an MDP with a metric compact state space X, a 1-Lispchitz transition and a uniformly continuous payoff function. Instead of concatenating strategies one after the other, we define a sequence of strategies, which guarantee the uniform value $v^*(x)$, from several states. Then, we split these strategies in streaks of actions and we build a 0-optimal strategy by playing these blocks in a proper order. In the following, we denote by x_1 the initial state.

3.4.1 Example of a commutative deterministic stochastic games without a 0-optimal pure strategy

In this section, we prove that Theorem 3.3.1(1) is tight in the sense that there exist a deterministic commutative game with a uniform value in pure strategies but without a 0-optimal pure strategy. Before going into details, we outline the structure of the example. The set of states is the countable set $\mathbb{N} \times \mathbb{N}$ and there exists a countable partition of the states such that the payoff is constant on each set of the partition. The payoff is 0 on set h^0 and $1 - \frac{1}{2^l}$ on set h^l for all $l \geq 1$. We will check first that for each $l \geq 1$, there exists a pure strategy from $(0, 0)$ that stays eventually in set h^l, so the game starting at $(0, 0)$ has a uniform value equal to 1. Then we will prove that any 0-optimal pure strategy has to visit all sets h^l and, that when switching from one set h^i to another set $h^{i'}$, the induced play has to stay many stages in set h^0. Moreover the payoff has to drop below $\frac{1}{2}$, which is absurd and there exists no 0-optimal pure strategies in the game starting at state $(0, 0)$.

Example 3.4.1 *The set of states is* $\mathbb{N} \times \mathbb{N}$ *and there are only two actions* R *and* T*.* R *incre-*

ments the first coordinate and T increments the second one

$$q((x,y),R) = (x+1,y),$$
$$q((x,y),T) = (x,y+1).$$

For each $l \geq 1$, we define the play h^l through the sequence of actions $R^{w_l}(TR^{4^{l-1}-1})^\infty$ where $w_l = \sum_{m=1}^{l}(4^{m-1}-1) = \frac{4^l-1}{3} - l$. The payoff is $1-\frac{1}{2^l}$ in every state on the play h^l and 0 if the state is not on a play h^l, $l \geq 1$,

$$\begin{cases} g(w_l,0) & = 1-\frac{1}{2^l}, \\ g(x,y) & = 1-\frac{1}{2^l} \text{ if } x \in \left[w_l + (y-1)\left(4^{l-1}-1\right), w_l + y\left(4^{l-1}-1\right)\right], \\ g(x,y) & = 0 \text{ otherwise.} \end{cases}$$

This MDP is commutative and the transition is deterministic but there is no 0-optimal pure strategy from $(0,0)$.

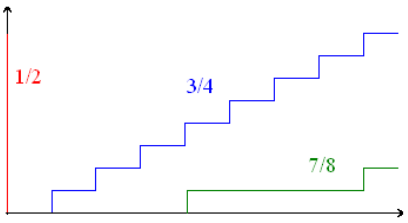

Figure 3.1: Payoff of the game on h^1,h^2 and h^3.

The transition is clearly deterministic and it is immediate to check that it is commutative. Along a sequence of actions, the sequence of states visited has the shape of a stair. Especially for $l \geq 2$, the set of states on h^l where the value is $1-\frac{1}{2^l}$ describes a stair, which begins at state $(0,w_l)$ with a constant rise height of 1 and a tread depth of $4^{l-1}-1$. For $l=1$, the stair starts at $(0,0)$ and is degenerate since there is a unique infinite riser. Figure 3.1 shows the play for $l=1$, $l=2$ and $l=3$ with their respective payoffs. For all $z=(x,y) \in \mathbb{N}\times\mathbb{N}$ and $l \geq 1$, we say that z is strictly in between h^l and h^{l+1}, denoted by $z \in (h^l, h^{l+1})$, if and only if $w_l + (y-1)(4^{l-1}-1) < x < w_{l+1} + (y-1)(4^l-1)$. We denote by $[h^l, h^{l+1})$, the set where the left inequality is not strict. Let us call h^0 the set of states, which are strictly inbetween these stairs, such that $(h^l)_{l\in\mathbb{N}^*}$ and h^0 induce a partition of X.

Proof: Let us first prove that in this example, the uniform value exists in the state $(0,0)$, is equal to 1 and the decision maker has ϵ-optimal pure strategies. Then we will show that he has no 0-optimal pure strategies. Given a play, we call the set of states in this history the path of the strategy and we say that two plays are crossing if there exists a common state on the

two paths. Let us check that there exists a uniform value in pure strategies in all states. For all $(p,q) \in \mathbb{N} \times \mathbb{N}$, there exists l such that $(p,q) \in [h^l, h^{l+1})$. Therefore for all $l' \geq l+1$, there exists a finite number $n(l')$ such that playing $n(l')$ times action R leads to the path $h^{l'}$. Thus by following the path $h^{l'}$ after this stage, the decision maker can guarantee $1 - \frac{1}{2^{l'}}$. The uniform value in pure strategies exists and is equal to 1.

We now show that there exists no 0-optimal pure strategy. Since there is only one player and the transition is deterministic, we can restrict to study 0-optimal sequence of actions and prove it by contradiction. Let $h = (z_1, ..., z_t, ...)$ be a 0-optimal play and n_0 be an integer. Since the strategy h guarantees 1, there exists $n_1 \geq n_0$ such that $g(z_{n_1}) \geq 1/2$. By definition of the payoff function, there exists an integer l such that $g(z_{n_1}) = 1 - \frac{1}{2^l}$ and play h crosses h^l at stage n_1. There also exists $l' > l$, such that play h crosses $h^{l'}$. It implies that h is crossing h^{l+1} at a stage denoted by n_2. Let n_1' and n_2' be two integers such that h crosses h^l at stage n_1', h^{l+1} at stage n_2' and stays in set h^0 inbetween.

Let us prove that by definition of h^l and h^{l+1}, the number of stages between n_1' and n_2' is strictly bigger than n_1'. The state $z_{n_1'}$ is on the play h^l but $z_{n_1'+1}$ is in (h^l, h^{l+1}), so $z_{n_1'}$ is a bottom corner of a stair: there exists y_1 such that $z_{n_1'} = (w_l + y_1\left(4^{l-1} - 1\right), y_1)$. The state $z_{n_2'}$ is on the path of h^{l+1} and $z_{n_2'-1}$ is in (h_l, h_{l+1}), so $z_{n_2'}$ is either the bottom of the stair $z_{n_2'} = (w_{l+1}, 0)$ or a top corner of the stair: there exists y_2 such that $z_{n_2'} = (w_{l+1} + (y_2 - 1)\left(4^l - 1\right), y_2)$ with $y_2 \geq \max(y_1, 1)$. If $y_2 = y_1 = 0$, then there exists a unique play between the two states and its size satisfies $w_{l+1} - w_l = 4^l - 1 \geq 4^{l-1}\frac{4}{3} - l - \frac{1}{3} = w_l \geq n_1'$. If $y_2 \geq 1$, there may be more than one play but they all have the same length given by the sum of the differences coordinate by coordinate, by using $y_2 \geq y_1$, the definition of w_l and

$$\begin{aligned}
n_2' - n_1' &= y_2 - y_1 + w_{l+1} + (y_2 - 1)\left(4^l - 1\right) - w_l - y_1\left(4^{l-1} - 1\right) \\
&= (y_2 - y_1) + (w_{l+1} - w_l - \left(4^l - 1\right)) + y_2\left(4^l - 1\right) - y_1\left(4^{l-1} - 1\right) \\
&\geq 0 + 0 + 3y_1 4^{l-1} \\
&\geq y_1 4^{l-1} + y_l 4^{l-1}\frac{4}{3} \\
&\geq y_1 4^{l-1} + w_l \\
&= n_1'
\end{aligned}$$

Since 1 is an upper bound of the payoff function and the payoff between $n_1' + 1$ and n_2' is 0, the expected average payoff at stage $n_2 - 1$ is less than $\frac{1}{2}$. So for all $n_0 \in \mathbb{N}$, there exists $n_2' \geq n_0$ with $\gamma_{n_2'}((0,0), h) \leq \frac{1}{2}$. This contradicts the optimality of h. The decision maker has no 0-optimal pure strategies. □

3.4.2 Existence of ε-optimal strategies with a constant value on the induced play

We first show that in commutative deterministic MDPs, there exists ε-optimal pure strategies such that the value is constant along the play. Lehrer and Sorin [LS92] showed that in deterministic MDPs, given a sequence of actions the value is always non increasing along the play induced. We need to prove that for all $\varepsilon > 0$, there exists a ε-optimal pure strategy such that the value is non decreasing.

By the commutativity assumption, the order the couple of actions are played is not important, thus we can consider a sequence of strategies that are ε_n-optimal with ε_n converging to 0 and such that each action is played more and more as n goes to infinity.

We define formally this notion. For each $n \in \mathbb{N}$ and finite sequence of length n, h_n, we denote by $M(h_n)$ a vector in $\mathbb{N}^I$ counting how many times each action is played. This function can be extended to infinite sequence of actions by taking the limit in $(\mathbb{N} \cup \{+\infty\})^I$ and for each $h \in H_\infty$, we denote by $M(h)$ the vector counting how many times each action is played. Let $i \in I$, if $M(h)(i) = l < +\infty$ then action i is played l times in h. If $M(h)(i) = +\infty$, then action i is played infinitely in h. If we denote by h_n the projection of h on the set of n-stage histories, we have: for all $A \in \mathbb{N}$, there exists an integer n such that for all $i \in I$ and $n' \geq n$,

$$M(h_{n'})(i) \geq \min(M(h)(i), A). \tag{3.1}$$

Lemma 3.4.2 *We consider a commutative deterministic MDP with a uniform value in pure strategies. For all $x_1 \in X$ and $\varepsilon > 0$, there exists an ε-optimal strategy in $\Gamma(x_1)$ such that the value is non decreasing on the play.*

Proof: Let $(\varepsilon_l)_{l\in\mathbb{N}}$ be a decreasing sequence of positive numbers, which converges to 0, and for each $l \in \mathbb{N}$, h^l an ε_l-optimal sequence of actions in $\Gamma(x_1)$. Since the number of action is finite, we can define a subsequence $(\varepsilon_{\psi(l)})_{l\in\mathbb{N}}$ such that for all $i \in I$, $M(h_{\psi(l)})(i)$ is increasing in l. Actions are played more often in the following sense. If an action is played a finite number of times in the strategy $h_{\psi(l)}$, then this action is played as many times in each $h_{\psi(l')}$ for $l' > l$. If it is played infinitely often in the strategy $\sigma_{\psi(l)}$ then this action is also played infinity often in all $\sigma_{\psi(l')}$ for $l' > l$.

In the following, we forget the initial sequence and just keep the subsequence: $(\eta_l = \varepsilon_{\psi(l)})_{l\in\mathbb{N}}$ and for each $l \in \mathbb{N}$, $s^l \in I^\infty$, an η_l-optimal strategy in $\Gamma(x_1)$ such that sequence $M(s^l)$ of vectors is non decreasing component by component.

Let $l' > l$ be two integers, we show that for all $n \in \mathbb{N}$ we can complete an n-stage history of s^l in order to cross $s^{l'}$: there exists $n' \in \mathbb{N}$ and w a sequence of actions of length $n' - n$ such that the state after (s^l_n, w) and after $s^{l'}_{n'}$ are the same. Let $n \in \mathbb{N}$, we have the component-wise inequality

$$M(s^l_n) \leq M(s^l) \leq M(s^{l'}),$$

so by definition of the convergence of $(M(s^{l'}_n))_{n\in\mathbb{N}}$ (3.1) and taking $A = M(s^l_n)$, there exists a

stage n' such that

$$M(s^{l'}_{n'}) \geq \min(M(s^{l'}), M(s^l_n)) = M(s^l_n).$$

The vector $w^{l,l'} = M(s^l_n) - M(s^{l'}_{n'}) \in \mathbb{N}^I$ represents the sequence of missing actions in s^l_n compared to $s^{l'}_{n'}$ and by commutativity the state after (s^l_n, w) and after $s^{l'}_{n'}$ are the same.

Given $l > l' \in \mathbb{N}$, for all $n \in \mathbb{N}$, there exists a strategy from x^l_n, a state on s^l, which guarantees $v(x_1) - 2\eta_{l'}$ by playing some actions $w^{l,l'}$ until crossing $s^{l'}$ and then following $s^{l'}$. Since it is true for all $l' > l$, the uniform value in x^l_n is equal to $v(x_1)$ and the uniform value is non decreasing. □

3.4.3 Existence of a 0-optimal strategy in the general case

In this subsection, we prove that in every commutative MDP with a uniform value in pure strategies, there exists a 0-optimal strategy and thus Theorem 3.3.1(1). The 0-optimal strategy σ is given by a countable concatenation of ε_l-optimal strategies where the value is constant on the induced play.

A naive approach in order to build a 0-optimal strategy at x_1 is to start by following an ε_1-optimal strategy where the value is constant on the induced play, then at some stage n_1 to switch to an ε_2-optimal strategy at the current state, and so forth so on. Since the value is constant along each play, providing the decision maker follows each strategy for long enough, there exists a subsequence of average payoffs which converges to the value. Nevertheless as seen in Example 3.4.1, the payoff may not converge: playing an ε_2-optimal strategy from stage n_1 will eventually guarantee the value with an error ε_2 but it may leads to a streak of bad payoffs. In Example 3.4.1, any pure strategy which changes a flag of stairs only finitely many times can only be ϵ-optimal so a pure 0-optimal strategy has to change infinitely often. But the stairs are going away one from each others, so every pure strategy that changes a flag of stairs infinitely often yields a low payoff infinitely often.

In order to prevent the payoff from dropping, one way is to choose between various pure strategies, all of which switching infinitely often between ε_l-optimal strategies, yet at a given time only one of them is switching between to ε-optimal strategies and still giving a bad payoffs. This ensures that the long-run average payoffs converges to the value. By Kuhn's theorem a probability on the set of pure strategies is equivalent to a proper strategy. In order to ensure these properties, we define the switching stages as random variable whose support are defined by induction.

We denote by $(u_i)_{i\in\mathbb{N}}$ a sequence of increasing stopping time. For all $i \in \mathbb{N}$, at time u_i, the decision maker forgets the past and starts playing an ε_i-optimal strategy in $\Gamma(x_{u_i})$. We will call a realization of u_i, a value of u_i that is actually observed. Given a realization of u_i, the strategy, starting to play ε_i-optimal in $\Gamma(x_{u_i})$, may give a bad payoff in short games but it will eventually gives a payoff higher than the uniform value with an error ε_i.

Let us first define formally the concatenation of two strategies. Given a stopping time u

and two strategies σ, σ' we define $\sigma u \sigma'$ as follows: play σ until u, then switch to σ' (and forget the history up to u). For every $t \in \mathbb{N}$ and every $h_t = (x_1, i_1, j_1, ..., x_t)$, $(\sigma u \sigma')(h_t) = \sigma(h_t)$ if $u(h_t) > t$ and $(\sigma u \sigma')(h_t) = \sigma'(h_t^u)$ if $u(h_t) \leq t$ where $h_t^u = (x_u, i_u, j_u ..., x_t)$. At stage u, the decision maker forgets the past and starts following σ' as if it was the first stage of the game $\Gamma(x_u)$. The definition for a finite number of stopping time is the same and we will show in the proof that we can give a meaning to a countable number of concatenation.

Definition of the strategy: Let $x_1 \in X$ and $(\varepsilon_l)_{l \in \mathbb{N}}$ be a decreasing sequence converging to 0. We define a sequence σ_l of ε_l-optimal strategies such that the value is constant and we will define the 0-optimal strategy by switching between them at times u_l. For each $l \geq 1$, the stopping time u_l will be given by a uniform distribution over a finite set T_l. The sets $(T_l)_{l \geq 1}$ are defined by induction.

For each $x \in X$ and integer l, we denote by $\sigma_l(x)$ an ε_l-optimal strategy in $\Gamma(x)$ such that the value is constant on the play and $N(l, x)$ an integer such that

$$\forall n \geq N(l,x),\ \gamma_n(x, \sigma_l(x)) \geq v^*(x) - \varepsilon_l.$$

If we consider games longer than $N(l, x)$ stages, the expected average payoff is close to the value but the payoff in shorter games is not controlled.

We define recursively, $(T_j)_{j \in \mathbb{N}}$ the support of the stopping times. Let $t_0 = 1$, $T_0^1 = 0$. We assume that the set T_j exists. We denote $t_{j+1} = \left[\frac{1}{\varepsilon_{j+1}}\right] + 1$ and define the next set T_{j+1} by:

$$\begin{aligned}
T_{j+1}^1 &= T_j^{t_j} + N(j, x_{T_j^{t_j}}) + \frac{1}{\varepsilon_j} T_j^{t_j}, \\
T_{j+1}^2 &= T_{j+1}^1 + N(j+1, x_{T_{j+1}^1}), \\
&\ldots\ldots, \\
T_{j+1}^{t_{j+1}} &= T_{j+1}^{t_{j+1}-1} + N\left(j+1, x_{T_{j+1}^{t_{j+1}-1}}\right).
\end{aligned}$$

Between each possible realization of u_j, one waits enough in order for a strategy, which starts playing like $\sigma_j(x_t)$ at stage $t \in T_j^i$, to give an expected average payoff greater than $v^*(x_1) - \varepsilon_j$. Moreover the first possible realization of u_j is big enough in order to outweigh everything that happened before.

For each set T_j, we define the stopping time u_j such that for all $m \in \{1, ..., t_j\}$, $P(u_j = T_j^m) = \frac{1}{t_j}$. Each realization has a probability less than ε_j. Let

$$\sigma_j^*(x_1) = \sigma_0(x_1) u_1 \sigma_1(x_{u_1}) .... u_j \sigma_j(x_{u_j})$$

and σ^* be the strategy, which coincides for each $j \in \mathbb{N}$ with σ_j^* on the set $\{n \leq u_{j+1}\}$,

$$\sigma^*(x_1) = \sigma_0(x_1) u_1 \sigma_1(x_{u_1}) .... u_j \sigma_j(x_{u_j}) ...$$

Lemma 3.4.3 *$\sigma^*(x_1)$ is a 0-optimal strategy.*

Proof: For each strategy, the value is constant so all the states with a positive probability have the same uniform value $v^*(x_1)$. Let $j \in \mathbb{N}$ and $n \in [T^1_{j+1}; T^1_{j+2} - 1]$, we consider the n-stage value and we study the payoff of σ^*_j, then of σ^*_{j+1} and finally of σ^*.

By definition, $\sigma^*_j = \sigma^*_{j-1} u_j \sigma_j(x_{u_j})$ and the number of stages between a realization of u_j and n is greater than between u_j and T^1_{j+1}, thus the payoff is close to the expected average payoff of $\sigma_j(x_{u_j})$. For $m \in \{1, ..., t_j\}$, we have

$$\frac{T^m_j - 1}{n} \leq \frac{T^{t_j}_j}{T^1_{j+1}} \leq \frac{T^{t_j}_j}{T^{t_j}_j + N(j, x_{T^{t_j}_j}) + \frac{1}{\varepsilon_j} T^{t_j}_j} \leq \varepsilon_j,$$

so if we consider the strategy that switches with probability 1 at T^m_j, we get

$$\begin{aligned}
\gamma_n(x_1, \sigma^*_{j-1} T^m_j \sigma_j) &= E\left[\frac{T^m_j - 1}{n} \gamma_{T^m_j - 1}(x_1, \sigma^*_j) + \frac{n - (T^m_j - 1)}{n} \gamma_{T^m_j/n}(x_1, \sigma^*_j)\right], \\
&\geq E\left[\gamma_{T^m_j/n}(x_1, \sigma^*_j) - \frac{T^m_j - 1}{n}\right], \\
&\geq E\left[\gamma_{n - T^m_j + 1}(x_{T^m_j}, \sigma_j(x_{T^m_j}))\right] - \varepsilon_j.
\end{aligned}$$

Moreover $\sigma_j(x_{T^m_j})$ is an ε_j-optimal strategy and $n - T^m_j \geq N(x_{T^m_j}, \varepsilon_j)$, so

$$\begin{aligned}
\gamma_n(x_1, \sigma^*_{j-1} T^m_j \sigma_j) &\geq E\left[v^*(x_{T^m_j}) - \varepsilon_j\right] - \varepsilon_j \\
&\geq v^*(x_1) - 2\varepsilon_j,
\end{aligned}$$

since the value is constant along all strategies. For any realization of u_j, the payoff of σ^*_j is greater than $v^*(x_1) - 2\varepsilon_j$ at stage n.

Since $\sigma^*_{j+1} = \sigma^*_j u_{j+1} \sigma_{j+1}(x_{u_{j+1}})$, we can now study σ^*_{j+1}. The strategies σ^*_j and σ^*_{j+1} coincide until the realization of u_{j+1} so

$$\begin{aligned}
&\gamma_n(x_1, \sigma^*_{j+1}) \\
&= E\left[\left(\frac{u_{j+1} - 1}{n} \gamma_{u_{j+1} - 1}(x_1, \sigma^*_{j+1}) + \frac{n - (u_{j+1} - 1)}{n} \gamma_{u_{j+1}/n}(x_1, \sigma^*_{j+1})\right) \mathbf{1}_{u_{j+1} < n} + \left(\gamma_n(x_1, \sigma^*_{j+1})\right) \mathbf{1}_{u_{j+1} \geq n}\right] \\
&= E\left[\sum_{t \in T_{j+1}} \left(\left(\frac{t-1}{n} \gamma_{t-1}(x_1, \sigma^*_j) + \frac{n - (t-1)}{n} \gamma_{t/n}(x_1, \sigma^*_{j+1})\right) \mathbf{1}_{u_{j+1} = t < n} + \left(\gamma_n(x_1, \sigma^*_j)\right) \mathbf{1}_{u_{j+1} = t \geq n}\right)\right].
\end{aligned}$$

For each stage $n \in [T^1_{j+1}, T^1_{j+2} - 1]$, there exists a unique integer $m \in \{1, ..., t_{j+1}\}$ such that

$n \in [T^m_{j+1}, T^{m+1}_{j+1} - 1]$ and we can separate the histories into 3 cases (we identify m and the stage T^m_{j+1}): the realization of u_{j+1} is smaller than $m-1$, it is equal to m or bigger than $m+1$. If $u_{j+1} > m$ then the strategy σ^*_{j+1} coincides with σ^*_j on the histories of length n and thus guarantees $v^*(x_1) - 2\varepsilon_j$. If $u_{j+1} < m$ then there exists $l \leq m-1$ such that $u_{j+1} = T^l_{j+1}$, so σ_{j+1} has been played for a long time

$$n - T^l_{j+1} \geq N(j+1, x_{T^l_{j+1}})$$

and therefore

$$\begin{aligned}
\frac{T^l_{j+1}-1}{n}\gamma_{T^l_{j+1}-1}(x_1, \sigma^*_j) + \frac{n+1-T^l_{j+1}}{n}\gamma_{T^l_{j+1}/n}(x_1, \sigma^*_{j+1}) \\
&\geq \frac{T^l_{j+1}-1}{n}(v^*(x_1) - 2\varepsilon_j) + \frac{n+1-T^l_{j+1}}{n}(v^*(x_{T^l_{j+1}}) - \varepsilon_{j+1}) \\
&\geq v^*(x_1) - 2\varepsilon_j.
\end{aligned}$$

Finally the probability of the event $\{u_j = T^m_{j+1}\}$ is less than ε_j, so

$$\begin{aligned}
\gamma_n(x_1, \sigma^*_{j+1}) &\geq P(u_j \leq m)(v^*(x_1) - 2\varepsilon_j) + P(u_j \geq m+1)(v^*(x_1) - 2\varepsilon_j) \\
&\geq v^*(x_1) - 2\varepsilon_j - P(u_j = m) \\
&\geq v^*(x_1) - 3\varepsilon_j.
\end{aligned}$$

The strategy σ^* and σ^*_{j+1} coincide on $[T^1_{j+1}; T^1_{j+2} - 1]$, so for all $j \in \mathbb{N}$, for all $n \in [T^1_{j+1}; T^1_{j+2} - 1]$, the strategy σ^* guarantees the payoff $v^*(x_1) - 3\varepsilon_j$ and therefore is 0-optimal. □

3.4.4 Existence of a pure 0-optimal strategy in the non-expansive framework

In this subsection, we will prove that there exists a pure 0-optimal strategy under the topological assumption of Theorem 3.3.1(2): the set of state X is precompact, the transition is non-expansive deterministic and commutative, and the payoff function is uniformly continuous. In the proof we will first prove that we can assume the set of states to be compact. Then we define recursively a sequence of states $(x^l)_{l \in \mathbb{N}^*}$ such that $x^1 = x_1$ and x^{l+1} is a limit point of states along an ε_l-optimal pure strategy $\sigma^l(x^l)$ starting from x^l where the value is constant on the induced play. It implies especially that for all $l \geq 1$, the uniform value at x^l is equal to $v^*(x_1)$.

For each $l \in \mathbb{N}^*$, we will define by induction $(n^l_k)_{k \in \mathbb{N}^*}$ a sequence of stages satisfying several properties, especially that the sequence of the states on σ^l at these stages converges to the limit point x^{l+1}. This sequence of stages splits the strategy σ^l in a finite sequence of states. Given

$k \geq 1$, we call an elementary block the sequence of actions played between stage n_k^l and stage n_{k+1}^l. Note that it has $n_k^l - n_{k-1}^l$ actions. By convention, the first block starts at stage $n_0^l = 1$.

We define the 0-optimal strategy σ^* by playing these elementary blocks in a specific order. The strategy σ^* is defined as a succession of two types of blocks $(A_l)_{l\geq 1}$ and $(B_l)_{l\geq 1}$ such that for all $l \geq 1$, A_l is composed of $l+1$ elementary blocks taken on $\sigma^l(x^l)$ and B_l is composed of $l-1$ elementary blocks, one from each $\sigma^{l'}(x^{l'})$ for $1 \leq l' \leq l-1$:

$$\sigma^* = (A_1, B_1, A_2, B_2, A_3, ...).$$

The block B_{l-1} ensures that the state at the beginning of the block A_l is close to x^l. The block A_l guarantees an expected average payoff close to the value with an error depending on ε_l. The block A_l is long enough for the expected average payoff of σ^* at the end of A_l to be close to the value. The rest of the proof consists in showing that the expected average payoff is not dropping between these stages, neither during B_{l+1} nor during the first stages of A_{l+1}.

Remark 3.4.4 We first show that we can assume X to be compact without loss of generality by using the existence result of Renault [Ren11]. Let X be a precompact metric space, q be a non-expansive transition and g be a uniformly continuous payoff function. We denote by Γ this MDP. Let $\hat{X}$ be the Cauchy completion of X, $\hat{q}$ and $\hat{g}$ the extensions of q and g to the adherence of X in $\hat{X}$ that is $\hat{X}$. It defines another MDP $\hat{\Gamma}$ on a compact set with a non-expansive transition and a reward function uniformly continuous.

By Renault [Ren11], these two games have a uniform value for any initial probability. Moreover, if x_1 is a state in X, any play in the game $\hat{\Gamma}$ from x_1 stays in X by an immediate induction: $\hat{q}$ and q coincides on X and q has values in $\Delta_f(X)$, so starting in X, the state at the next stage is with probability 1 in X. Therefore, the two value functions are equal at x_1 and a 0-optimal strategy in $\hat{\Gamma}$ is also well defined in the game Γ. So $\Gamma(x_1)$ has a uniform value if and only if $\hat{\Gamma}(x_1)$ has a uniform value.

Note that the assumption that r is uniformly continuous and not simply continuous is necessary in order to prove the existence of the uniform value. In the following example, the state space is metric compact, the transition is 1-Lipschitz and the payoff function continuous but there exists no uniform value.

Example 3.4.5 *Let X be the set of non negative integers with the following metric*

$$d(p,q) = \begin{cases} 0 \text{ if } p = q, \\ \frac{1}{2^p} + \frac{1}{2^q} \text{ if } p \neq q. \end{cases}$$

We consider an MDP with only one action, the transition is given by $q(n) = n+1$ for all $n \in \mathbb{N}$. Any payoff function is continuous on (X,d), the transition is 1-Lipschitz from (X,d) to (X,d), and (X,d) is metric compact.

There exist payoff functions such that the uniform value does not exist: consider a sequence of 0 *and* 1 *such that the sequence of average payoffs diverges. But if we restrict to uniformly continuous payoff function, then the sequence* $(g(n))_{n\in\mathbb{N}}$ *has to converge and the sequence of average payoffs converges.*

Definition of the strategy: In the following we will assume that X is compact. Let $x_1 \in X$ and $(\varepsilon_l)_{l\geq 1}$ be a decreasing sequence of positive numbers, which converges to 0. For each $x \in X$ and $l \geq 1$, we denote by $\sigma_l(x)$ an ε_l-optimal strategy in $\Gamma(x)$ such that the value is constant and by $N(l,x)$ an integer such that

$$\forall n \geq N(l,x),\ \gamma_n(x,\sigma_l(x)) \geq v^*(x) - \varepsilon_l$$

Given a sequence of actions σ, we denote by $(x_t)_{t\in\mathbb{N}^*}$ the sequence obtained from x by playing σ and $(x'_t)_{t\in\mathbb{N}^*}$ the sequence obtained from x' by playing σ. Since g is uniformly continuous there exists $(\eta_l)_{l\geq 1}$ such that

$$\forall x,x' \in X,\ d(x,x') \leq \eta_l,\ \forall a \in \Delta(I),\ |g(x,a) - g(x',a)| \leq \varepsilon_l.$$

Moreover since q is non-expansive, given a sequence of actions σ, $\forall t \geq 1$ we have $d(x_t,x'_t) \leq d(x,x')$ by an immediate induction and we have

$$\forall x,x' \in X,\ d(x,x') \leq \eta_l,\ \forall \sigma,\ \forall n \geq 1,\ |\gamma_n(x,\sigma) - \gamma_n(x',\sigma)| \leq \varepsilon_l.$$

Let $x^1 = x_1$ and given $(x^j)_{1\leq j\leq l}$ we define x^{l+1} a limit point of the play $(x^l,\sigma_l(x^l))$. Since the value is constant on the play $(x^l,\sigma_l(x^l))$, the uniform value in x^{l+1} is also equal to $v(x_1)$. To construct our 0-optimal strategy, we split each play $\sigma_j(x^j)$ in blocks by induction on j. Let us assume that the sequence of stages are defined for all $j \leq l-1$, we define the sequence for $j = l$. Set $L_l = 1 + \sum_{j\leq l-1}(n^j_l - 1)$, which depends only on the sequences for $j \leq (l-1)$, and denote by $(x^l_t)_{t\geq 1}$ the sequence of states along $(x^l,\sigma_l(x^l))$. We define the sequence $(n^l_k)_{k\geq 1}$ such that it satisfies 4 properties. First the strategy $\sigma_l(x^l)$ guarantees in $\Gamma(x^l)$ the value with an error less than ε_l in all games longer than $n^l_{l+1} - 1$,

$$n^l_{l+1} - 1 \ \ \geq N(l,x^l). \tag{3.2}$$

Secondly L_l is small compared to n^l_{l+1},

$$\frac{L_l}{n^l_{l+1}} \leq \varepsilon_l. \tag{3.3}$$

Thirdly at the beginning of the l block of this decomposition the state is near the limit point

$$d(x^l_{n^l_k}, x^{l+1}) \leq \frac{\eta_k}{k-1}. \tag{3.4}$$

Finally we assume that

$$\frac{N(l+1,x^{l+1})+\sum_{j=1}^{l-1}\left(n_{l+1}^{j}-n_{l}^{j}\right)}{n_{l+1}^{l}}\leq\varepsilon_{l}. \tag{3.5}$$

Let $l\in\mathbb{N}^{*}$. We define A_l the finite sequence of actions given by $\sigma^l(x^l)$ between stage 1 and stage n_{l+1}^l. In term of elementary blocks, it is composed of the first $l+1$ elementary blocks of $\sigma^l(x^l)$ and is composed of n_{l+1}^l-1 actions. We define B_l as the sequence of actions where the decision maker is playing, for each $l'<l$, the elementary block of $\sigma^{l'}(x^{l'})$ between stages $n_l^{l'}$ and $n_{l+1}^{l'}$. Thus B_l is the concatenation of $l-1$ elementary blocks. Moreover the number of actions on B_l is $b_l=\sum_{j=1}^{l-1}\left(n_{l+1}^{j}-n_{l}^{j}\right)$, which appeared in (3.5). The strategy σ^* is the sequence of actions given by the alternating sequence $(A_l,B_l)_{l\geq1}$.

Let us show that this strategy is 0-optimal. We prove that, for each $l\geq1$, the state at the beginning of A_l is close to x^l and we deduce that the expected average payoff of σ^* at the end of A_l is bigger than $v^*(x_1)-3\varepsilon_l$. Then we check that this payoff is not dropping between these stages. We have to study two cases. First, in a finite game that ends at a stage in B_l or after less than $N(l+1,x^{l+1})$ stages after the start of A_{l+1}, the number of stages is almost the same as if we have considered the game finishing at the end of A_l so the expected average payoff is greater than $v^*(x_1)-4\varepsilon_l$. Secondly in a longer game, σ^* guarantees a good payoff until the end of A_l and on the part of A_{l+1} played. Since the weight of B_l is small, the expected average payoff is also greater than $v^*(x_1)-4\varepsilon_l$.

Lemma 3.4.6 *The payoff at the end of A_l is greater than $v^*(x_1)-3\varepsilon_l$.*

Proof: Let us denote by $(x_t)_{t\geq1}$ the sequence of states when σ^* is played. One can check that the first stage of A_l is the stage $L_l=1+\sum_{j\leq l-1}(n_l^j-1)$. By definition, at stage L_l for each $l'\leq l-1$, all first l elementary blocks of $\sigma^{l'}(x^{l'})$ have been played: the first $l'+1$ on block l' and then one after the other for each $j\in[l'+1,l-1]$. By commutativity, the state does not depend on the order and the state is the same as after the sequence σ' where the decision maker plays: $\sigma_1(x^1)$ for n_l^1-1 stages, $\sigma_2(x^2)$ for n_l^2-1 stages,..., $\sigma_{l-1}(x^{l-1})$ for $n_l^{l-1}-1$ stages. For each strategy σ_j, equation (3.4) implies that the distance between x^{j+1} and the state after n_l^j-1 actions starting from x^j (so at stage n_l^j) is less than $\frac{\eta_l}{l-1}$ for each $j\in\{1,...,l-1\}$. The map q is non-expansive, so the distances sum up and an immediate induction implies that

$$d(x_{L_l},x^l)\leq\eta_l. \tag{3.6}$$

If we consider the game until the last actions played in A_l, it has $L_l+n_{l+1}^l-1$ stages. The

size of A_l outweighs the past stages and $L_l \geq 1$ so we have by equation (3.3)

$$\begin{aligned}\gamma_{L_l+n_{l+1}^l-1}(x_1,\sigma^*) &= \frac{L_l-1}{L_l+n_{l+1}^l-1}\gamma_{L_l-1}(x_1,\sigma^*) + \frac{n_{l+1}^l}{L_l+n_{l+1}^l-1}\gamma_{L_l/L_l+n_{l+1}^l-1}(x_1,\sigma^*)\\ &\geq \frac{n_{l+1}^l}{L_l+n_{l+1}^l-1}\gamma_{L_l/L_l+n_{l+1}^l-1}(x_1,\sigma^*)\\ &\geq \gamma_{L_l/L_l+n_{l+1}^l-1}(x_1,\sigma^*) - \frac{L_l}{L_l+n_{l+1}^l-1}\\ &\geq \gamma_{L_l/L_l+n_{l+1}^l-1}(x_1,\sigma^*) - \varepsilon_l.\end{aligned}$$

Moreover σ^* is playing like an ε_l-optimal strategy in $\Gamma(x^l)$ between stages L_l and $L_l+n_{l+1}^l-1$ so for long enough by equation (3.2), and the distance between x_{L_l} and x^l is less than η_l by equation (3.6) thus

$$\begin{aligned}\gamma_{L_l+n_{l+1}^l-1}(x_1,\sigma^*) &\geq \gamma_{n_{l+1}^l-1}(x_{L_l},\sigma_l(x^l)) - \varepsilon_l\\ &\geq \gamma_{n_{l+1}^l}(x^l,\sigma_l(x^l)) - 2\varepsilon_l\\ &\geq v^*(x_1) - 3\varepsilon_l.\end{aligned}$$

□

For each $l \geq 1$, the expected average payoff is good at the end of the block A_l. We now prove that it is not dropping until the strategy played in A_{l+1} ensures a good payoff itself.

Lemma 3.4.7 *The payoff in any n-stage game stopping either in the middle of B_l or after less than $N(l+1,x^{l+1})$ stages after the beginning of A_{l+1} is greater than $v^*(x_1)-4\varepsilon_l$*

Proof: If $n \in [L_l+n_{l+1}^l, L_{l+1}+N(l+1,x^{l+1})]$ the number of stages is close to $L_l+n_{l+1}^l-1$ by equation (3.5),

$$\begin{aligned}n - L_l - n_{l+1}^l + 1 &\leq N(l+1,x^{l+1}) + \sum_{j=1}^{l-1} n_{l+1}^j - n_l^j\\ &\leq \varepsilon n_{l+1}^l.\end{aligned}$$

So the payoff is close to the previous case,

$$\begin{aligned}
\gamma_n(x_1,\sigma^*) &= \frac{L_l + n^l_{l+1} - 1}{n}\gamma_{L_l+n^l_{l+1}-1}(x_1,\sigma^*) + \frac{n - L_l - n^l_{l+1} + 1}{n}\gamma_{L_l+n^l_{l+1}/n}(x_1,\sigma^*)\\
&\geq \frac{L_l + n^l_{l+1} - 1}{n}\gamma_{L_l+n^l_{l+1}-1}(x_1,\sigma^*)\\
&\geq \gamma_{L_l+n^l_{l+1}-1}(x_1,\sigma^*) - \frac{n - L_l - n^l_{l+1} + 1}{n}\\
&\geq v^*(x_1) - 3\varepsilon_l - \frac{n - L_l - n^l_{l+1} + 1}{n^l_{l+1}}\\
&\geq v^*(x_1) - 4\varepsilon_l.
\end{aligned}$$

□

Lemma 3.4.8 *The payoff in any n-stage game stopping after more than $N(l+1, x^{l+1})$ stages after the beginning of A_{l+1} and before the end is greater than $v^*(x_1) - 4\varepsilon_l$.*

Proof: Finally if we consider a stage n in the middle of A_{l+1}, σ^* is following a good strategy until L_{l+1} and a good strategy from stage L_{l+1}, so

$$\begin{aligned}
\gamma_n(x_1,\sigma^*) &= \frac{L_{l+1} - 1}{n}\gamma_{L_{l+1}-1}(x_1,\sigma^*) + \frac{n - (L_{l+1} - 1)}{n}\gamma_{L_{l+1}/n}(x_1,\sigma^*)\\
&= \frac{L_{l+1} - 1}{n}\gamma_{L_{l+1}-1}(x_1,\sigma^*) + \frac{n - (L_{l+1} - 1)}{n}\gamma_{n-L_{l+1}+1}(x_{L_{l+1}},\sigma_{l+1}(x_{L_{l+1}})\\
&\geq \frac{L_{l+1} - 1}{n}(v^*(x_1) - 4\varepsilon_l) + \frac{n - (L_{l+1} - 1)}{n}(v^*(x^{l+1}) - \varepsilon_{l+1})\\
&\geq v^*(x_1) - 4\varepsilon_l.
\end{aligned}$$

and the expected mean payoff is greater than $v^*(x_1) - 4\varepsilon_l$. □

This is true for each $l \in \mathbb{N}$, thus the strategy σ^* is pure and 0-optimal in x_1, which concludes the proof.

3.5 Existence of the uniform value in Commutative deterministic stochastic games.

We divide the section about stochastic games and repeated games into three subsections. In the first one, we show that the classic class of absorbing game can be embedded into the class of commutative game. We show that each absorbing state can be replaced by a non absorbing state leading to some new states, which are useless from a strategic point of view but designed in order to fulfill the commutativity assumption. Then, we prove the existence of the uniform

value in stochastic games with a deterministic commutative 1-Lipschitz transition (Theorem 3.3.4). In the last subsection, we provide some generalizations.

3.5.1 Reduction of the class of absorbing games to the class of commutative games

Absorbing games were introduced by Kohlberg [Koh74]. An absorbing game is a stochastic game $\Gamma = (\{\alpha\} \cup X, I, J, q, r)$ where for each $x \in X$, x is absorbing and the payoff in x does not depend on the actions. The state α is the only one where the players have an influence on the payoff and on the future states. For each couple $(i, j) \in I \times J$, we will denote by $q(\alpha, i, j)(X)$ the total probability to reach an absorbing state by playing the couple of action (i, j).

Proposition 3.5.1 *Let $\Gamma = (\{\alpha\} \cup X, I, J, q, g)$ be an absorbing game, then there exists a commutative game $\Gamma' = (X', I, J, q', g')$ and a state $(\alpha', \alpha') \in X'$ such that for all $n \in \mathbb{N}^*$, $v_n(\alpha) = v'_n(\alpha', \alpha')$. Moreover a player can guarantee w in $\Gamma'(\alpha', \alpha')$ if and only if he can guarantee w in the original game from state α.*

Proof: We consider $\Gamma = (\{\alpha\} \cup X, I, J, q, g)$ an absorbing game with I and J disjoints and we define $a(i, j) = 1 - q(\alpha, i, j)(\alpha)$ the probability of absorption in Γ if the couple of actions (i, j) is played and $q(\alpha, i, j|X)$ the conditional probability on X if there has been absorption. We denote by $\Gamma' = (X', I', J', q', g')$ the auxiliary commutative game. The actions spaces are defined by $I' = I$ and $J' = J$. For each $i \in I$, we define a new state x_i and similarly for each $j \in J$. The state space is given by $X' = X_I \times X_J$, with $X_I = \{\alpha'\} \cup \{x_i | \forall i \in I\} \cup \{\omega\}$ and $X_J = \{\alpha'\} \cup \{x_j | \forall j \in J\} \cup \{\omega\}$ and the payoff function is defined by,

$$\begin{aligned} \forall i, i' \in I,\ j, j' \in J, \quad g'((\alpha', \alpha'), i, j) &= g(\alpha, i, j) \\ g'((x_{i'}, x_{j'}), i, j) &= I\!E_{q(\alpha, i', j'|X)}(g(x)) \\ g'((x_{i'}, \omega), i, j) &= 1 \\ g'((\omega, x_{j'}), i, j) &= 0 \\ g'((\omega, \omega), i, j) &= 1/2. \end{aligned}$$

Before the definition of q', let us precise the role of the different states by looking at the payoffs. The state (α', α') is the substitute of the state α and for each couple (i', j'), the state $(x_{i'}, x_{j'})$ replaces the absorption occurring in state α by playing the couple (i', j'). This state will not be absorbing but an equilibrium at $(x_{i'}, x_{j'})$ is to stay in this state. If player 1 deviates, with some probability the state will still be $(x_{i'}, x_{j'})$ and with some probability the new state will be $(w, x_{j'})$ where player 2 can guarantee a payoff of 0. If player 2 deviates, the new state will still be $(x_{i'}, x_{j'})$ with some probability and with some probability it will be $(x_{i'}, \omega)$ where player 1 can guarantee a payoff of 1. The transition q' is defined in three steps: we define one controlled Markov chain s_I on X_I controlled by player 1 and another one s_J on X_J controlled by player 2, then we consider s the product transition corresponding to the absorbing part of q and finally we define q'. At each step, we check that the transition is commutative. We define s_I and s_J by

$$\forall i,i'\in I,\quad s_I(\alpha,i)=x_i \qquad\qquad \forall j,j'\in J,\quad s_J(\alpha,j)=x_j$$
$$s_I(x_{i'},i)=\begin{cases}x_{i'} & \text{if } i=i'\\ \omega & \text{if } i\neq i'\end{cases} \qquad s_J(x_{j'},j)=\begin{cases}x_{j'} & \text{if } j=j'\\ \omega & \text{if } j\neq j'\end{cases}$$
$$s_I(\omega,i)=\omega \qquad\qquad s_J(\omega,j)=\omega.$$

Each player controls a separate Markov chain, so we can check the commutativity assumption separately for s_I and s_J. We prove it only for s_I. If player 1 plays twice the same action then the order is irrelevant and the formula is satisfied. If player 1 plays two different actions then the state after two stages is ω, whatever is the initial state, so the commutativity assumption is also fulfilled. Let s be defined for all $(x,y)\in X_I\times X_J$ by $s((x,y),(i,j))=(s_I(x,i),s_J(y,j))$. For a fixed couple $(i,j)\in I\times J$, the transition s on the set $\{(x_i,x_j),(x_i,\omega),(\omega,x_j),(\omega,\omega)\}$ can be written as

$$\begin{array}{cc} & \overset{j}{\downarrow} \\ \overset{i}{\rightarrow} & \left(\begin{array}{c|c|c} (\omega,\omega) & (\omega,x_j) & (\omega,\omega)\\ \hline (x_i,\omega) & \circlearrowleft & (x_i,\omega)\\ \hline (\omega,\omega) & (\omega,x_j) & (\omega,\omega)\end{array}\right) \\ & (x_i,x_j)\end{array} \qquad \begin{array}{c} \overset{j}{\downarrow} \\ \left(\begin{array}{ccc} (\omega,\omega) & (\omega,\omega) & (\omega,\omega)\\ \hline \circlearrowleft & \circlearrowleft & \circlearrowleft \\ \hline (\omega,\omega) & (\omega,\omega) & (\omega,\omega)\end{array}\right) \\ (x_i,\omega)\end{array}$$

$$\begin{array}{cc} \overset{i}{\rightarrow} & \left(\begin{array}{c|c|c} (\omega,\omega) & \circlearrowleft & (\omega,\omega)\\ (\omega,\omega) & \circlearrowleft & (\omega,\omega)\\ (\omega,\omega) & \circlearrowleft & (\omega,\omega)\end{array}\right) \\ & (\omega,x_j)\end{array} \qquad \begin{array}{c} \left(\begin{array}{ccc} \circlearrowleft & \circlearrowleft & \circlearrowleft\\ \circlearrowleft & \circlearrowleft & \circlearrowleft\\ \circlearrowleft & \circlearrowleft & \circlearrowleft\end{array}\right) \\ (\omega,\omega)\end{array}$$

Since s_I and s_J are commutative, s is also a commutative transition.
Finally let q' be defined by $q'(x,i,j)=(1-a(i,j))\delta_x+a(i,j)\delta_{s(x,i,j)}$ for all $x\in X'$, $i\in I$ and $j\in J$. Thus for all $x\in X$, $i,i'\in I$ and $j,j'\in J$, we have

$$\begin{aligned}\tilde{q'}(q'(x,i,j),i',j') &= (1-\alpha(i,j)(1-a(i',j'))\delta_x+(1-a(i,j))a(i',j')\delta_{s(x,i',j')}\\ &\quad+(1-a(i',j'))a(i,j)\delta_{s(x,i,j)}+a(i,j)a(i',j')\delta_{s(s(x,i,j),i',j')}.\end{aligned}$$

The same computation if the actions are played in the other order leads to a symmetric result except for the last term where appears $s(s(x,i',j'),i,j)$. Therefore the commutativity of s implies the commutativity of q'. Note that q' is not the product of two independent controlled Markov chains if $a:I\times J\to[0,1]$ is not the product of one function on I and one function on J.

Fix $n\in\mathbb{N}^*$. We prove that the value in α, in the absorbing game, and in (α',α'), in the commutative game, are equal. The state (ω,ω) is absorbing so the value is equal to the payoff and $v_n((\omega,\omega))=1/2$. For all i' in I, the state $(x_{i'},\omega)$ is controlled by player 1 and he can either stay in $(x_{i'},\omega)$ with a payoff of 1 or generate a random law on $(x_{i'},\omega),(\omega,\omega)$. Thus his optimal action is i' and $v_n((x_{i'},\omega))=1$. The situation is symmetric for $(\omega,x_{j'})$, so for all $j'\in J$, $v_n((\omega,x_{j'}))=0$. Let $(i',j')\in I\times J$, then i' is an optimal action for player 1 and j' is an optimal action for player 2, thus $v_n(x_{i',j'})=I\!E_{q(\alpha,i',j'|X)}(g(x))$. By replacing all these states by their

values, the situation in (α', α') is the same as in $\Gamma(\alpha)$ and the value in α and in (α', α') are equal.

Moreover if σ is a strategy for player 1 in the absorbing game, which guarantees w, we define σ' in Γ' by $\sigma'(\alpha', \alpha') = \sigma(\alpha)$ and for all $i' \in I$, $\sigma'(x_{i'}) = i'$. For all $i' \in I$ and $j' \in J$, this strategy guarantees the payoff $I\!E_{q(\alpha,i',j'|X)}(g(x))$ in the state $(x_{i'}, x_{j'})$, so it guarantees w from state (α', α'). Reciprocally if σ' guarantees w' in the commutative game, then let σ'' be the strategy in the commutative game, which plays like σ' in (α', α') and optimally outside. σ'' guarantees also w' and σ defined by $\sigma(\alpha) = \sigma'(\alpha', \alpha')$ guarantees the same payoff in the absorbing game. The two games are completely equivalent from a strategic point of view. □

3.5.2 Proof of the existence of the uniform value

In this section, we prove Theorem 3.3.4. Let $\Gamma = (X, I, J, q, g)$ be a stochastic game such that X is a compact subset of $\mathbb{R}^m$, I and J are finite sets, q is commutative, deterministic and 1-Lipschitz for $\|.\|_1$, and g is continuous. Let us prove that for all $z \in \Delta_f(X)$, the stochastic game $\Gamma(z)$ has a uniform value. Is equivalent to prove simply that for all $x \in X$, $\Gamma(x)$ has a uniform value.

We first outline the main steps of the proof. For each $x \in X$ we separate the couples of actions in two different sets. A couple of actions $(i, j) \in I \times J$ is cyclic in x if the play, obtained by repeating (i, j) starting from x, comes back in x in a finite number of stages. If (i, j) does not satisfy this property, we say that it is non-cyclic. For each x, $I \times J$ is the union of $\mathcal{C}(x)$ the set of cyclic action couples in x and of $\mathcal{NC}(x)$ the set of non-cyclic action couples in x.

We denote by Φ_k the set of states with more than k cyclic action couples and we prove that the uniform value exists for all initial points $x \in X$, by decreasing induction on the number of cyclic action couples.

The initial case is when all couples of actions are cyclic. We prove that the game $\Gamma(x)$ can be expressed with only a finite number of states and thus has a value by the result of Mertens and Neyman [MN81]. Note that this minimal set of states, necessary to formulate the game starting in x, depends on the initial state x.

For the induction step, given a state x_1 with $k-1$ cyclic action couples, we study a family $\dot{\Gamma}(\varepsilon, x_1)$ of games, which approximate $\Gamma(x_1)$ more and more precisely, and, which have a uniform value. For each $\varepsilon > 0$, the game $\dot{\Gamma}(\varepsilon, x_1)$ is defined as follows: in the neighbourhood of Φ_k, states with more than k cyclic action couples, there is absorption and the payoff is the uniform value in one state of Φ_k close to the current state, otherwise the transition and the payoff are the same as in Γ. For all $\varepsilon > 0$, we prove that $\dot{\Gamma}(\varepsilon, x_1)$ with initial state x_1, has a finite number of states and therefore a uniform value in x_1 denoted by $v(\varepsilon)(x_1)$. Finally, when ε converges to 0, we consider smaller and smaller neighbourhood of Φ_k and $v(\varepsilon)(x_1)$ has to converge to a limit value v, which is the uniform value of $\Gamma(x_1)$.

First we denote by $q_{i,j}$ the operator from X to X defined by $q_{i,j}(x) = q(i, j, x)$. The map q is deterministic, so we can define the play along a sequence of actions. Let $n \in \mathbb{N}$ and

$h = (i_1, j_1, ..., i_n, j_n) \in (I \times J)^n$, for all integers $l \leq n$, we denote $x_{l+1}(h) = q_{i_l,j_l}...q_{i_1,j_1}x_1 = \prod_{t=1}^{l} q_{i_t,j_t}x_1$ and we say that x is reachable from x_1 if there exists a play from x_1 to x.

Proposition 3.5.2 *When there are only cyclic action couples, the game has a uniform value.*

Proof: Let M be such that for all couples of actions (i, j), the play cycles in less than M stages. Let $x \in X$, we prove by contradiction that all states reachable from x can be reached in less than $(M-1)\sharp\mathcal{C}(x)$ stages. By contradiction let x^* be a state, which is not reached in $(M-1)\sharp\mathcal{C}(x)$ stages. We define

$$t^* = \inf_{t\geq 1} \left\{ t,\ \exists h = (i_l, j_l)_{l=1...t} \in (I \times J)^t,\ x_t(h) = x^* \right\}$$

the minimum number of stages needed to reach x^*. By assumption, $t^* > (M-1)\sharp\mathcal{C}(x)$ and

$$\begin{aligned} &\sum_{(i,j)\in\mathcal{C}(x)} \sharp\{l, (i_l, j_l) = (i, j)\} = t^* \\ \Rightarrow\ &\exists (i^*, j^*) \in \mathcal{C}(x)\ \sharp\{l, (i_l, j_l) = (i^*, j^*)\} \geq \frac{t^*}{\sharp\mathcal{C}(x)} \\ \Rightarrow\ &\exists (i^*, j^*) \in \mathcal{C}(x)\ \sharp\{l, (i_l, j_l) = (i^*, j^*)\} \geq M. \end{aligned}$$

So one couple of actions is repeated more than M times. By definition, there exists $d^* \leq M$ such that $q_{i,j}^{d^*}x = x$. Hence the state at stage $t^* - d$ along the sequence of actions deduced from h, by deleting d^* times the couple of actions (i^*, j^*), is x^*. This contradicts the definition of t^*. Therefore all states are reached in less than $(M-1)\sharp\mathcal{C}(x)$ stages and since I and J are finite, the game $\Gamma(x)$ can be defined only with a finite number of states. Formally it is a stochastic game with a finite set of states and finite sets of actions, thus it has a uniform value by the theorem of Mertens and Neyman [MN81]. □

We now prove the step of induction. Let $k \in \mathbb{N}$ be such that every games starting in a state with more than k cyclic action couples has a uniform value and $x_1 \in X$ a state with $k-1$ cyclic action couples. We define for each $\eta > 0$, the set of states reachable from x_1 such that there is no state with more than k cyclic action couples in the η-neighbourhood ,

$$\Phi(\eta) = \{x \text{ reachable from } x_1,\ \forall x' \in X\ s.t. \|x - x'\|_1 \leq \eta,\ \sharp\mathcal{C}(x') \leq k-1\}.$$

We now prove that this set is finite.

We deduce from a theorem of Sine[Sin90] that the play, where a non-cyclic action couple in x_1 is iterated, converges to a periodic orbit of states with more than k cyclic action couples and then the finiteness of $\Phi(\eta)$. The theorem of Sine implies immediately the following result.

Lemma 3.5.3 *Let $m \in \mathbb{N}$, there exists $f(m) \in \mathbb{R}$ such that for all maps M from $X \subset \mathbb{R}^m$ to X non-expansive for $\|.\|_1$, there exists an integer $L \leq f(m)$ and a family of maps $B_0, \cdots, B_{L-1}$ such that*

$$\forall l \in \{0, ..., L-1\}, \lim_{t \to +\infty} M^{tL+l} = B_l.$$

A classic example is the case where M is the transition of a Markov chain on a finite set. If λ is a complex eigenvalue of M then $|\lambda| \leq 1$ since the map is non-expansive. Moreover the theorem of Perron-Frobenius ensures that if $|\lambda| = 1$ then there exists $l \leq m$ such that $\lambda^l = 1$. The integer L is then the smallest common multiple and we can take $f(m) = m!$.

Applied to our framework, we deduce that, if iterated, a non cyclic action couple (i, j) becomes cyclic at the limit. Moreover previous cyclic action couples are still cyclic at the limit, by commutativity.

Lemma 3.5.4 *Let $x \in X$, $(i, j) \in \mathcal{NC}(x)$ be a couple of non-cyclic actions at x, and $\varepsilon > 0$, there exists an integer n such that*

$$\forall t \geq n, \exists x' \in X, \ \|q_{i,j}^t x - x'\|_1 \leq \varepsilon \text{ and } \sharp\mathcal{C}(x') \geq \sharp\mathcal{C}(x) + 1.$$

Proof: Let $x \in X$, $(i, j) \in \mathcal{NC}(x)$ be a couple of non cyclic actions and ε be a positive real. We show three properties: first the sequence has a finite number of limit points, then a cyclic action couple in x is still a cyclic action couple in the limit points and finally the couple (i, j) becomes cyclic in the limit points. Therefore, the number of cyclic action couples increases strictly.

By lemma 3.5.3 applied to $Q = q_{i,j}$, there exist an integer L and some operators B_0,..., B_{L-1} such that

$$\forall l \in \{0, ..., L-1\} \lim_{t \to +\infty} Q^{tL+l} = B_l.$$

Let $y = B_0 x$ then the sequence $(Q^t x)_{t \in \mathbb{N}}$ of iterated converges to the family $(Q^l y)_{l=0..L-1}$. By compacity of X, y is in X. There exists an integer n such that

$$\forall t \geq n, \|Q^{tL} x - y\|_1 \leq \varepsilon$$

and as Q is non-expansive for the norm 1, $\|Q^{tL+l} x - Q^l y\|_1 \leq \varepsilon$. We denote $n' = n(L+1)$ and we have

$$\forall t \geq n', \ \exists x' \in \{B^l x, \ l = 0, \ldots, L-1\}, \ \|x Q^t - x'\|_1 \leq \varepsilon.$$

The play has a finite number of limit points.

Let (i', j') be a couple of cyclic actions in x and d an integer such that $q_{i',j'}^d x = x$. We check

that (i', j') is still cyclic in the limit points. For all $l \in \{0, ..., 1\}$, we have

$$\begin{aligned} q^d_{i',j'} y = q^d_{i',j'} B_l x &= \lim_t q^d_{i',j'} Q^{tL+l} x \\ &= \lim_t Q^{tL+l} q^d_{i',j'} x = \lim_t Q^{tL+l} x = y, \end{aligned}$$

by commutativity. Therefore (i', j') is still a cyclic action couple on the set $\{Q^l y, l = 0..L-1\}$.

The iterated couple of actions (i, j), which was non-cyclic in x, becomes cyclic in x' for all $x' \in \{Q^l y, l = 0..L-1\}$. For all $l \in \{0, ..., 1\}$, we have

$$Q^L y = Q^L B_l x = \lim_t Q^L Q^{tL+l} x = \lim_n Q^{(t+1)L+l} x = B_l x = x.$$

All cyclic action couples are still cyclic and (i, j) becomes cyclic, so the number of cycling action couples is strictly increasing. □

Example 3.5.5 Let $X = \Delta(\mathbb{Z}/2\mathbb{Z})$, $x_1 = (1, 0)$, $I = \{i_1\}$, $J = \{j_1\}$ and

$$A = A(i_1, j_1) = \begin{pmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{pmatrix}.$$

Then for all $t \in \mathbb{N}$, $A^t x_1$ has no cyclic action couples but it converges to $x_\infty = (1/2, 1/2)$ where the couple of actions (i_1, j_1) is cyclic.

Corollary 3.5.6 *There exists a state $x \in X$ such that all the couples of actions are cyclic.*

It is immediate since the number of actions is finite. Starting from any initial state $x_1 \in X$, we apply the previous lemma to one couple of non-cyclic actions and we get a state $x_2 \in X$ with more cyclic action couples. Then we can repeat from this new state and iterate the lemma until all the couples of actions are cyclic.

Proposition 3.5.7 *For all $\eta > 0$, the set $\Phi(\eta)$ is finite.*

Proof: Let $H = \{h \in (I \times J)^{\mathbb{N}} |\ \exists t \geq 1, x_t(h) \in \Phi\}$ be the set of possible histories associated to states in Φ. For all $x \in \Phi$, we denote $t^*(x) = \inf\{t |\ \exists h \in (I \times J)^t\ x_t(h) = x\}$, the least number of stages necessary to reach x. Let us prove that the set $F = \{t^*(x) | x \in \Phi\}$ of minimal reached time is finite. Since at each stage there exists a finite number of actions, it would imply that the set Φ is finite. For each couple of actions (i, j) in $\mathcal{NC}(x_1)$, we denote by $u(i, j)$ the integer given by lemma 3.5.4. Since there is a finite number of couple of actions, there exists M' an integer such that for all $(i, j) \in \mathcal{NC}(x_1)$, $u(i, j) \leq M'$ and for all $(i, j) \in \mathcal{C}(x_1)$, the minimal period of (i, j) is smaller than M' and we prove that F is bounded by $N = M' \sharp (I \times J)$. Let $x \in \Phi$ be such that $t^* = t^*(x) \geq N$ and h be an history associated to x and t^*, then one action (i^*, j^*) is repeated more than M' times and this action is either cyclic or not cyclic. If this action is

cyclic, the history can be shortened, as in the proof of proposition 3.5.2, which is absurd with respect to the definition of t^*. If this action is not cycling, there exists $\overline{x} \in X$ such that

$$\begin{aligned} \|q_{i^*,j^*}^{M'} x_1 - \overline{x}\|_1 &\leq \varepsilon \\ \text{and } \sharp\mathcal{C}(\overline{x}) &> k-1 \end{aligned}$$

But the transition is non-expansive and $\mathcal{C}$ is increasing along the orbits. So if we denote by h' the sequence of actions where (i^*, j^*) has been deleted M' times, and we define x' the state obtained from $\overline{x}$ by playing h', we have

$$\begin{aligned} \|x - x'\|_1 &\leq \varepsilon, \\ \text{and } \sharp\mathcal{C}(x') &> k-1, \end{aligned}$$

which contradicts the definition of x. Thus there exists an integer M'' such that each state can be reached in less than M'' stages. At each stage, the number of actions is finite, so $\Phi(\eta)$ is finite. □

We now define the auxiliary game by choosing for each $\varepsilon > 0$, an η-neighbourhood small enough. First we check that the 1-Lipschitz transition and the uniform continuity of the payoff imply the continuity of the maximal payoff that a player can guarantee, then we describe the family of auxiliary games and conclude the proof.

Lemma 3.5.8 *Given $x \in X$ and $\varepsilon > 0$, there exists $\eta > 0$ such that if player 1 guarantees w in $\Gamma(x')$ then for all x, such that $\|x - x'\|_1 \leq \eta$, he guarantees $w - \varepsilon$ in $\Gamma(x)$.*

Proof: Given $\varepsilon > 0$, for all $(i, j) \in I \times J$ the map $g(., i, j)$ is uniformly continuous. Moreover the number of maps is finite, so there exists $\eta > 0$ such that for all $x, x' \in X$ with $\|x - x'\|_1 \leq \eta$, we have

$$\forall (i,j) \in (I \times J),\ |g(x,i,j) - g(x',i,j)| \leq \varepsilon.$$

Let $\sigma \in \Sigma$ be a pure strategy, we define the strategy σ^*, which plays as if the game were $\Gamma(x)$ whatever is the initial state. Especially this strategy does not depend on the state and only on the actions. Let $\tau \in J^{\mathbb{N}}$ be a sequence of actions of player 2.

We denote by x_t the state at stage t along (x, σ^*, τ) and x'_t the state at stage t along (x', σ^*, τ). For all $(i, j) \in I \times J$, q is a non-expansive function so for all $n \in \mathbb{N}$, $\|x_t - x'_t\|_1 \leq \|x - x'\|_1 \leq \eta$ and

$$\begin{aligned} |\gamma_n(x, \sigma^*, \tau) - \gamma_n(x', \sigma^*, \tau)| &\leq \frac{1}{n} \sum_{t=1}^{n} |g(x_t, i_t, j_t) - g(x'_t, i_t, j_t)| \\ &\leq \varepsilon. \end{aligned}$$

Given a mixed strategy σ, we define the strategy σ^* by associating to each pure strategy

with positive probability, the one defined in the previous paragraph. We then have

$$|\gamma_n(x,\sigma^*,\tau)-\gamma_n(x',\sigma^*,\tau)| \le I\!E_\sigma\left(\frac{1}{n}\sum_{t=1}^{n}|g(x_t,i_t,j_t)-g(x'_t,i_t,j_t)|\right)$$
$$\le \varepsilon.$$

If player 1 guarantees w in $\Gamma(x')$ then he guarantees $w-\varepsilon$ in the game $\Gamma(x)$. □

Let $\varepsilon>0$ and η be given by lemma 3.5.8. By proposition 3.5.7, the set of states reachable from x_1 and at more than η of any state with more than k cyclic action couples, i.e. $\Phi(\eta)$, is finite. We denote by $q(\Phi(\eta))$ the set of all states obtained by one transition from one of these states and, which are not already in $\Phi(\eta)$. $q(\Phi(\eta))$ is finite and for each $x\in q(\Phi(\eta))$, there exists $\xi(x)$ such that $d(x,\xi(x))\le\eta$ and the game $\Gamma(\xi(x))$ has a uniform value denoted by $v^*(\xi(x))$, by the induction assumption. We define the auxiliary game $\dot\Gamma(\varepsilon,x_1)$ as follows: the initial state is x_1, the set of actions are I et J and the transition and reward functions are given by:

$$\dot q(x,i,j)=\begin{cases} q_{i,j}x & \text{if } x\in\Phi(\eta)\\ x & \text{if } x\in q(\Phi(\eta))\\ x & \text{otherwise ,}\end{cases}$$

$$\text{and } \dot r(x,i,j)=\begin{cases} g(x,i,j) & \text{if } x\in\Phi(\eta)\\ v^*(\xi(x)) & \text{if } x\in q(\Phi(\eta))\\ 0 & \text{otherwise .}\end{cases}$$

The sets of strategy for player 1 and 2 are the same as in the game Γ. In the game starting in x_1, all the states are in $\Phi(\eta)$ or $q(\Phi(\eta))$. Since both sets are finite, this game is formally a stochastic game with a finite set of states and finite sets of actions. Therefore $\dot\Gamma(\varepsilon,x_1)$ has a uniform value by the theorem of Mertens and Neyman [MN81].

Proposition 3.5.9 *$\dot\Gamma(\varepsilon,x_1)$ has a uniform value in x_1 denoted by $v^*(\varepsilon)(x_1)$.*

Moreover let us check that the value of the auxiliary game is a good approximation of what the players can guarantee in $\Gamma(x_1)$.

Proposition 3.5.10 *If player 1 can guarantee w in $\dot\Gamma(\varepsilon,x_1)$ then he can guarantee $w-3\varepsilon$ in $\Gamma(x_1)$.*

Proof: By assumption, there exists $\dot\sigma$ a strategy of player 1 in $\dot\Gamma(\varepsilon,x_1)$ and a stage $\dot N$ such that

$$\forall n\ge\dot N,\ \forall\dot\tau\ \dot\gamma_n(x_1,\dot\sigma,\dot\tau)\ge w-\varepsilon.$$

For each state $x\in q(\Phi(\eta))$, we denote by $\sigma^{\xi,x}$ the strategy given by proposition 3.5.10 with respect to the point $\xi(x)$ and to an ε-optimal strategy in $\Gamma(\xi(x))$ such that

$$\exists N(x),\ \forall n\ge N(x),\ \forall\tau,\ \gamma_n(x,\sigma^{\xi,x},\tau)\ge v(\xi(x))-2\varepsilon.$$

Let $\overline{N} = \max(N(x), x \in \Phi(\eta))$ be an upper bound.

Let θ be a function from $(X \times I \times J)^{\mathbb{N}}$ to $\mathbb{N}$, giving the first time where the distance between the state and a state with more than k cyclic action couples, is less than η.

$$\theta(h) = \inf_{t \geq 1}\{t | x_t(h) \in q(\Phi(\eta))\},$$

We define the strategy σ, which plays optimally in $\dot{\Gamma}$ until a state $x' \in q(\Phi(\eta))$, and then optimally as if the remaining game started from $\xi(x')$:

$$\sigma_n(h) = \begin{cases} \dot{\sigma}_n(h) & \text{if } n \leq \theta(h) \\ \sigma_{n-\theta(h)}^{\xi, x_{\theta(h)}(h)} & \text{if } n > \theta(h). \end{cases}$$

Let us show that this strategy guarantees $w - 3\varepsilon$. If τ is a strategy of player 2, we denote by x_t the state at stage t and we define $\widetilde{\theta}$ a stopping time checking the exit of $\Phi(\eta)$:

$$\widetilde{\theta} = \inf_{t \geq 1}\{t | x_t \in q(\Phi(\eta))\}.$$

Let $N^* \in \mathbb{N}$ such that $N^* \geq \dot{N}$ and $\frac{\overline{N}}{N^*} \leq \varepsilon$. For all $n \geq N^*$, for each history there are two cases. On one hand if $n - \theta(h) > \overline{N}$, the definition of $\overline{N}$ implies that σ has played optimally in the game from the state $\xi(x_\theta)$ for long enough in order for the payoff to be above $v^*(\xi(x_\theta)) - \varepsilon$. On the other hand if $n - \theta(h) < \overline{N}$, then the part of the play after $\theta(h)$ weights for less than ε of the total. We split the payoff depending on this criteria, and we study both parts separately. We first focus on histories where the switch occurred early: $n - \widetilde{\theta} \geq \overline{N}$. Since $\|x_{\widetilde{\theta}} - \xi(x_{\widetilde{\theta}})\| \leq \eta$, if σ^{h_n} and τ^{h_n} are the strategies induced by σ and τ after $\widetilde{\theta}$ given h_n, σ guarantees in average the value $v(\xi(x_{\widetilde{\theta}})) - 2\varepsilon$ between $\widetilde{\theta}$ and N,

$$\begin{aligned} I\!E_{x,\sigma,\tau}\left(\sum_{t=\widetilde{\theta}+1}^{n} g(x_t, i_t, j_t)\right) &= I\!E_{x,\sigma,\tau}\left(\gamma_{n-\widetilde{\theta}}(x_{\widetilde{\theta}}, \sigma^{h_{\widetilde{\theta}}}, \tau^{h_{\widetilde{\theta}}})(n - \widetilde{\theta})\right) \\ &\geq I\!E_{x,\sigma,\tau}\left(\left(v(\xi(x_{\widetilde{\theta}})) - 2\varepsilon\right)(n - \widetilde{\theta})\right), \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{n} I\!E_{x,\sigma,\tau}\left(\sum_{t=1}^{n} g(x_t, i_t, j_t)\mathbf{1}_{n-\widetilde{\theta} \geq \overline{N}}\right) \\ &= \frac{1}{n} I\!E_{x,\sigma,\tau}\left(\sum_{t=1}^{\widetilde{\theta}} g(x_t, i_t, j_t) + \sum_{t=\widetilde{\theta}+1}^{n} g(x_t, i_t, j_t)\right) \\ &\geq I\!E_{x,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\widetilde{\theta}} g(x_t, i_t, j_t) + v(\xi(x_{\widetilde{\theta}}))(n - \widetilde{\theta})\right)\mathbf{1}_{n-\widetilde{\theta} \geq \overline{N}} - 2\varepsilon \mathbf{1}_{n-\widetilde{\theta} \geq \overline{N}}\right). \end{aligned}$$

If we consider the other set of histories, the following lower bound is always true since the payoffs are in $[0,1]$:

$$\forall x \in X, i \in I,\ j \in J,\ g(x,i,j) \geq v(\xi(x)) - 2,$$

and on these histories, $\frac{n-\widetilde{\theta}}{N} \leq \frac{\overline{N}}{N} \leq \varepsilon$, thus

$$\begin{aligned}
&\frac{1}{n}I\!E_{x,\sigma,\tau}\left(\sum_{t=1}^{n} g(x_t,i_t,j_t)\mathbf{1}_{n-\widetilde{\theta}<\overline{N}}\right)\\
&= I\!E_{x,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{n=1}^{\widetilde{\theta}} g(x_t,i_t,j_t) + \sum_{t=\widetilde{\theta}+1}^{n} g(x_t,i_t,j_t)\right)\mathbf{1}_{n-\widetilde{\theta}<\overline{N}}\right)\\
&\geq I\!E_{x,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\widetilde{\theta}} g(x_t,i_t,j_t) + v(\xi(x_{\widetilde{\theta}}))(n-\widetilde{\theta}) - 2(n-\widetilde{\theta})\right)\mathbf{1}_{n-\widetilde{\theta}<\overline{N}}\right)\\
&\geq I\!E_{x,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\widetilde{\theta}} g(x_t,i_t,j_t) + v(\xi(x_{\widetilde{\theta}}))(n-\widetilde{\theta})\right)\mathbf{1}_{n-\widetilde{\theta}<\overline{N}} - 2\varepsilon\mathbf{1}_{t-\widetilde{\theta}<\overline{N}}\right).
\end{aligned}$$

Therefore by summing the two inequalities, we get the result

$$\gamma_n(x,\sigma,\tau) \geq \dot{\gamma}_n(x,\dot{\sigma},\tau) - 2\varepsilon \geq w - 3\varepsilon.$$

□

To conclude the proof, we show that $v(\varepsilon)(x_1)$, the value of the auxiliary game $\dot{\Gamma}(x_1,\varepsilon)$, converges when ε converges to 0, and the limit is the value of the game $\Gamma(x_1)$. By proposition 3.5.10, for all $\varepsilon > 0$, player 1 can guarantee $v(\varepsilon)(x_1) - 3\varepsilon$ in the game $\Gamma(x_1)$. So he can guarantee the superior limit when ε converges to 0: for all $\delta > 0$, there exists n_1 and a strategy $\sigma^* \in \Sigma$ such that for all $\tau \in \mathcal{T}$, for all $n' \geq n_1$,

$$\gamma_{n'}(x_1,\sigma^*,\tau) \geq \limsup v(\varepsilon)(x_1) - \delta.$$

The same argument shows that player 2 can guarantee the inferior limit. Therefore for all $\delta > 0$, there exists n_2 and a strategy $\tau^* \in \mathcal{T}$ such that for all $\sigma \in \Sigma$, for all $n' \geq n_2$,

$$\gamma_{n'}(x_1,\sigma,\tau^*) \leq \liminf v(\varepsilon)(x_1) + \delta.$$

Thus given $\delta > 0$ and $n' \geq \max(n_1,n_2)$, we have

$$\limsup v(\varepsilon)(x_1) - \delta \leq \gamma_{n'}(x_1,\sigma^*,\tau^*) \leq \liminf v(\varepsilon)(x_1) + \delta,$$

so $v(\varepsilon)(x_1)$ converges and the limit is the uniform value of the game $\Gamma(x_1)$. This proves the induction hypothesis at the next step and finishes the proof. For all $x \in X$, the game $\Gamma(x)$ has a uniform value and for all initial probability with finite supports, $z \in \Delta_f(X)$, $\Gamma(z)$ has a uniform value.

3.5.3 Extensions.

The proof of Theorem 3.3.4 can be extended by replacing some of the lemmas with more general results. The result of Sine [Sin90], for example, applies to more general norms than the norm $\|.\|_1$.

Definition 3.5.11 *A norm on $\mathbb{R}^n$ is polyhedral if the unit ball has a finite number of extreme points.*

For example the norm $\|.\|_1$ and the sup norm are polyhedral norms but not the euclidean norm. For polyhedral norm, the application of the theorem of Sine [Sin90] to compact sets gives the following results,

Lemma 3.5.12 *Let $N(.)$ be a polyhedral norm and $K \subset \mathbb{R}^m$ be a compact set. There exists $\phi(N,m)$ such that for all mappings T non-expansive for N, there exists $t \leq \phi(N,m)$ such that $(T^{tn})_{n\in\mathbb{N}}$ converges.*

Theorem 3.5.13 *Let $\Gamma = (X, I, J, q, g)$ be a stochastic game, such that X is a compact set of $\mathbb{R}^m$, I and J are finite sets, q is commutative deterministic non-expansive for a polyhedral norm, and g is continuous. For all $z \in \Delta_f(X)$, the stochastic game $\Gamma(z)$ has a uniform value.*

This theorem does not apply to the Example 3.2.2 on the circle and that the existence of a uniform value in this model is still an open question.

We can also change the result by replacing the theorem from Mertens and Neyman [MN81] with other existence results. First, Vieille [Vie00a][Vie00b] proves the existence of an equilibrium payoff in every two-player stochastic games. So our proof, adapted to the non zero-sum case leads to the following result :

Theorem 3.5.14 *Let $\Gamma = (X, I, J, q, g_1, g_2)$ be a two-player non zero-sum stochastic game such that X is a compact subset of $\mathbb{R}^m$, I and J are finite sets of actions, q is commutative deterministic non-expansive for $\|.\|_1$ and g_1 and g_2 are continuous. Then, for all $z \in \Delta_f(X)$, the stochastic game $\Gamma(z)$ has an equilibrium payoff.*

Secondly, there exist some specific classes of m-player stochastic games where the existence of an equilibrium has been proven. For example, Flesch, Schoenmakers and Vrieze [FSV08][FSV09] prove the existence of an equilibrium for m-player stochastic games where each player controls a finite Markov chain and the payoffs depend on the m states and the m actions at stage n. Note that the commutativity assumption here is reduced to a condition player by player. As in our proof, the commutativity assumption implies that we can study deterministic transitions non-expansive for the norm $\|.\|_1$.

Theorem 3.5.15 *Let $\Gamma = \Big((X_j, I_j, q_j)_{j\in\{1,\dots,m\}}, g\Big)$ be a m-player product-state space stochastic game such that for all $j \in \{1, ..., m\}$, X_j is a compact subset of $\mathbb{R}^{m_j}$, I_j is a finite set of actions, q_j is commutative deterministic non-expansive for $\|.\|_1$ and $g : \prod(X_j \times I_j) \to [0,1]^m$ is continuous. For all $z \in \Delta_f(\prod_j X_j)$, the stochastic game $\Gamma(z)$ has an equilibrium payoff.*

Acknowledgements. I thank J.Flesch, S. Gaubert, J. Renault, S.Sorin and two anonymous referees for their comments on this article. The suggestions they provided were extremely helpful.

Chapter 4

Repeated games with a more informed controller

Résumé : Renault [Ren12b] donne un ensemble de conditions pour l'existence de la valeur uniforme dans un jeu stochastique avec un ensemble d'états compact, des ensembles d'actions compacts et une transition controlée par un joueur. Il utilise ensuite ce résultat pour prouver l'existence de la valeur uniforme dans les jeux répétés avec un contrôleur parfaitement informé. On montre que la valeur uniforme existe si le contrôleur n'est pas parfaitement informé mais simplement plus informé que le joueur 2. La démonstration suit les grandes lignes de la preuve de Renault [Ren12b] et utilise son théorème sur les jeux stochastiques avec un espace d'états compact.

Ce chapitre est extrait d'un article écrit en collaboration avec Miquel Oliu-Barton et Fabien Gensbittel.

Abstract: Renault [Ren12b] gives a set of sufficient conditions for the existence of the uniform value in a stochastic game with a compact state space, compact actions spaces and a transition controlled by one player. Then, he uses this result in order to deduce the existence of the uniform value in repeated games with an informed controller. We show that the uniform value exists if the controller is only more informed than the second player. The proof follows a similar approach to Renault [Ren12b] and uses his result on stochastic games with a compact state space.

This chapter is extracted from an article written in collaboration with Miquel Oliu-Barton and Fabien Gensbittel.

4.1 Introduction

Finite stochastic games were introduced by Shapley [Sha53] in 1953 in order to study repeated interaction between two players. The first focus was set on a discounted evaluation of the payoff and Shapley proved the existence of the value. The focus was then extended by Gillette [Gil57] with another way to evaluate the payoff: the undiscounted version where we consider as a payoff the limit of the Cesàro mean which stayed an open problem for 20 years. The breakthrough was done by two works. First Bewley and Kohlberg [BK76b][BK76a] showed that the discounted value is an algebraic mapping of λ the discount factor. Moreover it implies that the discounted value converges when λ goes to 0 and that the value of the n-stage games converges also to the same limit when their length goes to infinity. Then Mertens and Neyman [MN81] build some uniform strategies which guarantee the limit. In order to play uniformly, at each stage a player play optimally in a discounted game but the discount factor changes at each stage depending on the new state and the previous payoffs.

Simultaneously stochastic games appeared as a tool in order to understand models of repeated games with incomplete information introduced by Aumann and Maschler [AMS95]. In this model, at stage 1 a matrix game is chosen according to a probability. Whereas one player is informed of the matrix game played, the other one only knows the probability. They show that these games have a uniform value which was latter interpreted as the value of an auxiliary stochastic game with an underlying Borel state space: the space of beliefs of the uninformed player. This model has then been generalized to different models especially with signals on the actions or incomplete informations on both sides.

The proximity and the links between both types of models lead naturally to a general class of repeated games where a stochastic game is played and players have private information. This information can concern either the state or past actions. Kohlberg [Koh74] showed that private monitoring of actions leads in general to the absence of the uniform value. Additional work on signalling about actions has been done in the framework of absorbing game by Coulomb [Cou92], [Cou03] or Rosenberg, Solan and Vieille [RSV03]. The problem of information about the state has first been formulated with a collection of stochastic games: at stage 0 a stochastic game is chosen according to a probability and the players receive some information then the stochastic game is played normally. The existence of the uniform value has been shown in several cases: when the information is symmetric by Geitner [Gei02] or when one player is perfectly informed and controls the transition by Rosenberg, Solan and Vieille [RSV04]. One can also directly consider a stochastic game and assume that players have some partial information on the state. Renault [Ren06] proved the existence of the uniform value for Markov chain games where nobody controls the transition of the stochastic game, then he extended the existence to the case of an informed controller (Renault [Ren12b]) by using the existence of the uniform value in dynamic programming problems (Renault [Ren11]). Note that in this latter article he also proves the existence of the uniform value for Markov Decision Processes with partial observation.

The aim of this article is to relax the assumption that the controller is fully informed. By allowing the controller to be partially informed, we especially include the case where the information is symmetric and Markov Decision Process with partial observation. The first part of the article is dedicated to the model and to define properly the notion of a better informed controller. Then we state the result and in the second part we prove the existence of the uniform value. The proof uses the main theorem of Renault [Ren12b] which gives sufficient conditions for the existence of the uniform

value. We introduce an auxiliary stochastic game on an auxiliary state variable which is the belief of player 2 about the belief of player 1. This game has a uniform value and moreover player 1 can use an ε-optimal strategy of the auxiliary game in order to play optimally in the original game. Finally we prove that player 2 can also guarantee this value by playing by blocks and thus the original game has a uniform value.

4.2 General model

A two-player zero-sum repeated game Γ is defined by a 7-tuple $\Gamma = (K, I, J, C, D, q, g)$, where K is a finite set of states, C (resp. D) is a finite set of signals for player 1 (resp. 2), I and J are finite sets of actions for player 1 and 2 respectively, $g : K \times I \times J \to [0, 1]$ is the payoff function and $q : K \times I \times J \to \Delta(K \times C \times D)$ is the transition function. Both mappings can be naturally extended to $\Delta(I) \times \Delta(J)$.

Given a probability $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$, the game $\Gamma(\pi)$ is played by stages as following. At stage 1 a triple (k_1, c', d') is chosen according to π. Player 1 is informed of c' and player 2 of d'. Then both players choose independently and simultaneously respectively an action i_1 and j_1. Player 1 obtains an unobserved payoff $r(k_1, i_1, j_1)$ and player 2 the opposite $-r(k_1, i_1, j_1)$. Then a new triple (k_2, c_1, d_1) is chosen accordingly to $q(k_1, i_1, j_1)$. Player 1 observes c_1 whereas player 2 observes d_1 and the game proceeds to the next stage. At stage $t \geq 1$ the current state is k_t and the players receive respectively c_{t-1} and d_{t-1}, they choose simultaneously some actions i_t and j_t, and a new triple (k_{t+1}, c_t, d_t) is chosen according to $q(k_t, i_t, j_t)$.

Thus at stage $t \geq 1$ the current state is k_t and the information held by the players is $h_t^1 = (c', i_1, c_1, .., i_{t-1}, c_{t-1})$ (resp. $h_t^2 = (d', j_1, d_1, .., j_{t-1}, d_{t-1})$) for player 1 (resp. 2). That is, every player remembers his own past signals and actions. The set of finite histories for player 1 is $H^1 = \bigcup_{t\geq 1} H_t^1$, where $H_t^1 = \mathbb{N} \times (I \times C)^{t-1}$ (resp. H^2 for player 2). Notice that after stage $t \geq 1$, the signal c_t (resp. d_t) contains essentially two types of information: (1) information about the opponent's past move j_t (resp. i_t), (2) information about the current state k_t. A behavior strategy for player 1 (resp. 2) is a mapping $\sigma : H^1 \to \Delta(I)$ (resp. $\tau : H^2 \to \Delta(J)$). We consider the discrete topology on the sets K,I, J, C and D and the product topology on the set of infinite histories. Together with the initial lottery π and with the transition function q, a strategy profile (σ, τ) induces probabilities on the set of finite histories with respect to the Borelian algebra. It can be extended by Kolmogorov extension theorem in a unique probability distribution over $K \times \mathbb{N} \times \mathbb{N} \times (K \times I \times C \times J \times D)^{\mathbb{N}}$ with respect to the Borelian algebra, that we call $\mathbb{P}_{\pi,\sigma\tau}$. The set of strategies are denoted by Σ and $\mathcal{T}$, and the payoff is given by

Definition 4.2.1 *For any $\theta \in \Delta(\mathbb{N}^*)$, we define the payoff under evaluation θ as*

$$\gamma_\theta(\pi, \sigma, \tau) \quad = \quad \mathbb{E}_{\pi,\sigma,\tau}[\sum_{t\geq 1} \theta_t g(k_t, i_t, j_t)] \tag{4.1}$$

We denote by $\Gamma_\theta(\pi)$ the 8-tuple defined above together with an evaluation θ and an initial distribution π. For any $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$ and any $\theta \in \Delta(\mathbb{N}^*)$, it is well known that $\Gamma_\theta(\pi)$ has a value denoted by $v_\theta(\pi)$. The classical evaluations of the n-stage repeated game or of the λ-discounted game belong to this approach and we denote by $v_n(\pi)$ and $v_\lambda(\pi)$ the corresponding values. Here, the initial lottery π is thought of as the parameter of the game.

Definition 4.2.2 *Let v be a real number,*

- *Player 1 can guarantee v in $\Gamma(\pi)$ if for any $\varepsilon > 0$ there exists a strategy $\sigma \in \Sigma$ of player 1 and an integer $N \in \mathbb{N}$, such that*

$$\forall n \geq N,\ \forall \tau \in \mathcal{T},\ \gamma_n(\pi, \sigma, \tau) \geq v - \varepsilon.$$

  We say that such a strategy σ guarantees $v - \varepsilon$ in $\Gamma(\pi)$.
- *Player 2 can guarantee v in $\Gamma(\pi)$ if for any $\varepsilon > 0$ there exists a strategy $\tau \in \mathcal{T}$ of player 2 and an integer $N \in \mathbb{N}$, such that*

$$\forall n \geq N,\ \forall \sigma \in \Sigma,\ \gamma_n(\pi, \sigma, \tau) \leq v + \varepsilon.$$

  We say that such a strategy τ guarantees $v + \varepsilon$ in $\Gamma(\pi)$.
- *If both players can guarantee the same value, the game has a uniform value denoted by $v^*(\pi)$.*

For any behavior strategy for player 1 (resp. 2) σ (resp. τ) and any $c' \in \mathbb{N}$ (resp. $d' \in \mathbb{N}$), let $\sigma(c')$ (resp. $\tau(d')$) denote the strategy after stage 1, that is once the private signal c' (resp. d') has been revealed to him (or equivalently the restriction of the map σ to the subset of histories beginning by c'). These strategies may be interpreted as strategies in the game where players have no initial signals, more formally for $\pi = \delta_{(k,c',d')}$ or $\pi' = p \otimes \delta_{(c',d')}$, we will write

$$\gamma_\theta(k, \sigma(c'), \tau(d')) \triangleq \gamma_\theta(\pi, \sigma, \tau) \quad \gamma_\theta(p, \sigma(c'), \tau(d')) \triangleq \gamma_\theta(\pi', \sigma, \tau).$$

The payoff can then be written as

$$\gamma_\theta(\pi, \sigma, \tau) = \int_{K \times \mathbb{N} \times \mathbb{N}} \gamma_\theta(k, \sigma(c'), \tau(d'))d\pi(k, c', d').$$

4.3 Model with a more informed player

The initial distribution are probabilities with finite support over $K \times \mathbb{N} \times \mathbb{N}$. Given an initial distribution, we can replace $\mathbb{N}$ and $\mathbb{N}$ by two finite subsets C' and D' of $\mathbb{N}$. We will also write abusively that $\pi \in \Delta(K \times C' \times D')$, and the initial signals will be denoted (c', d'). Reciprocally, given finite sets C' and D', any $\pi \in \Delta(K \times C' \times D')$ can be seen as an element of $\Delta_f(K \times \mathbb{N} \times \mathbb{N})$ using some enumerations of C' and D'. The sets C' and D' are not fixed but only a convenient abstract notation, their precise definition will depend on the context and they will especially grow along the game but always be finite.

Our aim is to state rigorously the assumptions: $(A1)$ the first player always has a more accurate information than player 2 about the state variable, $(A2)$ he can compute the information of player 2, and $(A3)$ finally he controls the evolution of their beliefs. We then prove that stochastic games satisfying the three assumptions have a uniform value. Let us define x_t the posterior belief[4] of player 1 about the state at stage t, z_t the posterior belief of player 2 about the belief of player 1 and η_t its law. The beliefs of the players about the state are called first-order beliefs whereas the beliefs of one

4. Formally, these are conditional laws and the word beliefs has to be understood as posterior beliefs when knowing the strategies used, since otherwise these "beliefs" are not necessarily accessible.

player about the first-order belief of the other are called second-order beliefs. The first and the second assumptions will imply that x_t, z_t and η_t are the key variables and that we can introduce a stochastic game on $\Delta_f(\Delta(K))$. The last assumption ensures that we can solve this new stochastic game. After the theorem we give a set of stronger, but easier to check, assumptions.

Let us state at first a very useful definition for the sequel, and the formal definition of x_t,z_t and η_t.

Definition 4.3.1 *For any random variable U defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathcal{F}$ a sub σ-algebra of $\mathcal{A}$, $\mathcal{L}_{\mathbb{P}}(U \mid \mathcal{F})$ denotes the conditional law of U given $\mathcal{F}$ which is seen as a $\mathcal{F}$-measurable random variable*[5] *and $\mathcal{L}_{\mathbb{P}}(U)$ the distribution of U.*

Definition 4.3.2 *Given $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$ and a pair (σ, τ) of behavior strategies, we define for $t \in \mathbb{N}^*$,*

$$x_t := \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1) \in \Delta(K),$$

$$z_t = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}\left(\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t|h_t^1)|h_t^2\right) \in \Delta_f(\Delta(K)),$$

and

$$\eta_t = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(z_t) \in \Delta_f(\Delta_f(\Delta(K))).$$

4.3.1 Player 1 is better informed than player 2

We first present an abstract version $(A1)$, which is more intuitive. Then we give an equivalent formulation on the initial probability $(A1a)$ and the transition $(A1b)$.

(A1) Player 1 is always better informed than player 2 about the state variable:

$$\forall t \in \mathbb{N}^*, \forall \sigma \in \Sigma, \forall \tau \in \mathcal{T}, \qquad \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1, h_t^2) = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_t \mid h_t^1).$$

This equation is equivalent to the conditional independence of k_t and h_t^2 given h_t^1 under the probability $\mathbb{P}_{\pi,\sigma,\tau}$. It means that the information of player 2 does not contain any information about the state variable that is not already contained in the information of player 1. Clearly stochastic games as in the original model of Shapley [Sha53] or repeated games with incomplete information on one side as in Aumann and Maschler [AMS95] fulfill this assumption since player 1 always has all the information available to player 2. This assumption is also satisfied by repeated games with incomplete information on one-and-a-half side studied by Sorin and Zamir [SZ85].

Example 4.3.3 We consider a stochastic game (K, I, J, C, D, q, g) such that $K = \{\alpha, \beta\}$, I and J are finite, $C = K$ and $D = \{d_1, d_2\}$. The payoff function is anything and the transition q does not depend on the state or the action and is given by

5. All random variables appearing here take only finitely many values so that the definition of such conditional laws does not require any additional care about measurability.

$$
\begin{array}{cccc}
 & & D & D \\
 & & d_1 \quad d_2 & d_1 \quad d_2 \\
C & \alpha & \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\
 & \beta & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \\
 & & \alpha & \beta \\
 & & K &
\end{array}
$$

This game is the repetition of a one stage game with incomplete information on one-and-a-half side. At each stage, player 1 learns the state but does not know what is the belief of player 2. He is better informed but he is not able to compute the belief of player 2.

Assumption $(A1)$ can be reformulated as a couple of assumptions $(A1a)$ and $(A1b)$ which treat directly with π, i.e. the initial information, and q, i.e. the transition function for stages $t \geq 2$ respectively.

(A1a) The probability $\pi \in \Delta(K \times C' \times D')$ is such that

$$\forall k, c', d' \in K \times C' \times D', \qquad \pi(c')\pi(k, c', d') = \pi(k, c')\pi(c', d') \tag{4.2}$$

and

(A1b) There exists a map F from $\Delta(K) \times I \times C$ to $\Delta(K)$ such that

$$\forall p, i, j, c, d, k \in \Delta(K) \times I \times J \times C \times D \times K, \quad q(p, i, j)[k, c, d] = F(p, i, c) \sum_{k' \in K} q(p, i, j)[k', c, d] \tag{4.3}$$

Lemma 4.3.4 *Assumptions* $(A1)$ *and* $(A1a + A1b)$ *are equivalent. Furthemore, the map* F *from* $\Delta(K) \times I \times C$ *to* $\Delta(K)$ *defined in* $(A1b)$ *is such that for all* $t \geq 2$,

$$x_t = F(x_{t-1}, i_{t-1}, c_{t-1}) \quad \mathbb{P}_{\pi,\sigma,\tau}\text{-almost surely.}$$

Proof: Using the characterization of conditional independence, assumption $(A1)$ for $t = 1$ is equivalent to $(A1a)$. Thus we have to prove that $(A1)$ for $t \geq 2$ implies $(A1b)$ and the converse. Assume that σ_1, τ_1 have full support. By construction, we have

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_2, c_1, d_1 \mid k_1, c', i_1, d', j_1) = q(k_1, i_1, j_1).$$

It follows, using the tower property of conditional expectation that

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_2, c_1, d_1 \mid c', i_1, d', j_1) = q(x_1, i_1, j_1),$$

where x_1 can be written as a function of c' using $(A1a)$. By disintegration, we obtain

$$\mathbb{P}(k_2 = k \mid c_1, d_1, c', i_1, d', j_1)(\sum_{\widetilde{k} \in K} q(x_1, i_1, j_1)[\widetilde{k}, c_1, d_1]) = q(x_1, i_1, j_1)[k, c_1, d_1].$$

The conditional law $\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(k_2 \mid c', i_1, c_1)$ has the following expression

$$\begin{aligned}\mathbb{P}(k_2 = k \mid c', i_1, c_1)\Big(\sum_{\widetilde{k},\widetilde{d'},\widetilde{d_1},\widetilde{j_1}} \pi(c', \widetilde{d'})\tau_1(\widetilde{d'})[\widetilde{j_1}]q(x_1(c'), i_1, \widetilde{j_1})[\widetilde{k}, c_1, \widetilde{d_1}]\Big)\\ = \sum_{\widetilde{d'},\widetilde{d_1},\widetilde{j_1}} \pi(c', \widetilde{d'})\tau_1(\widetilde{d'})[\widetilde{j_1}]q(x_1(c'), i_1, \widetilde{j_1})[k, c_1, \widetilde{d_1}].\end{aligned}$$

Assumption $(A1)$ for $t = 2$ implies that these two conditional probabilities are equal, which in turn implies

$$\frac{q(x_1, i_1, j_1)[k, c_1, d_1]}{\sum_{\widetilde{k}\in K} q(x_1, i_1, j_1)[\widetilde{k}, c_1, d_1]} = \frac{\sum_{\widetilde{d'},\widetilde{d_1},\widetilde{j_1}} \pi(c', \widetilde{d'})\tau_1(\widetilde{d'})[\widetilde{j_1}]q(x_1(c'), i_1, \widetilde{j_1})[k_2, c_1, \widetilde{d_1}]}{\sum_{\widetilde{k},\widetilde{d'},\widetilde{d_1},\widetilde{j_1}} \pi(c', \widetilde{d'})\tau_1(\widetilde{d'})[\widetilde{j_1}]q(x_1(c'), i_1, \widetilde{j_1})[\widetilde{k}, c_1, \widetilde{d_1}]} \tag{4.4}$$

whenever the left-hand side is well-defined. Since τ_1 has full support, the right-hand side is also well-defined and does not depend on d_1, j_1. Moreover, for all $p \in \Delta(K)$, we can choose an initial distribution π such that $\pi(x_1 = p) > 0$. It follows that there exists a map F such that

$$F(p, i, c)[k] = \frac{q(p, i, j)[k, c, d]}{\sum_{k'\in K} q(p, i, j)[k', c, d]},$$

whenever the right hand side is well-defined and extended by $1/K$ otherwise.

For the converse assertion, we already mentioned that $(A1a)$ implies $(A1)$ for $t = 1$. We are therefore allowed to write down the formula for the conditional laws,

$$\mathbb{P}(k_2 = k \mid c_1, d_1, c', i_1, d', j_1) = \frac{q(x_1, i_1, j_1)[k, c_1, d_1]}{\sum_{\widetilde{k}\in K} q(x_1, i_1, j_1)[\widetilde{k}, c_1, d_1]} \tag{4.5}$$

It follows therefore that

$$\mathbb{P}(k_2 = k \mid c_1, d_1, c', i_1, d', j_1) = F(x_1, i_1, c_1) \quad \mathbb{P}_{\pi,\sigma,\tau}\text{-almost surely},$$

and since the right-hand-side is measurable with respect to the history of player 1, we have the equality

$$\begin{aligned}\mathbb{P}(k_2 = k \mid c_1, c', i_1) &= \mathbb{E}\left(\mathbb{P}(k_2 = k \mid c_1, d_1, c', i_1, d', j_1) \,|c', i_1, c_1\right)\\ &= F(x_1, i_1, c_1)\\ &= \mathbb{P}(k_2 = k \mid c_1, d_1, c', i_1, d', j_1).\end{aligned}$$

This proves $(A1)$ and our last assertion for $t = 2$. Note finally that the law of $(k_2, (c', i_1, c_1), (d', j_1, d_1))$, seen as an element of $\Delta(K \times \mathbb{N} \times \mathbb{N})$, fulfills $(A1a)$. Applying exactly the same argument with these new initial signals allows us therefore to conclude by induction on t and prove $(A1)$ for all $t \geq 1$. □

Under the assumption $(A1)$, player 1 can compute at each stage his beliefs about the state without knowing the strategy of player 2. Indeed F is fixed by the transition q and not by the initial probability. Then at each stage, player 1 knows his previous belief, the actions he has played and the signal he received, thus he can compute his new belief. Therefore, strategies defined as a function of these beliefs can be effectively played.

4.3.2 Player 1 can compute the beliefs of player 2 about himself

Our second assumption focuses on the knowledge of player 1 about player 2. We first restrict the set of strategies allowed for player 1. A reduced strategy of player 1 is a strategy which depends at each stage $t \in \mathbb{N}^*$ only on the couple of random variables (x_t, z_t). For $t = 1$, we define the beliefs like in any state by the projection of π on $\Delta(\Delta(\Delta(K)))$. A reduced strategy is Markovian with respect to the projection on $\Delta(K)$ and $\Delta(\Delta(K))$. If player 1 follows one of these strategies, we assume that he can compute the beliefs of player 2 about his first-order belief , directly from his signals without knowing the strategy of the second player.

(A2a) The probability π is such that there exists a map $f_\pi^1 : C' \to \Delta(\Delta(K))$ such that

$$z_1 = f_\pi^1(c') \quad \pi\text{-almost surely.}$$

(A2b) If $(A2a)$ is true and σ_1 is a reduced strategy, then there exists a sequence of maps $(f^t_{\pi,\sigma_1})_{t\in\mathbb{N}^*}$ such that for all $t \in \mathbb{N}^*$, $f^t_{\pi,\sigma_1} : H_1^t \to \Delta(\Delta(K))$ and for all τ

$$z_t = f^t_{\pi,\sigma_1}(h_1^t) \quad \mathbb{P}_{\pi,\sigma,\tau}\text{-almost surely.}$$

The introduction of reduced strategies for player 1 is necessary in order to exclude non relevant correlations between the players. Player 1 could base his decisions on what he observed from the past actions of player 2, so that his posterior beliefs will also depend on the behavior of player 2. As shown in lemma 4.4.1, player 1 does not need these strategies so we focus on reduced strategies. Note that for $t = 1$, assumption $(A2a)$ only concerns the initial information structure which is given by π.

Assumption $(A2)$ is independent of assumption $(A1)$, as shown in the next example, nevertheless it really makes sense only when player 1 is better informed. If assumptions $(A1)$, $(A2a)$ and $(A2b)$ are true, then player 1 can compute the beliefs of player 2 about the state and so is at the same time better and more informed. Furthermore, player 1 does not need to know the strategy of player 2 in order to compute the first-order belief and the second-order belief of player 2.

Example 4.3.5 Let $\Gamma = (K, I, J, C, D, q, g)$ be a stochastic game such that player 1 is in the dark and player 2 is perfectly informed: $K = \{\alpha, \beta\}$, I and J are finite, C is a singleton $\{c\}$ and $D = K$. The payoff mapping is anything and q is constant equal to $\frac{1}{2}\delta_{\alpha,c,\alpha} + \frac{1}{2}\delta_{\beta,c,\beta}$. At each stage the state is drawn with probability $(\frac{1}{2}, \frac{1}{2})$, player 1 gets no information and player 2 learns the state. Let us assume that q is also the initial probability. It is clear that player 1's signal is less accurate than player 2's signal so that assumption $(A1)$ is not satisfied. On the other hand, $(A2a)$ is satisfied since player 1 knows the belief of player 2 about himself which is $(\frac{1}{2}, \frac{1}{2})$ whatever is the initial signal.

In the rest of the article we restrict to initial probabilities which satisfy $(A1a)$ and $(A2a)$.

Definition 4.3.6 *Let $\Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$ denote the set of probabilities π meeting the two previous assumptions $(A1a)$ and $(A2a)$, i.e.*

$$\mathcal{L}_\pi(k_1 \mid c', d') = \mathcal{L}_\pi(k_1 \mid c') \quad \pi\text{-almost surely,}$$

and there exists a map $f_\pi^1 : C' \to \Delta(\Delta(K))$ such that

$$z_1 = f_\pi^1(c') \quad \pi\text{-almost surely.}$$

4.3.3 Player 1 controls the relevant information

The last assumption that will be needed in order to prove the existence of the uniform value requires that player 1 controls the evolution of information about the state during the game. In fact, we restrict to player 1 playing reduced strategies.

(A3) For all $\pi \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$, for all σ_1 reduced strategy of player 1, τ_1 and τ_1' strategies of player 2, the law of z_2 is the same under (π, σ_1, τ_1) and under (π, σ_1, τ_1').

A natural example is a stochastic game with complete information where the transition is a mapping from $K \times I$ to $\Delta(K)$. More generally, any repeated game such that the transition is a mapping from $K \times I$ to $\Delta(K \times C \times D)$ fulfills assumption $(A3)$. Since player 2 has no influence neither on the state nor on the signals, he does not influence the evolution of the beliefs.

4.3.4 Result

Let us now state the main result of this work.

Theorem 4.3.7 *Let $\Gamma = (K, I, J, C, D, q, g)$ be a finite repeated game such that assumptions $(A1, A2, A3)$ are true. We say that Γ is a stochastic game with a more informed controller. For all $\pi \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$, $\Gamma(\pi)$ has a uniform value.*

Our model answers to a particular case of the following conjecture, from Mertens, Sorin and Zamir [MSZ94]: *If player 1's information includes player 2's at every stage, then there exists a limit value and it is equal to the maximal payoff that player 1 can guarantee,* $\lim_{n\to\infty} v_n = \lim_{\lambda\to 0} v_\lambda = \underline{v}$.

To conclude this section, we state a couple of stronger but easier to check assumptions: player 1 can deduce exactly the signal received by player 2 and player 2 can not influence the joint law of his signal and the belief of player 1. First, we define the mapping $H_{x,i}$ which associates to any couple of signals $(c, d) \in C \times D$, the couple $(x, d) \in \Delta(K) \times D$ consisting in the belief of player 1 and in the signal of player 2.

Definition 4.3.8 *For all, $x, i, j \in \Delta(K) \times I \times J$, let $q_{C\times D}(x, i, j)$ denote the marginal distribution on $C \times D$ induced by $q(x, i, j)$. Let also $H_{x,i}$ the map defined on $C \times D$ by*

$$H_{x,i}(c, d) = (F(x, i, c), d) \in \Delta(K) \times D.$$

With this notation, we can define a set of assumption on the marginal of q. The assumption $(A1)$, $(A2a)$ are unchanged and we define $(A'2b)$ and $(A'3)$.

($A'2b$) Player 1 knows the signal of player 2 i.e. there exists a map $h : C \to D$ such that for all $(k, i, j) \in K \times I \times J$, $\sum_{c\in C} q(k, i, j)[c, h(c)] = 1$.

($A'3$) The image probability $\phi(x, i)$ of $q_{C\times D}(x, i, j)$ by the map $H_{x,i}$ does not depend on j.

Corollary 4.3.9 *Let $\Gamma = (K, I, J, C, D, q, g)$ be a repeated game such that assumptions $A1, A2a, A'2b$ and $A'3$ are true.*

$$\textit{For all } \pi \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N}), \quad \Gamma(\pi) \textit{ has a uniform value.}$$

Lemma 4.3.10 *If* $(A1)$ *and* $(A2a)$ *hold, then* $(A'2b)$ *and* $(A'3)$ *imply* $(A2b)$ *and* $(A3)$.

Proof: First, we prove that $(A'3)$ implies $(A3)$ and $(A2)$ for $t = 2$ then we check that along reduced strategies, for all $t \geq 2$, z_t does not depend on player 2 and player 1 can compute the random variable z_t without knowing the strategy of player 2.

It follows from the definitions and from lemma 4.3.4 that

$$\begin{aligned}\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_2, d_1 \mid c', d', i_1, j_1) &= \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(F(x_1, i_1, c_1), d_1 \mid c', d', i_1, j_1) \\ &= \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(H_{x_1,i_1}(c_1, d_1) \mid c', d', i_1, j_1) = \phi(x_1, i_1),\end{aligned}$$

since (x_1, i_1) is measurable with respect to (c', d', i_1, j_1) and $\phi(x_1, i_1)$ is the image probability of $q_{C\times D}(x_1, i_1, j_1)$ by the map H_{x_1,i_1}. Therefore the conditional probability, on the joint density of (x_2, d_1), does not depend on the strategy of player 2. Precisely, we have

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_2, d_1 \mid d', j_1) = \mathbb{E}[\phi(x_1, i_1) \mid d', j_1].$$

Since j_1 and (x_1, i_1) are independent conditionally on d', we have

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_2, d_1 \mid d', j_1) = \mathbb{E}[\phi(x_1, i_1) \mid d'].$$

The right hand side depends neither on τ nor on j_1, so $\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_2, d_1 \mid d', j_1)$ does not depend on player 2 actions and therefore the joint law (x_2, z_2) does not depend on the actions of player 2. Taking the marginal on z_2 proves that η_2 does not depend on player 2's actions and $(A3)$. We now prove that player 1 can compute the auxiliary random variable z_2 by continuing the computation. Using $(A2a)$ and that σ_1 is reduced, i_1 can be written as a function of (x_1, z_1) and of an independent random variable U uniformly distributed on $[0, 1]$. Recall that the law of x_1 is z_1, thus we have

$$\begin{aligned}\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_2, d_1 \mid d', j_1) &= \mathbb{E}[\phi(x_1, i_1(x_1, z_1, u)) \mid d'] \\ &= \int_{\Delta(K)\times[0,1]} \phi(x, i_1(x, z_1, u) dz_1[x] du.\end{aligned}$$

Player 1 can compute the joint law of (x_2, d_1) conditionally on (d', j_1) since it depends only on (z_1, σ_1). Moreover by assumption $A'2$, he knows d_1 the signal of player 2, so he is able to compute z_2 which proves $(A2)$ for $t = 2$.

Let us prove by induction that for all $t \geq 2$, the joint law of (x_t, z_t) does not depend on player 2, and player 1 can compute the realization (x_t, z_t). Let $t \geq 2$ and σ be a reduced strategy such that both properties are true for t, then we have

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_{t+1}, d_t \mid h_t^1, h_t^2) = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(F(x_t, i_t, c_t), d_t \mid h_t^1, h_t^2) = \mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(H_{x_t,i_t}(c_t, d_t) \mid h_t^1, h_t^2) = \phi(x_t, i_t),$$

and

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_{t+1}, d_t \mid h_t^2) = \mathbb{E}[\phi(x_t, i_t) \mid h_t^2]$$

By induction, the joint law of (x_t, z_t) does not depend on player 2. Since player 1 is playing a reduced strategy, the law of i_t does not depend on the actions played by player 2:

$$\mathcal{L}_{\mathbb{P}_{\pi,\sigma,\tau}}(x_{t+1}, d_t \mid h_t^2) = \mathbb{E}[\phi(x_t, i_t) \mid d', d_1, ..., d_t]$$

Player 1 can compute the joint law of d_t and x_{t+1}, and knows d_t by $A'2$ so he can compute the random variable z_{t+1} and the joint law does not depend on player 2's actions. □

4.4 Proof of the existence of the uniform value

The proof is divided into three steps. First we show that the value v only depends of π through his projection $\eta_1 \in \Delta_f(\Delta_f(\Delta(K)))$. We define the value $\widetilde{v}$ on $\Delta_f(\Delta_f(\Delta(K)))$. Secondly we introduce an auxiliary stochastic game Ψ on $\Delta_f(\Delta(K))$ and check that he satisfies some weakened assumptions of Renault [Ren12b] which imply the existence of a uniform value. Finally we show that both players can guarantee this uniform value in the original game: player 2 by playing by blocks and player 1 by using an optimal strategy of the auxiliary stochastic game Ψ.

4.4.1 The canonical value function $\widetilde{v}_\theta$

Let us first show that if assumptions $(A1a)$ and $(A2a)$ hold, then $v_\theta(\pi)$ can be factorized by an equivalence relation. Then we prove that the mapping defined on the quotient is concave and 1-Lipschitz under a suitable metric.

Lemma 4.4.1 *Let $\pi, \pi' \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$ and let $\eta_1, \eta_1' \in \Delta_f(\Delta_f(\Delta(K)))$ be defined as*

$$\eta_1 := \mathcal{L}_\pi\left(\mathcal{L}_\pi\left(\mathcal{L}_\pi(k_1|c')|d'\right)\right) \text{ and } \eta_1' := \mathcal{L}_{\pi'}\left(\mathcal{L}_{\pi'}\left(\mathcal{L}_{\pi'}(k_1|c')|d'\right)\right).$$

If both projections are equal, $\eta_1 = \eta_1'$, then $v_\theta(\pi) = v_\theta(\pi')$.

Proof: Let (σ, τ) be a pair of behavior strategies in $\Gamma_\theta(\pi)$. It is enough to show that $v_\theta(\pi)$ depends on π only through η_1. Let $x_1 := \mathcal{L}_\pi(k_1|c')$ and $z_1 := \mathcal{L}_\pi\left(\mathcal{L}_\pi(k_1|c')|d'\right)$. Note that z_1 is observed by player 2 since it is a function of d' and π, and that x_1 is observed by player 1. We can also assume without loss of generality that z_1 is observed by player 1 using assumption $(A2a)$, i.e. there exists a map $f_\pi^1 : C' \to \Delta(\Delta(K))$ such that

$$f_\pi^1(c') = z_1 \quad \pi\text{-almost surely.}$$

Let us construct a reduced version of the game $\Gamma_\theta(\pi)$ in which player 1 and player 2 are constrained to choose strategies that depend only on c' and d' through the variables (x_1, z_1) and z_1 respectively, and keeping the same payoff function. This game has a value since the set of possible values of (x_1, z_1) is finite and this value is exactly the value of $\Gamma_\theta(\tilde{\pi})$ where $\tilde{\pi}$ is the joint distribution of $(k_1, (x_1, z_1), z_1)$ seen as an element of $\Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$. The sets of strategies in $\Gamma_\theta(\tilde{\pi})$ (denoted Σ' and $\mathcal{T}'$) can be seen as subsets of Σ and $\mathcal{T}$ via the previous identification and we will prove that both games have the same value and that $v_\theta(\tilde{\pi})$ depends only on η_1.

Assume at first that $\tau \in \mathcal{T}'$ and $\sigma \in \Sigma$ and let μ denote the joint law of (k_1, c', d', x_1, z_1) induced by π. By disintegration, we have

$$\begin{aligned}
\gamma_\theta(\pi,\sigma,\tau) &= \int_{K\times\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \gamma_\theta(k_1,\sigma(c'),\tau(z_1))d\mu(k_1,c',x_1,z_1),\\
&= \int_{\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \left(\int_K \gamma_\theta(k_1,\sigma(c'),\tau(z_1))d\mu(k_1|c',x_1,z_1)\right) d\mu(c',x_1,z_1)\\
&= \int_{\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \int_K \gamma_\theta(k_1,\sigma(c'),\tau(z_1))d\mu(k_1|c')d\mu(c',x_1,z_1)\\
&= \int_{\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \langle\gamma_\theta(.,\sigma(c'),\tau(z_1)),x_1\rangle_{\mathbb{R}^K} d\mu(c',x_1,z_1),
\end{aligned}$$

where we used that $\mu(k_1|c',x_1,z_1) = \mu(k_1|c') = x_1$ using that (x_1,z_1) are c'-measurable and the notations $\gamma_\theta(.,\sigma(c'),\tau(z_1))$ for $(\gamma_\theta(k,\sigma(c'),\tau(z_1)))_{k\in K} \in \mathbb{R}^K$ and $\langle .,.\rangle_{\mathbb{R}^K}$ for the scalar product in $\mathbb{R}^K$.

Taking the supremum over all strategies of player 1, we obtain

$$\sup_{\sigma\in\Sigma} \gamma_\theta(\pi,\sigma,\tau) = \int_{\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \sup_{\sigma(c')} \langle\gamma_\theta(.,\sigma(c'),\tau(z_1)),x_1\rangle_{\mathbb{R}^K} d\mu(c',x_1,z_1).$$

The supremum can be written inside the integral and is achieved by strategies depending only on (x_1,z_1) since these variables are c'-measurable. It means that there exists an optimal σ in Σ', which proves

$$\inf_{\tau\in\mathcal{T}'} \sup_{\sigma\in\Sigma} \gamma_\theta(\pi,\sigma,\tau) = \inf_{\tau\in\mathcal{T}'} \sup_{\sigma\in\Sigma'} \gamma_\theta(\pi,\sigma,\tau).$$

Moreover the value of the reduced game depends only on η_1 since taking the infimum over $\tau \in \mathcal{T}'$

$$\inf_{\tau\in\mathcal{T}'} \sup_{\sigma\in\Sigma'} \gamma_\theta(\pi,\sigma,\tau) \tag{4.6}$$

$$= \int_{\Delta(\Delta(K))} \left[\inf_{\tau(z_1)} \int_{\Delta(K)} \left(\sup_{\sigma(x_1,z_1)} \langle\gamma_\theta(.,\sigma(x_1,z_1),\tau(z_1)),x_1\rangle_{\mathbb{R}^K}\right) d\mu(x_1 \mid z_1)\right] d\mu(z_1) \tag{4.7}$$

which depends only on the law of z_1. Indeed using that z_1 is d'-measurable, it follows that $z_1 = \mu(x_1 \mid d') = \mu(x_1 \mid z_1)$.

Let us prove a dual equality starting with $\sigma \in \Sigma'$ and $\tau \in \mathcal{T}$:

$$\begin{aligned}
\gamma_\theta(\pi,\sigma,\tau) &= \int_{K\times\mathbb{N}\times\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \gamma_\theta(k_1,\sigma(x_1,z_1),\tau(d'))d\mu(k_1,c',d',x_1,z_1),\\
&= \int_{\mathbb{N}\times\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \left(\int_K \gamma_\theta(k_1,\sigma(x_1,z_1),\tau(d'))d\mu(k_1|c',d',x_1,z_1)\right) d\mu(c',d',x_1,z_1)\\
&= \int_{\mathbb{N}\times\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \langle\gamma_\theta(.,\sigma(x_1,z_1),\tau(d')),x_1\rangle_{\mathbb{R}^K} d\mu(c',d',x_1,z_1)\\
&= \int_{\mathbb{N}\times\Delta(K)\times\Delta(\Delta(K))} \langle\gamma_\theta(.,\sigma(x_1,z_1),\tau(d')),x_1\rangle_{\mathbb{R}^K} d\mu(d',x_1,z_1)\\
&= \int_{\mathbb{N}\times\Delta(\Delta(K))} \left(\int_{\Delta(K)} \langle\gamma_\theta(.,\sigma(x_1,z_1),\tau(d')),x_1\rangle_{\mathbb{R}^K} d\mu(x_1 \mid d',z_1)\right) d\mu(d',z_1).
\end{aligned}$$

For the second equality, we used that $\mu(k_1 \mid c', d', x_1, z_1) = \mu(k_1 \mid c', d') = \mu(k_1 \mid c') = x_1$ which follows from the fact that (x_1, z_1) is c'-measurable and assumption $(A1a)$.

Taking the infimum over all $\tau \in \mathcal{T}$, it follows that

$$\inf_{\tau\in\mathcal{T}} \gamma(\pi,\sigma,\tau) = \int_{\mathbb{N}\times\Delta(\Delta(K))} \inf_{\tau\in\mathcal{T}} \left(\int_{\Delta(K)} \langle \gamma_\theta(.,\sigma(x_1,z_1),\tau(d')), x_1\rangle_{\mathbb{R}^K} d\mu(x_1 \mid d') \right) d\mu(d', z_1). \tag{4.8}$$

The infimum inside the integral is achieved for strategies depending only on $z_1 = \mu(x_1 \mid d')$ since z_1 is d'-measurable, and we have proved

$$\sup_{\sigma\in\Sigma'} \inf_{\tau\in\mathcal{T}} \gamma_\theta(\pi,\sigma,\tau) = \sup_{\sigma\in\Sigma'} \inf_{\tau\in\mathcal{T}'} \gamma_\theta(\pi,\sigma,\tau).$$

Finally, using that $\mathcal{T}' \subset \mathcal{T}$ and $\Sigma' \subset \Sigma$, it follows that

$$v_\theta(\pi) = \sup_{\sigma\in\Sigma} \inf_{\tau\in\mathcal{T}} \gamma_\theta(\pi,\sigma,\tau) \geq \sup_{\sigma\in\Sigma'} \inf_{\tau\in\mathcal{T}} \gamma_\theta(\pi,\sigma,\tau) = v_\theta(\tilde{\pi}),$$

$$v_\theta(\pi) = \inf_{\tau\in\mathcal{T}} \sup_{\sigma\in\Sigma} \gamma_\theta(\pi,\sigma,\tau) \leq \inf_{\tau\in\mathcal{T}'} \sup_{\sigma\in\Sigma} \gamma_\theta(\pi,\sigma,\tau) = v_\theta(\tilde{\pi}),$$

which proves the equality. Since $v_\theta(\tilde{\pi})$ depends only on η_1, the proof is complete. □

In view of this result, it is appropriate to work directly on the set $\Delta(\Delta(\Delta(K)))$, i.e. for any π, π' such that $\eta = \eta'$ the game is essentially the same. On the other hand, for any $\eta \in \Delta_f(\Delta_f(\Delta(K)))$ there is a canonical way to generate the distribution η: it is enough to take some finite sets of signals $C' \subset \Delta(K)$ and $D' \subset \Delta_f(\Delta(K))$, and a distribution $\pi \in \Delta_f^*(K \times C' \times D')$ such that $\pi(k,p,z) = \eta(z)z(p)p(k)$. The following definition captures these ideas.

Definition 4.4.2 *For any $\eta \in \Delta_f(\Delta_f(\Delta(K)))$ we may define $\widetilde{\Gamma}(\eta) := \Gamma(C', D', \pi)$, where C', D' and π generates η canonically. If $\eta = \delta_z$ for some $z \in \Delta_f(\Delta(K))$, we denote $\widetilde{\Gamma}(z) = \widetilde{\Gamma}(\delta_z)$. The value is denoted by $\widetilde{v}_\theta(\eta)$ (resp. $\widetilde{v}_\theta(z)$).*

Reciprocally, for any $\pi \in \Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$, let $\Phi(\pi) \in \Delta_f(\Delta_f(\Delta(K)))$ be the distribution generated by π:

$$\Phi(\pi) = \sum_{d'\in\mathbb{N}} \pi(d') \delta_{\left(\sum_{c'\in\mathbb{N}} \pi(c'|d')\delta_{\pi(.|c',d')}\right)}.$$

An important consequence of Lemma 4.4.1 is that, if $\pi, \pi' \in \Delta_f^*(K\times\mathbb{N}\times\mathbb{N})$ are such that $\Phi(\pi) = \Phi(\pi')$, we have that $v_\theta(\pi) = v_\theta(\pi') = \widetilde{v}_\theta(\Phi(\pi))$.

Less formally, the game $\widetilde{\Gamma}(\eta)$ corresponds to the game $\Gamma(\pi)$, where stage 1 has been replaced by the following procedure: η is common knowledge, player 2 is informed about the realization z of a random variable of law η (player 2 learns his belief), and player 1 is informed about z (his opponent's belief) and about the realization p of a random variable of law z(his own belief), the state variable is finally selected according to p, but none of the players observes it. If $\eta = \delta_z$, for some $z \in \Delta_f(\Delta(K))$, then the set of initial signals for player 2 is reduced to a singleton. That is, player 1 receives a partial information about the state, whereas player 2 only knows the joint distribution over the state and player 1's signal.

Let (Z,d) be a compact metric space and $E_1(Z)$ the set of 1-Lipschitz functions on Z. The function

$$d_{KR} : \Delta(Z) \times \Delta(Z) : (\mu,\nu) \to \sup_{f \in E_1(Z)} \int_Z f d\mu - \int_Z f d\nu$$

is a metric on $\Delta(Z)$ which makes $\Delta(Z)$ compact. Moreover, for all $\mu,\nu \in \Delta(Z)$

$$d_{KR}(\mu,\nu) = \min_{\chi \in \Pi(\mu,\nu)} \int_{Z\times Z} d(x,y) d\chi(x,y),$$

where $\Pi(\mu,\nu)$ is the set of probabilities on $Z \times Z$ having for marginals μ and ν.

Let f be a measurable function on Z and define $\tilde{f} : \Delta(Z) \to \mathbb{R}$ by for all $\mu \in \Delta(Z)$, $\tilde{f}(\mu) = \int_Z f d\mu$, then $\tilde{f} \in E_1(\Delta(Z))$ with respect to d_{KR}) if and only if $f \in E_1(Z)$ with respect to d.

In the following, $X = \Delta(K)$ is endowed with the norm $\|.\|_1$ induced by $\mathbb{R}^K$, $Z = \Delta(\Delta(K))$ is endowed with the Kantorovitch Rubinstein metric d_{KR} induced by the metric space $(\Delta(K), \|.\|_1)$, and finally, $\Delta(Z)$ is endowed with the Kantorovitch Rubinstein metric D_{KR} induced by the metric space (Z, d_{KR}).

Lemma 4.4.3 *Let $\eta \in \Delta_f(\Delta_f(\Delta(K)))$ and $z \in \Delta_f(\Delta(K))$. Then $\widetilde{v}_\theta(\eta)$ is linear on $\Delta_f(\Delta_f(\Delta(K)))$ and as a mapping on $\Delta(\Delta(K))$, $\widetilde{v}_\theta(z)$ is 1-Lipschitz for the Kantorovitch-Rubinstein metric d_{KR}.*

Proof: The first assertion is immediate since by definition the players learn the realization of η. We focus on the second assertion. Let z,$z' \in \Delta_f(\Delta(K))$ and $\theta \in \Delta(\mathbb{N}^*)$. By definition of the Kantorovitch-Rubinstein metric, there exists $\chi \in \Delta(\Delta(K) \times \Delta(K))$ such that the first marginal is z, the second is z' and

$$d_{KR}(z,z') = \int_{\Delta(K)\times\Delta(K)} \|p - p'\|_1 d\chi(p,p').$$

We denote by $\chi(p|p')$ the conditional law of p given p'.

Let $\sigma \in \Sigma$ be a behavior strategy for player 1 in the game $\widetilde{\Gamma}(z)$. As in Section 4.3, let $\sigma(p)$ denote the strategy, once the private signal p has been revealed to him. Let us construct a general strategy for player 1 as follows. Let $(\Omega,\mathbb{P}) = ([0,1],dx)$ be the auxiliary probability space[6] that will be used as a "tossing coin". The classical representation result of Blackwell-Dubins (see [BD83]) asserts that there exists a jointly Borel-measurable map $\phi : \Omega \times \Delta(\Delta(K)) \mapsto \Delta(K)$ such that for all $\nu \in \Delta(\Delta(K))$, $\phi(\cdot,\nu)$ is a ν-distributed random variable. Therefore, the map $\sigma'(\omega,p') = \sigma(\phi(\omega,\chi(p|p')))$ defines a general strategy which is equivalent to a behavior strategy by Kuhn's theorem (see [Kuh53]). It follows

6. Using a continuum of alternatives is clearly unnecessary but allows to unify the treatment, note also that the above proof of Lipschitz continuity can be generalized to compact metric spaces without modifying the presentation.

that

$$\begin{aligned}\gamma_\theta(z',\sigma',\tau) &= \int_{\Delta(K)\times\Omega} \gamma_\theta(p',\sigma'(\omega,p'),\tau)dz'(p')\otimes d\mathbb{P}(\omega),\\ &= \int_{\Delta(K)}\left(\int_\Omega \gamma_\theta(p',\sigma(\phi(\omega,\chi(p|p'))),\tau)d\mathbb{P}(\omega)\right)dz'(p'),\\ &= \int_{\Delta(K)}\left(\int_{\Delta(K)} \gamma_\theta(p',\sigma(p),\tau)d\chi(p|p')\right)dz'(p'),\\ &= \int_{\Delta(K)\times\Delta(K)} \gamma_\theta(p',\sigma(p),\tau)d\chi(p,p'),\end{aligned}$$

where the last equality follows from $dz'(p') = d\chi(p')$. Recall that by assumption $g(k,i,j)\in[0,1]$, $\forall(k,i,j)\in K\times I\times J$. Consequently for all $p\in\Delta(K)$, $\sigma\in\Sigma$ and $\tau\in\mathcal{T}$, $\gamma_\theta(p,\sigma(p),\tau)$ takes values in $[0,1]$. Hence

$$\begin{aligned}|\gamma_\theta(z',\sigma',\tau)-\gamma_\theta(z,\sigma,\tau)| &\le \int_{\Delta(K)\times\Delta(K)} |\gamma_\theta(p,\sigma(p),\tau)-\gamma_\theta(p',\sigma(p),\tau)|d\chi(p,p'),\\ &\le \int_{\Delta(K)\times\Delta(K)} \|p-p'\|_1 d\chi(p,p'),\\ &= d_{KR}(z,z').\end{aligned}$$

It follows that $|\widetilde{v}_\theta(z)-\widetilde{v}_\theta(z')|\le d_{KR}(z,z')$, for any $z,z'\in\Delta_f(\Delta(K))$. □

Remark 4.4.4 Note that usually, the underlying space is $\Delta(K)$ with an underlying finite space K and with the discrete metric. In order to prove that the value is 1-Lipschitz, in this case, we can use the same strategy in $\Gamma(z)$ and in $\Gamma(z')$. Here the state is $\Delta_f(Z)$ and we consider the norm 1 on Z, so two states may be close with very different supports. An optimal strategy σ in $\Gamma(z)$ may have no sense in z'. Therefore, we can not just play σ in $\Gamma(z)$ and we need given σ in $\Gamma(z)$, to build σ' in $\Gamma(z')$ which behaves in z' like σ in z.

Example 4.4.5 *Let $z=\delta_{(\frac12,\frac12)}$ and $z'=\frac12\delta_{(\frac12-\varepsilon,\frac12+\varepsilon)}+\frac12\delta_{(\frac12+\varepsilon,\frac12-\varepsilon)}$ be two initial distributions. An optimal strategy in $\Gamma(z)$ is bonded only at $(\frac12,\frac12)$ and can play anyhow in $(\frac12-\varepsilon,\frac12+\varepsilon)$ and in $(\frac12+\varepsilon,\frac12-\varepsilon)$. The good way to use the proximity between z and z' is to always play as if the initial distribution was $(\frac12,\frac12)$: let σ' be defined by $\sigma'(z)=\sigma(\frac12,\frac12)$ for all $z\in\Delta(K)$.*

Lemma 4.4.6 *The mapping $\widetilde{v}_\theta(z)$ is concave on $\Delta_f(\Delta(K))$.*

Proof: We follow the proof of Aumann and Maschler. Let λ be a real in $[0,1]$ and let Y be a random variable with values in $\{0,1\}$, such that $\mathbb{P}(Y=0)=\lambda$. Let $z,z'\in\Delta_f(\Delta(K))$. The random variable P is selected according to the distribution z if $Y=0$ and z' if $Y=1$, the state variable k_1 is finally selected according to p if $P=p$. Compare now the two following situations: on one hand, the game with initial signals (Y,P) for player 1 and nothing for player 2 and on the other hand the game with initial signals (Y,P) for player 1 and Y for player 2. These two distributions of initial signals and states fulfill our assumptions and it's clear that the value of the second is less or equal than the value

of the first for any evaluation $\theta \in \Delta_f(\mathbb{N}^*)$ since the set of behavior strategies of player 2 in the second game is larger than in the first game. Translating this inequality using $\widetilde{v}$, we deduce directly

$$\widetilde{v}_\theta(\delta_{\lambda z+(1-\lambda)z'}) \geq \widetilde{v}_\theta(\lambda\delta_z + (1-\lambda)\delta_{z'}) = \lambda\widetilde{v}_\theta(z) + 1 - \lambda\widetilde{v}_\theta(z'),$$

which proves the lemma. □

4.4.2 The auxiliary stochastic game Ψ

Recall that $Z = \Delta(\Delta(K))$ and let it be our new state space, which corresponds to player 2's belief about player 1's belief about the current state. It is a convex compact subset of a normed vector space and we are going to express the auxiliary game and the recursive formula on this state space.

Let Ψ be the stochastic game defined by

- the state space $Z = \Delta_f(\Delta(K))$,
- the action space $A = \{f : \Delta(K) \to \Delta(\mathcal{I}), \text{measurable}\}$,
- the action space $B = \Delta(J)$,
- the reward function $\widetilde{g} : Z \times A \times B \to [0,1]$ defined, for any $z \in \Delta_f(\Delta(K))$ by

$$\widetilde{g}(z,a,b) = \sum_{p\in\text{supp}(z)} \sum_{(i,j)\in I\times J} b(j)a(p,i)g(p,i,j)z(p),$$

  where $\text{supp}(z)$ stands for the support of z.
- the transition function $\widetilde{q} : Z \times A \times B \to \Delta_f(Z)$ is defined as $\widetilde{q}(z,a,b) = \Phi(Q(z,a,b))$, where $Q(z,a,b) \in \Delta_f((K) \times (\Delta(K) \times C) \times (D))$ is the induced joint distribution of $(k_2, (p, c_1), (d_1))$ in the canonical game $\widetilde{\Gamma}(\delta_z)$ where players play at the first stage $\sigma_1 = a$ and $\tau_1 = b$. The sets C, D, K and $\text{supp}(z)$ being finite, we may consider Q as an element in $\Delta_f(K \times \mathbb{N} \times \mathbb{N})$. Moreover using assumptions $(A1)$ and $(A2)$, it is an element in the restricted set $\Delta_f^*(K \times \mathbb{N} \times \mathbb{N})$, of initial probabilities with a more informed controller.

We aim to apply a weakened version of Renault [Ren12b] to the game Ψ, thus let us first recall these hypotheses as they appear in the original article. Given an evaluation $\theta \in \Delta(\mathbb{N}^*)$, recall that θ^+ denote the evaluation after one stage renormalized : if $\theta_1 = 1$ then for all $t \geq 1$, $\theta_t^+ = 0$ and if $\theta_1 < 1$ then for all $t \geq 1$, $\theta_t^+ = \frac{\theta_{t+1}}{1-\theta_1}$.

Hypotheses 4.4.1

H1) The map $\widetilde{q}$ does not depend on b.

H2) Z is a compact convex subset of a normed vector space,

H3) A and B are convex compact subsets of a topological vector space,

H4) $(a \mapsto \widetilde{g}(z,a,b))$ is concave upper semi-continuous $\forall (z,b) \in Z \times B$ and $(b \mapsto \widetilde{g}(z,a,b))$ is convex and lower semi-continuous $\forall (z,a) \in Z \times A$.

H5) There exists a subset $\mathcal{C}$ of 1-Lipschitz functions such that for all f in $\mathcal{C}$, $\alpha \in [0,1)$, the function $\phi(\alpha, f)$ defined as following

$$\forall z \in \Delta_f(Z)\ \phi(\alpha,f)(z) = \sup_{a\in A} \min_{b\in B} \{\alpha\widetilde{g}(z,a,b) + (1-\alpha)\widetilde{q}(z,a,b)\,[w_{\theta^+}]\}$$

is in $\mathcal{C}$.

H6) The mapping $a \mapsto \widetilde{q}(z,a)$ is concave for the Choquet order and continuous.

H7) (Splitting assumption) Let z be a convex combination in $\Delta_f(\Delta(K))$, $z = \sum_{s=1}^{S} \lambda_s z_s$ and $(a_s)_{s\in S}$ be a family of actions in A^S. Then there exists $a \in A$ such that

$$\widetilde{q}(z,a) \geq \sum_{s\in S} \lambda_s \widetilde{q}(z_s,a_s) \text{ and } \widetilde{g}(z,a) \geq \sum_{s\in S} \lambda_s \widetilde{g}(z_s,a_s).$$

The main consequence of assumption $(A3)$ is that the player 2 can not influence the transition in the auxiliary game so the map $\widetilde{q}$ does not depend on b, i.e.

$$\forall (z,a) \in Z \times A,\ \forall b,b' \in B,\ \widetilde{q}(z,a,b) = \widetilde{q}(z,a,b').$$

$(H1)$ is satisfied and from now on, we will work under the shorter notation $\widetilde{q}(z,a)$ for $\widetilde{q}(z,a,b)$.

The hypotheses $(H2, H3, H4, H6, H7)$ ensure the application of Sion's theorem in several steps of Renault's proof. Here they are not all satisfied since, for example, the set A is not compact. However, it is well known that adding some geometrical hypotheses allows to weaken the topological assumptions in Sion's theorem (see, for instance, Proposition A.8 in Sorin's monography [Sor02]). For instance, if A is a convex set, B is a compact convex subset of a topological vector space, $(a \mapsto \widetilde{g}(z,a,b))$ is concave $\forall (z,b) \in Z \times B$ and $(b \mapsto \widetilde{g}(z,a,b))$ is convex and lower semi-continuous $\forall (z,a) \in Z \times A$, Sion's result applies to the one-stage game and it has a value. They can be replaced without altering the proof of Renault [Ren12b] by

Hypotheses 4.4.2

H2) Z is a compact convex subset of a normed vector space.

H3') B is a convex compact subset of a topological vector space, A is a convex set.

H4') $(a \mapsto \widetilde{g}(z,a,b))$ is concave $\forall (z,b) \in Z \times B$ and $(b \mapsto \widetilde{g}(z,a,b))$ is convex and lower semi-continuous $\forall (z,a) \in Z \times A$.

H6') The mapping $a \mapsto \widetilde{q}(z,a)$ is concave for the Choquet order.

H7') (Splitting assumption) Let z be a convex combination in $\Delta_f(\Delta(K))$, $z = \sum_{s=1}^{S} \lambda_s z_s$ and $(a_s)_{s\in S}$ be a family of actions in A^S. Then there exists $a \in A$ such that for all $b \in B$:

$$\widetilde{q}(z,a) \geq \sum_{s\in S} \lambda_s \widetilde{q}(z_s,a_s) \text{ and } \widetilde{g}(z,a,b) \geq \sum_{s\in S} \lambda_s \widetilde{g}(z_s,a_s,b).$$

Assumption $(H2)$ is satisfied since the Kantorovitch-Rubinstein metric can be extended to a norm on the space of finite signed measures. Moreover assumptions $(H3')$ and $(H4')$ are clearly satisfied. Therefore, we need to prove $(H6')$ and $(H7')$.

Lemma 4.4.7 *The game Ψ fulfills $H6'$ and $H7'$.*

Proof: Let z be a convex combination in $\Delta_f(\Delta(K))$, $z = \sum_{s=1}^{S} \lambda_s z_s$ and $(a_s)_{s\in S}$ be a family of actions in A^S. Denote $\mu(z_s,a_s) \in \Delta(\Delta(K) \times I)$ the joint law induced on $\Delta(K) \times I$ by (z_s,a_s). There exists $a \in A$ such that $\mu(z,a) = \sum_{s\in S} \lambda_s \mu(z_s,a_s)$.

At first, we prove that $Q(z,a,b) = \sum_{s\in S} \lambda_s Q(z_s, a_s, b)$.

$$\begin{aligned} Q(z,a,b) &= \sum_{p,i} Q(\delta_p, \delta_i, b)\mu(z,a)[p,i] \\ &= \sum_{p,i}\sum_{s\in S} Q(\delta_p, \delta_i, b)\lambda_s\mu(z_s,a_s)[p,i] \\ &= \sum_{s\in S}\lambda_s\left(\sum_{p,i} Q(\delta_p,\delta_i,b)\mu(z_s,a_s)[p,i]\right) \\ &= \sum_{s\in S}\lambda_s Q(z_s,a_s,b) \end{aligned}$$

Let $\pi = Q(z,a,b)$ (resp. $\pi_s = Q(z_s,a_s,b)$) and $\widetilde{c} = (p,i_1,c_1)$ (resp. $\widetilde{d} = (d_1,j_1)$) considered as an element in $\mathbb{N}$. By construction,

$$\widetilde{q}(z,a) = \Phi(\pi) = \mathcal{L}_\pi(\mathcal{L}_\pi(\mathcal{L}_\pi(k_2 \mid \widetilde{c}) \mid \widetilde{d})) = \sum_{\widetilde{d}\in D'} \pi(\widetilde{d})\delta_{\mathcal{L}_\pi(\mathcal{L}_\pi(k_2|\widetilde{c})|\widetilde{d})}.$$

Using lemma 4.3.4, the conditional law $\mathcal{L}_\pi(k_2 \mid \widetilde{c},\widetilde{d})$ does not depend on (z,a,b) but only on q and $\widetilde{c} = (p,i_1,c_1)$. Indeed, for all $\pi = \widetilde{q}(z,a,b)$ the following equality is π-almost surely true

$$\mathcal{L}_\pi(k_2 \mid p, i_1, c_1) = F(p,i_1,c_1).$$

The map $\pi \mapsto \mathcal{L}_\pi(\mathcal{L}_\pi(k_2 \mid \widetilde{c}),\widetilde{d}) = \mathcal{L}_\pi(F(\widetilde{c}),\widetilde{d}) \in \Delta_f(\Delta_f(\Delta(K)\times D')$ depends on π only once and therefore is linear. It was proved in Renault [Ren12b] that the following disintegration map denoted $\phi_{D'}$ is concave for the Choquet order :

$$\mathcal{L}_\pi(F(\widetilde{c}),\widetilde{d}) \mapsto \mathcal{L}_\pi(\mathcal{L}_\pi(F(\widetilde{c}) \mid \widetilde{d})) = \widetilde{q}(z,a).$$

We conclude that the first part of (H7') holds since

$$\widetilde{q}(z,a) = \phi_{D'}\left(\sum_{s\in S}\lambda_s\mathcal{L}_{\pi_s}(F(\widetilde{c}),\widetilde{d})\right) \geq \sum_{s\in S}\lambda_s\phi_{D'}\left(\mathcal{L}_{\pi_s}(F(\widetilde{c}),\widetilde{d})\right) = \sum_{s\in S}\lambda_s\widetilde{q}(z_s,a_s).$$

For the second part of H7, it is sufficient to note that (with abusive notations) $\mu(z,a) \mapsto \widetilde{g}(z,a,b)$ is linear so that for all $b \in B$

$$\widetilde{g}(z,a,b) = \sum_{s\in S}\lambda_s\widetilde{g}(z_s,a_s,b),$$

which implies the result. Finally, in case $z_s = z$ for all s, the same arguments also imply $(H6')$ since in this case one can choose $a = \sum_{s\in S}\lambda_s a_s$ in the above proof. □

In order to prove the last assumption $(H5)$, we first prove that the value of the game Ψ is equal to the canonical value function. Since we proved that the canonical value is 1-Lipschitz, it will imply that the set of mappings $\mathcal{C} = \{v_\theta, \theta \in \Delta(\mathbb{N}^*)\}$ satisfies $(H5)$. The proof is classic and consists to show that both functions fulfill the same recursive formula. From Sion minmax theorem we have

Proposition 4.4.8 *For any $\theta \in \Delta(\mathbb{N}^*)$ and any $\eta \in \Delta_f(\Delta_f(\Delta(K)))$, the game $\Psi_\theta(\eta)$ has a value*

$w_\theta(\eta)$ such that

$$\forall z \in \Delta_f(\Delta(K)),\ w_\theta(z) = \sup_{a\in A}\min_{b\in B}\left\{\theta_1\widetilde{g}(z,a,b) + (1-\theta_1)w_{\theta^+}(\widetilde{q}(z,a))\right\}, \tag{4.9}$$

$$= \min_{b\in B}\sup_{a\in A}\left\{\theta_1\widetilde{g}(z,a,b) + (1-\theta_1)w_{\theta^+}(\widetilde{q}(z,a))\right\}, \tag{4.10}$$

Moreover, in $\Psi_\theta(\eta)$ both players have optimal Markov strategies.

We can deduce the equality of the two sets of mappings.

Proposition 4.4.9 *For all $\theta \in \Delta(\mathbb{N}^*)$ and for any $z \in \Delta_f(\Delta(K))$, $w_\theta(z) = \widetilde{v}_\theta(z)$.*

Corollary 4.4.10 *The game Ψ fulfills $(H5)$.*

Proof:(Proposition 4.4.9) Since the payoffs are bounded, it is sufficient to prove the equality for probabilities with a finite support and then approximate infinite distributions by a sequence of probabilites with finite support. Notice that $\widetilde{v}_1(z) = w_1(z)$ for all z from the definition:

$$\begin{aligned}\widetilde{v}_1(z) &= \sup_{\sigma_1:\Delta(K)\to\Delta(I)}\inf_{b\in\Delta(J)}\int_{\Delta(K)} g(p,\sigma_1(p),b)dz(p)\\ &= \sup_{a\in A}\min_{b\in\Delta(J)}\widetilde{g}(z,a,b)\\ &= \min_{b\in\Delta(J)}\sup_{a\in A}\widetilde{g}(z,a,b)\\ &= w_1(z).\end{aligned}$$

It is enough to prove that w and $\widetilde{v}$ satisfy the same recurrence formula, or equivalently that $\widetilde{v}$ satisfies the following recurrence formula in Γ,

$$\begin{aligned}\widetilde{v}_\theta(z) &= \sup_{a\in A}\min_{b\in B}\theta_1\widetilde{g}(z,a,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))\\ &= \min_{b\in B}\sup_{a\in A}\ \theta_1\widetilde{g}(z,a,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))\end{aligned}$$

We prove the recursive formula by induction on the greatest element in the support of θ. If $\theta = \delta_1$, it follows from the preceding equality. Fix now $n \geq 2$, and assume that the proposition is true for every θ supported by $\{1,...,n-1\}$. Let $z \in \Delta_f(\Delta(K))$. We first prove that player 1 can defend in $\widetilde{\Gamma}_\theta(z)$ the quantity

$$\min_{b\in B}\sup_{a\in A}\left(\theta_1\widetilde{g}(z,a,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))\right).$$

Using the canonical representation $\widetilde{\Gamma}$, $\widetilde{v}_\theta(z) = v_\theta(\mu)$ where $\mu \in \Delta_f(K\times\Delta(K)\times\Delta_f(\Delta(K)))$ is defined by

$$\forall(k,p,u) \in K\times\Delta(K)\times\Delta_f(\Delta(K))\ \mu(k,p,u) = p(k)z(p)\mathbb{1}_{u=z}.$$

Consider the game $\Gamma_\theta(\mu)$ and let τ be a strategy of player 2. We denote by b the law induced by τ_1, $a^* \in A$ an action which realizes the supremum up to ε in the previous equation and σ^* an ε-optimal

strategy in the game $\Gamma_{\theta^+}(\widetilde{q}(z,a^*,b))$. Let σ be defined such that $\sigma_1 = a^*$ and for all $n \in \mathbb{N}^*$, $h_t^1 = (p, i_1, c_1, ...i_{t-1}, c_{t-1}) \in H_t^1$, $\sigma_t(h_t^1) = \sigma_{t-1}^*(c', h_{t-1}^{1,+})$ where $c' = (p, i_1, c_1)$ and $h_{t-1}^{1,+} = (i_2, c_2, ..i_t, c_t)$. We have

$$\gamma_\theta(\mu,\sigma,\tau) = \theta_1 \widetilde{g}(z,a^*,b) + (1-\theta_1)\gamma_{\theta^+}(Q(z,a^*,b),\sigma^*,\tau^+),$$

where τ^+ is a continuation strategy. Precisely, for all $n \in \mathbb{N}^*$, $\tau_{t-1}^+(d', h_{t-1}^{2,+}) = \tau_t(h_t^2)$ with $h_{t-1}^{2,+} = (j_2, d_2, .., j_t, d_t)$, $h_t^2 = (d', h_{t-1}^{2,+})$ and $d' = (j_1, d_1)$ is the signal to player 2 given by $\widetilde{q}(z,a^*,b)$.

Therefore σ^* and τ^+ can be seen as behavior strategies in a new game with initial signals corresponding to the past history in the original game. Since σ^* is ε-optimal in $\Gamma(\widetilde{q}(z,a^*,b))$, we have

$$\begin{aligned}
\gamma_\theta(\mu,\sigma,\tau) &\geq \theta_1 \widetilde{g}(z,a^*,b) + (1-\theta_1) v_{\theta^+}(Q(z,a^*,b)) - \varepsilon \\
&= \sup_{a\in A} \theta_1 \widetilde{g}(z,a,b) + (1-\theta_1) v_{\theta^+}(Q(z,a,b)) - 2\varepsilon \\
&= \sup_{a\in A} \theta_1 \widetilde{g}(z,a,b) + (1-\theta_1) \widetilde{v}_{\theta^+}(\widetilde{q}(z,a)) - 2\varepsilon \\
&\geq \min_{b\in B} \sup_{a\in A} \theta_1 \widetilde{g}(z,a,b) + (1-\theta_1) \widetilde{v}_{\theta^+}(\widetilde{q}(z,a)) - 2\varepsilon.
\end{aligned}$$

It follows that $\widetilde{v}_\theta(z) \geq \min_{b\in B} \sup_{a\in A} \theta_1 \widetilde{g}(z,a^*,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))$ by sending ε to zero.

Let us show that player 2 can defend $\sup_{a\in A} \min_{b\in B}(\theta_1 \widetilde{g}(z,a,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))$ in $\Gamma(\mu)$. Fix a strategy of player 1 and let $a = \sigma_1$, there exists $b^* \in B$ achieving $\min_b \widetilde{g}(z,a,b)$. We also choose τ^* an optimal strategy for player 2 in the game $\Gamma_{\theta^+}(\widetilde{q}(z,a,b^*))$.

$$\begin{aligned}
\gamma_\theta(\mu,\sigma,\tau) &= \theta_1 \widetilde{g}(z,a,b^*) + (1-\theta_1)\gamma_{\theta^+}(Q(z,a,b^*),\sigma^+,\tau^*) \\
&\leq \theta_1 \widetilde{g}(z,a,b^*) + (1-\theta_1) v_{\theta^+}(Q(z,a,b^*)) \\
&= \theta_1 \widetilde{g}(z,a,b^*) + (1-\theta_1) \widetilde{v}_{\theta^+}(\widetilde{q}(z,a)) \\
&= \min_{b\in B} \theta_1 \widetilde{g}(z,a,b) + (1-\theta_1) \widetilde{v}_{\theta^+}(\widetilde{q}(z,a))
\end{aligned}$$

Thus $\widetilde{v}_\theta(z) \leq \sup_{a\in A} \min_{b\in B}(\theta_1 \widetilde{g}(z,a,b) + (1-\theta_1)\widetilde{v}_{\theta^+}(\widetilde{q}(z,a))$. Finally the maxmin is always smaller than the minmax, so all the intermediate inequalities are equalities. □

4.4.3 Back to the repeated game.

For all $\pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N})$, we denote by $\hat{\pi} = \Phi(\pi)$ the projection of π on $\Delta_f(Z)$. We denote by $v_{m,n}(\pi) = v_{\theta_{m,n}}(\pi)$ the value with respect to the uniform evaluation between stage $m+1$ and stage $m+n$:

$$\forall \pi \in \Delta_f(K \times \mathbb{N} \times \mathbb{N}), \forall \sigma \in \Sigma, \tau \in \mathcal{T},\ \gamma_{m,n}(\pi) = I\!E_{\pi,\sigma,\tau}\left(\frac{1}{n}\sum_{t=m+1}^{m+n} r(k_t, i_t, j_t)\right),$$

and for all $z \in Z$, we denote by $\widetilde{v}_{m,n}(z)$, the value on the quotient. In order to conclude the proof, we show that both players can guarantee

$$v^*(\pi) = \inf_{n\geq 1} \sup_{m\geq 0} v_{m,n}(\pi).$$

The proof follows the same approach as Renault [Ren12b].
We first prove that player 1 can guarantee $v^*(\pi)$. We recall Renaults theorem [Ren12b].

Proposition 4.4.11 (Renault [Ren12b]) *Assume that $H1$,...,$H7'$ hold. Then for any initial distribution $\eta \in \Delta_f(Z)$, the stochastic game $\Psi(\eta)$ has a uniform value $w^*(\eta)$. Moreover player 1 can guarantee $w^*(\eta)$ with a Markov strategy :*

$$\forall \varepsilon > 0,\ \exists \sigma \in \Sigma^M, \exists N_0 \in \mathbb{N},\ \forall N \geq N_0\ \forall \tau' \in \mathcal{T},\ \gamma_N(\eta, \sigma, \tau') \geq w^*(\eta) - \varepsilon,$$

and the uniform value is characterized by $w^*(\eta) = \inf_{n\geq 1} \sup_{m\geq 0} w_{m,n}(\eta)$.

The stochastic game $\Psi(z) = \Psi(\delta_z)$ satisfies assumptions $H1, ..., H7'$ so it has a uniform value given by

$$w^*(z) = \inf_{n\geq 1} \sup_{m\geq 0} w_{m,n}(z).$$

And by proposition 4.4.9, for all evaluations the value in Ψ and in the reduced game are equal, so if $\pi \in \Delta^*_f(K \times \mathbb{N} \times \mathbb{N})$ we have

$$v^*(\pi) = \inf_{n\geq 1} \sup_{m\geq 0} v_{m,n}(\pi) = \inf_{n\geq 1} \sup_{m\geq 0} \widetilde{v}_{m,n}(\hat{\pi}) = \inf_{n\geq 1} \sup_{m\geq 0} w_{m,n}(\hat{\pi}) = w^*(\hat{\pi}).$$

Player 1 can guarantee $v^*(\pi)$ in $\Psi(\hat{\pi})$ with a Markov strategy. Let us check that he can guarantee $v^*(\pi)$ in the game $\widetilde{\Gamma}(\hat{\pi})$ or equivalently in $\Gamma(\pi)$.

Proposition 4.4.12 *Any markovian strategy $\widehat{\sigma}$ of player 1 in Ψ induces a strategy σ in Γ guaranteeing the same amount.*

Proof: Let $\widetilde{\sigma}$ be a behavior strategy in $\Psi(z)$. Let us describe the strategy σ. Player 1 plays at the first stage in $\Gamma(z)$ the mixed action $\widetilde{\sigma}_1(z)(p)$ where p is his initial signal. At stage $t \geq 2$, he is able to compute the value of $x_t(h^1_t)$ and $z_t(h^1_t)$ without knowing the strategy of his opponent using assumptions $(A1)$,$(A2)$. He plays the mixed action $\widetilde{\sigma}_t(z_t)(x_t)$.

It remains to prove that this strategy guarantees the same quantity as $\widetilde{\sigma}$. Fix $n \in \mathbb{N}^*$, we prove that there exists a best reply $\widetilde{\tau}$ to σ in $\widetilde{\Gamma}_\theta(z)$ which derives from a strategy $\hat{\tau}$ in $\Psi_\theta(z)$ and such that

$$\gamma_\theta(z, \sigma, \widetilde{\tau}) = \widetilde{\gamma}_\theta(z, \widetilde{\sigma}, \widetilde{\tau}),$$

where $\widetilde{\gamma}$ is the payoff in the stochastic game Ψ_θ.

We proceed by backward induction. Let τ be a best reply to σ in $\widetilde{\Gamma}_\theta(z)$. We build a strategy $\widetilde{\tau}$ which depends at stage m on h^2_m only through z_m. Recall that σ is fixed so that $z_m(h^2_m)$ can be computed by player 2. We replace τ_t by

$$\widetilde{\tau}_t(z_t) = \mathbb{E}_{z,\sigma,\tau}[\tau(h^2_t) \mid z_t].$$

Note that this conditional expectation depends on the strategies σ, τ up to stage $t-1$. Then the payoff at the last stage t is not modified.

$$\begin{aligned}
\mathbb{E}_{z,\sigma,\tau}[g(k_t,i_t,j_t)] &= \mathbb{E}_{z,\sigma,\tau}[\mathbb{E}_{z,\sigma,\tau}[g(k_t,i_t,j_t) \mid h_t^1, h_t^2]] \\
&= \mathbb{E}_{z,\sigma,\tau}[g(x_t,\sigma_t(x_t,z_t),\tau_t(h_t^2))] \\
&= \mathbb{E}_{z,\sigma,\tau}[\mathbb{E}_{z,\sigma,\tau}[g(x_t,\sigma_t(x_t,z_t),\tau_t(h_t^2)) \mid h_t^2]] \\
&= \mathbb{E}_{z,\sigma,\tau}\left[\int g(x,\sigma_t(x,z_t),\tau_t(h_t^2))dz_t[x]\right] \\
&= \mathbb{E}_{z,\sigma,\tau}\left[\mathbb{E}_{z,\sigma,\tau}\left[\int g(x,\sigma_t(x,z_t),\tau_t(h_t^2))dz_t[x] \mid z_t\right]\right] \\
&= \mathbb{E}_{z,\sigma,\tau}\left[\int g\left(x,\sigma_t(x,z_t),\mathbb{E}_{z,\sigma,\tau}[\tau_t(h_t^2) \mid z_t]\right)dz_t[x]\right] \\
&= \mathbb{E}_{z,\sigma,\tau}[\int g(x,\sigma_t(x,z_t),\widetilde{\tau}_t(z_t))dz_t[x]] \\
&= \mathbb{E}_{z,\sigma,(\tau_1,\ldots,\tau_{t-1},\widetilde{\tau}_t)}[g(k_t,i_t,j_t)]
\end{aligned}$$

The above equations show that the expected payoff at stage t when player 2 is playing a best reply against σ is a function of σ and of the law of z_t. Assume now that at step m, we have proved that there exists a best reply to σ of player 2 such that the sum of expected payoffs for the stages $m+1,...,n$ is a function of σ and of the law of z_{m+1} only. We can replace $\tau_m(h_m^2)$ by $\widetilde{\tau_m}(z_m) = \mathbb{E}_{z,\sigma,\tau}[\tau_m(h_m^2) \mid z_m]$ without modifying the expected payoff of stage m, and using assumption $(A3)$, the law of z_{m+1} is not modified by this operation which proves that this modified strategy is still a best reply to σ. □

Secondly, we prove that player 2 can guarantee $v^*(\pi)$ by splitting the game in blocks and playing on each block separately. He has no influence on the transition so what he is playing on one block has no influence on the next one.

Lemma 4.4.13 *For every $\pi \in \Delta(K \times C' \times D')$, $n \geq 1$ and $m \geq 0$, $\forall\ \tau_1,...,\tau_m$, $\exists\ \tau_{m+1},...,\tau_{m+n}$ such that any strategy of player 2 starting by $\tau_1,...,\tau_m,...,\tau_{m+n}$ is optimal in the game $\Gamma_{m,n}$.*

Proposition 4.4.14 *For every $\pi \in \Delta(K \times C' \times D')$, player 2 can guarantee $v^*(\pi)$ in the game $\Gamma(\pi)$.*

Proof: We prove that for all $n \in \mathbb{N}$, player 2 can guarantee the payoff $\sup_{m\geq 0} v_{m,n}(p)$. Let $n \in \mathbb{N}$ be a number of stages, then for each $l \in \mathbb{N}$ we split the game of length nl in l blocks of length t: $B_1,...,B_l$. We define the strategy τ^* by induction.

Let τ be an optimal strategy in $\Gamma_{1,n}(\pi)$ then we set $\tau_i^* = \tau_i$ for all $i \in \{1,..,n\}$. Given $\tau_1^*,...,\tau_{nl-1}^*$, we define the game $\Gamma^{\#}(\pi)$ where the player 2 has to play τ_i^* for all $i \leq nl-1$. We have $v^{\#}_{nl+1,(n+1)l}(\pi) \geq v_{nl+1,(n+1)l}(\pi)$ and player 2 can defend $v_{nl+1,(n+1)l}(\pi)$. Let τ be an optimal strategy in $\Gamma^{\#}_{nl+1,(n+1)l}$

and set $\tau_i^* = \tau_i$ for all $i \in [nl+1, (n+1)l]$. We have

$$\begin{aligned}
\gamma_{ln}(\sigma, \tau^*) &= \frac{1}{ln} E_{(\pi,\sigma,\tau^*)} \left(\sum_{t=1}^{ln} g(k_t, i_t, j_t) \right) \\
&= \frac{1}{ln} \sum_{d=0}^{l-1} E_{(\pi,\sigma,\tau^*)} \left(\sum_{t=dn+1}^{(d+1)n} g(k_t, i_t, j_t) \right) \\
&\leq \frac{1}{l} \sum_{d=0}^{l-1} v_{dn+1,n}(\pi) \\
&\leq \frac{1}{l} \sum_{d=0}^{l-1} \sup_{m \geq 0} v_{m,n}(\pi) \\
&\leq \sup_{m \geq 0} v_{m,n}(\pi).
\end{aligned}$$

Therefore player 2 can guarantee the minimum with respect to n

$$v^*(\pi) = \inf_{n \in \mathbb{N}} \sup_{m \geq 0} v_{m,n}(\pi).$$

□

Remark 4.4.15 In the second chapter, we showed the existence of the stronger notion of general limit value and general uniform value for POMDPs and repeated games with an informed controller by using a special metric d_* on $Z = \Delta_f(\Delta(K))$ computed with respect to the L_1-norm on $\Delta(K)$. Maybe one could iterate this definition and define d_{**} on $\Delta(Z)$ computed with respect to (Z, d_*) and study the general uniform value in the framework of a more informed controller.

Chapter 5

Asymptotic properties of optimal trajectories in dynamic programming

Résumé : On montre dans un contexte de programmation dynamique que la convergence uniforme des valeurs des problèmes de longueur finis implique qu'asymptotiquement le paiement moyen sur les stratégies optimales est constant. On discute ensuite les extensions possibles au cas des jeux avec deux joueurs et à somme nulle.

Ce chapitre est extrait d'un article écrit en collaboration avec Sylvain Sorin et Guillaume Vigeral, et publié dans *Sankhya A - Mathematical Statistics and Probability, Volume 72, Number 1 (2010).*

Abstract: We show in a dynamic programming framework that uniform convergence of the finite horizon values implies that asymptotically the average accumulated payoff is constant on optimal trajectories. We analyze and discuss several possible extensions to two-person games.

This chapter is extracted from an article written in collaboration with Sylvain Sorin and Guillaume Vigeral, and published in *Sankhya A - Mathematical Statistics and Probability, Volume 72, Number 1 (2010).*

5.1 Presentation

Consider a dynamic programming problem as described in Lehrer and Sorin [LS92]. Given a set of states Z, a correspondence F from Z to itself with non empty values and a payoff function r from Z to $[0,1]$, a feasible play at $z \in Z$ is a sequence $\{z_t\}$ of states with $z_1 \in F(z)$ and $z_{t+1} \in F(z_t)$. It induces a sequence of payoffs $\{r_t = r(z_t)\}, t = 1, ..., n, ...$ Recall that starting from a standard problem with random transitions and/or signals on the state, this presentation amounts to work on the set of probabilities on Z and to consider expected payoffs.

Let $v_n(z)$ (resp. $v_\lambda(z)$) be the value of the n-stage program $G_n(z)$ (resp. λ discounted program $G_\lambda(z)$): maximize $\frac{1}{n}\sum_{t=1}^{n} r_t$ (resp. $\sum_{t=1}^{+\infty} \lambda(1-\lambda)^{t-1} r_t$) over the set of feasible plays at z. The **asymptotic approach** deals with asymptotic properties of the values v_n and v_λ as n goes to ∞ or λ goes to 0.

The **uniform approach** focuses on properties of the strategies that hold uniformly in long horizons. v_∞ is the uniform value if for each $\varepsilon > 0$ and each $z \in Z$, there exists N such that:
1) there is a feasible play $\{z_t\}$ at z with

$$\frac{1}{n}\sum_{t=1}^{n} r(z_t) \geq v_\infty(z) - \varepsilon, \qquad \forall n \geq N,$$

2) for any feasible play $\{z'_t\}$ at z and any $n \geq N$,

$$\frac{1}{n}\sum_{t=1}^{n} r(z'_t) \leq v_\infty(z) + \varepsilon.$$

Obviously the second approach is more powerful than the first (existence of a uniform value $v_\infty(z)$ implies existence of an asymptotic value: $v(z)$ which is the limit of $v_n(z)$ and $v_\infty(z) = v(z)$) but it is also more demanding: there are problems without uniform value where the asymptotic value exists (see Section 2).

Note that the existence of a uniform value says that the average accumulated payoff on optimal trajectories remains close to the value.

We will study a related phenomenon in the asymptotic framework and consider the following property **P**:

There exists $w : Z \to \mathbb{R}$ satisfying: for any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$, for any state z and any feasible play $\{z_t\}$ ε-optimal for $G_n(z)$ and for any $l \in [0,1]$:

$$-3\varepsilon \leq \frac{1}{n}\left(\sum_{t=1}^{[ln]} r_t - l v(z)\right) \leq 3\varepsilon. \tag{5.1}$$

where $[ln]$ stands for the integer part of ln.

This condition says that the average payoff remains close to the value on every almost-optimal trajectory with long duration (but the trajectory may depend on this duration). It also implies a similar property on every time interval.

Say that the dynamic programming problem is **regular** if :

1) $\lim v_n(z) = v(z)$ exists for each $z \in Z$.
2) the convergence is uniform.

This condition was already introduced and studied in Lehrer and Sorin [LS92] (see Section 2). Note that bf P implies regularity.

Our main result is:

Theorem 5.1.1 *Assume that the program is regular, then* **P** *holds (with* $w = v$*).*

5.2 Examples and comments

1) The existence of the asymptotic value is not enough to control the payoff as required in property **P**. An example is given in Lehrer and Sorin [LS92] (Section 2), where $\lim v_n(z) = v(z)$ exists on Z but where the asymptotic average payoff is not constant on the unique optimal trajectory, nor on ε-optimal trajectories at some state z_0: in $G_{2n}(z_0)$, an optimal play will induce n times 0 then n times 1 while $v(z_0) = 1/2$.

Note that this example is not regular: the convergence of v_n to v is not uniform.

2) In the framework of dynamic programming, regularity is also equivalent to uniform convergence of v_λ (and with the same limit v), see Lehrer and Sorin [LS92] (Section 3).

Note also that this regularity condition is not sufficient to obtain the existence of a uniform value, see Monderer and Sorin [MS93] (Section 2).

3) General conditions for regularity can be found in Renault [Ren11].

5.3 Proof of the main result

Take $w = v$ and let us start with the upper bound inequality in (5.1).
The result is clear for $l \leq \varepsilon$ (recall that that the payoff is in $[0, 1]$). Otherwise let n_1 large enough so that $n \geq n_1$ implies $||v_n - v|| \leq \varepsilon$ by uniform convergence. Then the required inequality holds for $n \geq n_2$ with $[\varepsilon n_2] \geq n_1$.

Consider now the lower bound inequality in (5.1). The result holds for $l \geq 1 - \varepsilon$ by the ε-optimal property of the play, for $n \geq n_1$. Otherwise we use the following lemma from Lehrer and Sorin [LS92] (Proposition 1).

Lemma 5.3.1 *Both* $\limsup v_n$ *and* $\limsup v_\lambda$ *decrease on feasible histories.*

In particular, starting from $z_{[ln]}$ the value of the program for the last $n-[ln]$ stages is at most $v(z_{ln]})+\varepsilon$ for $n \geq n_2$, by uniform convergence, hence less than the initial $v(z) + \varepsilon$, using the previous Lemma. Since the play is ε-optimal in $G_n(z)$, this implies that

$$\sum_{t=1}^{[ln]} r_t + (n - [ln])(v(z) + \varepsilon) \geq n(v_n(z) - \varepsilon) \geq n(v(z) - 2\varepsilon) \tag{5.2}$$

hence the required inequality. □

5.4 Extensions

5.4.1 Discounted case

A similar result holds for the program G_λ corresponding to the evaluation $\sum_{t=1}^{\infty} \lambda(1-\lambda)^{t-1} r_t$. Explicitly, one introduces the property $\mathbf{P'}$:

There exists $w : Z \to \mathbb{R}$ satisfying: for any $\varepsilon > 0$, there exists λ_0, such that for all $\lambda \leq \lambda_0$, for any state z and any feasible play $\{z_t\}$ ε-optimal for $G_\lambda(z)$ and for any $t \in [0,1]$:

$$-3\varepsilon \leq \sum_{t=1}^{n(l;\lambda)} \lambda(1-\lambda)^{t-1} r_t) - lw(z) \leq 3\varepsilon. \tag{5.3}$$

where $n(l;\lambda) = \inf\{p \in; \sum_{t=1}^{p} \lambda(1-\lambda)^{t-1} \geq l\}$. Stage $n(l;\lambda)$ corresponds to the fraction l of the total duration of the program G_λ.

Theorem 5.4.1 *Assume that the program is regular, then* **P'** *holds (with* $w = v$*).*

Proof. The proof follows the same lines than the proof of Theorem 5.1.1.

Recall that by regularity both v_n and v_λ converge uniformly to v. Moreover the discounted sums $(1-\lambda)^{-N} \sum_{t=1}^{N} \lambda(1-\lambda)^{t-1} r_t$ belong to the convex hull of the averages $\frac{1}{n}\sum_{t=1}^{n} r_t; 1 \leq n \leq N$, see Lehrer and Sorin [LS92]. The counterpart of equation (5.2) is now

$$\sum_{t=1}^{n(l;\lambda)} \lambda(1-\lambda)^{t-1} r_t + (1-l)(v(z)+\varepsilon) \geq (v_\lambda(z) - \varepsilon) \geq v(z) - 2\varepsilon \tag{5.4}$$

and the result follows. □

5.4.2 Continuous time

Similar results hold in the following set-up: let $v_T(x)$ be the value of the control problem $\Gamma_T(x)$ with control set A where the state variable in X is governed by a differential equation (or more generally a differential inclusion)

$$\dot{x}_t = f(x_t, a_t)$$

starting from x at time 0. The real payoff function is $g(x,a)$ and the evaluation is given by:

$$\frac{1}{T}\int_0^T g(x_t, a_t)dt.$$

Regularity in this framework amounts to uniform convergence (on X) of V_T to some V. (Sufficient conditions for regularity can be found in Quincampoix and Renault [QR11]). The corresponding property is now **P"**:

There exists $W : X \to \mathbb{R}$ satisfying: for any $\varepsilon > 0$, there exists T_0, such that for all $T \geq T_0$, for

any state x and any feasible trajectory ε-optimal for $\Gamma_T(x)$ and for any $l \in [0,1]$:

$$-3\varepsilon \leq \frac{1}{T}\int_0^{lT} g(x_t, a_t)dt - lW(x) \leq +3\varepsilon. \tag{5.5}$$

Theorem 5.4.2 *Assume that the optimal control problem is regular, then* **P"** *holds (with* $W = V$*).*

Proof Follow exactly the same lines than the proof of Theorem (5.2).□

Finally similar tools can be used for an evaluation of the form

$$\lambda \int_0^{+\infty} e^{-\lambda t} g(x_t, a_t)dt,$$

see Oliu-Barton and Vigeral [OBM].

5.5 Two-player zero-sum games

In trying to extend this result to a two-person zero-sum framework, several problems occur.

5.5.1 Optimal strategies on both sides

First it is necessary, to obtain good properties on the trajectory, to ask for optimality on both sides. For example consider the Big Match with no signals, which is a stochastic game described by the matrix

	α	β
a	1*	0*
b	0	1

where a $*$ denotes an absorbing payoff. Assume that the players receive no information during the play. Then the asymptotic properties of the repeated game can be analyzed trough an "asymptotic game" played on $[0,1]$, see Sorin [Sor02] (Section 5.3.2.) and Sorin [Sor05] (Section 4) and the optimal strategy of player 1 is to play "a before time l" with probability l. Obviously, without restriction on player 2's moves, the average payoff along the play will not be closed to the asymptotic value $v = \frac{1}{2}$. (For example if Player 2 plays α during the first half of the game the corresponding average payoff at time $t = \frac{1}{2}$ is $\frac{1}{4}$). However, the optimal strategy of player 2 is "always $(1/2, 1/2)$" hence time independent on $[0,1]$ and thus induces a constant payoff.

5.5.2 Player 1 controls the transition.

Consider a repeated game with finite characteristics (states, moves, signals, ...) and use the recursive formula for the values corresponding to the canonical representation with entrance laws being consistent probabilities on the universal belief space, see Mertens, Sorin and Zamir [MSZ94], Chapters III.1, IV.3. This representation preserves the values but in the auxiliary game, if player 1 controls the transition an optimal strategy of player 2 is to play a stage by stage best reply. Hence

the model reduces to the dynamic programming framework and the results of the previous sections apply.

A simple example corresponds to a game with incomplete information on one side where asymptotically an optimal strategy of the uniform player 1 is a splitting at time 0, while player 2 can obtain $u(p_t)$ at time t where u is the value of the non-revealing game and p_t the martingale of posteriors at time t, see Sorin [Sor02], 3.7.2.

Another class corresponds to the Markov games with incomplete information introduced by Renault [Ren06].

5.5.3 Example.

Back to the general framework of two person zero-sum repeated games, the following example shows that in addition one has to strengthen the conditions on the pair of ε-optimal strategies. We exhibit a regular game where for some state z with $v(z) = 0$ one can construct, for each n, optimal strategies in $\Gamma_n(z)$ inducing roughly a constant payoff 1 during the first half of the game.

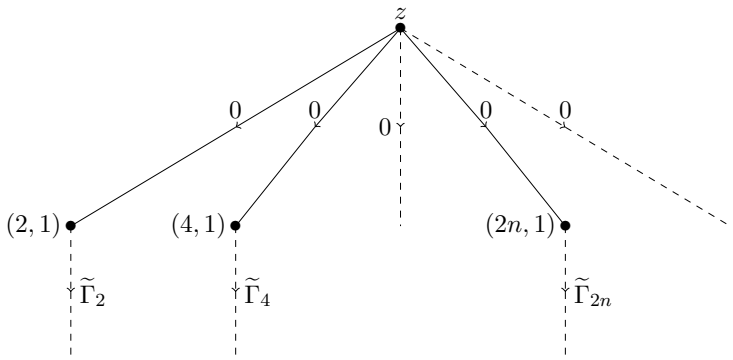

FIGURE 1. The game Γ starting from state z

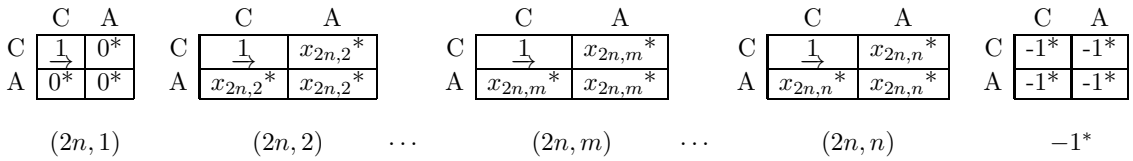

FIGURE 2. The subgame $\widetilde{\Gamma}_{2n}$ starting from state $(2n, 1)$

Starting from the initial state z, the tree representing the game Γ has countably many subgames $\tilde{\Gamma}_{2n}$, $n = 1, ...$, the transition being controlled by player 1 (with payoff 0). In $\tilde{\Gamma}_{2n}$ there are at most n stages before reaching an absorbing state. At each of these stages of the form $(2n, m), m = 1, ...n$, the players play a "jointly controlled" process leading either to a payoff 1 and the next stage $(2n, m + 1)$ (if they agree) or an absorbing payoff $x_{2n,m}$ with $(m - 1) + (2n - (m - 1))x_{2n,m} = 0$, otherwise. At

stage $(2n, n+1)$, the payoff is -1 and absorbing. Hence, every feasible path of length $2n$ in $\tilde{\Gamma}_{2n}$ gives a total payoff 0. Obviously the game is regular since each player can stop the game at each node $(2n, m)$, inducing the same absorbing payoff $x_{2n,m}$. The representation is as shown in figure 5.5.3.

Notice that in the $2n+1$ stage game, after a move of player 1 to $\tilde{\Gamma}_{2n}$, any play is compatible with optimal strategies, in particular those leading to the sequence of payoffs $2n$ times 0 or n times 1 then n times -1.

5.5.4 Conjectures.

A natural conjecture is that in any regular game (i.e. where v_n converges uniformly to v):
for any $\varepsilon > 0$, there exists n_0, such that for all $n \geq n_0$, for any initial state z, there exists a couple (σ_n, τ_n) of ε-optimal strategies in $G_n(z)$ such that for any $l \in [0,1]$:

$$-3\varepsilon \leq \frac{1}{n} I\!E_{z,\sigma_n,\tau_n} \left(\sum_{t=1}^{[ln]} r_t \right) - lv(z) \leq +3\varepsilon. \tag{5.6}$$

where $[ln]$ stands for the integer part of ln and r_t is the payoff at stage t.

A more elaborate conjecture would rely on the existence of an asymptotic game Γ^* played in continuous time on $[0,1]$ with value v (as in Section 5.1), see Sorin [Sor05] (Section 4). We use the representation of the repeated game as a stochastic game trough the recursive structure as above, see Mertens, Sorin, Zamir [MSZ94] (Chapter IV).

The condition is now the existence of a couple of strategies (σ, τ) in the asymptotic game that would depend only on the time $l \in [0,1]$ and on the current state z and which would generate a constant payoff along the play. Then the couple (σ_n, τ_n) would correspond to the strategies induced by (σ, τ) for a discretization of $[0,1]$ of width $\frac{1}{n}$

Acknowledgment: This work was done while the three authors were members of the Equipe Combinatoire et Optimisation. Sorin's research was supported by grant ANR-08-BLAN-0294-01 (France).

Bibliography

[ABFG+93] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus. Discrete-time controlled markov processes with average cost criterion: a survey. SIAM Journal on Control and Optimization, 31(2):282–344, 1993.

[AMS95] R.J. Aumann, M. Maschler, and R.E. Stearns. Repeated games with incomplete information. The MIT press, 1995.

[Ash72] R.B. Ash. Real analysis and probability, volume 239. Academic Press New York, 1972.

[Ast65] K.J. Aström. Optimal control of markov processes with incomplete state information. J. Math. Anal. Appl., 10:174–205, 1965.

[Aub77] J.P. Aubin. Applied abstract analysis. John Wiley & Sons, 1977.

[BD83] D. Blackwell and L.E. Dubins. An extension of Skorohod's almost sure representation theorem. Proceedings of the American Mathematical Society, 89(4):691–692, 1983.

[Bel57] R. Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

[BF68] D. Blackwell and T. S. Ferguson. The big match. Ann. Math. Statist, 39:159–163, 1968.

[Bir31] G.D. Birkhoff. Proof of the ergodic theorem. Proceedings of the National Academy of Sciences of the United States of America, 17(12):656, 1931.

[BK76a] T. Bewley and E. Kohlberg. The asymptotic solution of a recursion equation occurring in stochastic games. Mathematics of Operations Research, pages 321–336, 1976.

[BK76b] T. Bewley and E. Kohlberg. The asymptotic theory of stochastic games. Mathematics of Operations Research, pages 197–208, 1976.

[Bla62] D. Blackwell. Discrete dynamic programming. Ann. Math. Statist., 33:719–726, 1962.

[Bor00] V.S. Borkar. Average cost dynamic programming equations for controlled markov chains with partial observations. SIAM Journal on Control and Optimization, 39:673, 2000.

[Bor02] V.S. Borkar. Convex analytic methods in markov decision processes. Handbook of Markov decision processes, pages 347–375, 2002.

[Bor07] V.S. Borkar. Dynamic programming for ergodic control of markov chains under partial observations: a correction. SIAM Journal on Control and Optimization, 45(6):2299–2304, 2007.

[Cho56] G. Choquet. Existence et unicité des représentations intégrales au moyen des points extrémaux dans les cônes convexes. Séminaire Bourbaki, 4:33–47, 1956.

[Cou92] J.M. Coulomb. Repeated games with absorbing states and no signals. International Journal of Game Theory, 21(2):161–174, 1992.

[Cou03] J.M. Coulomb. Stochastic games without perfect monitoring. International Journal of Game Theory, 32(1):73–96, 2003.

[DF68] E.V. Denardo and B.L. Fox. Multichain markov renewal programs. SIAM Journal on Applied Mathematics, 16(3):468–487, 1968.

[DJ79] E.B. Dynkin and A.A. Juškevič. Controlled markov processes, volume 235. Springer New York, 1979.

[DS65] L.E. Dubins and L.J. Savage. How to gamble if you must: Inequalities for stochastic processes. McGraw-Hill New York, 1965.

[Eve57] H. Everett. Recursive games. Contributions to the Theory of Games III, 39:47–78, 1957.

[For82] F. Forges. Infinitely repeated games of incomplete information: Symmetric case with random signals. International Journal of Game Theory, 11(3):203–213, 1982.

[FSV08] J. Flesch, G. Schoenmakers, and K. Vrieze. Stochastic games on a product state space. Mathematics of Operations Research, 33(2):403–420, 2008.

[FSV09] J. Flesch, G. Schoenmakers, and K. Vrieze. Stochastic games on a product state space: The periodic case. International Journal of Game Theory, 38(2):263–289, 2009.

[Gei02] J. Geitner. Note Equilibrium payoffs in stochastic games of incomplete information: the general symmetric case. International Journal of Game Theory, 30(3):449–452, 2002.

[Gil57] D. Gillette. Stochastic games with zero stop probabilities. Ann. Math. Stud, 39:178–187, 1957.

[Har67] J.C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part i. the basic model. Management science, pages 159–182, 1967.

[HK79] A. Hordijk and LCM Kallenberg. Linear programming and markov decision chains. Management Science, pages 352–362, 1979.

[HL89] O. Hernández-Lerma. Adaptive Markov control processes, volume 79. Springer New York, 1989.

[HRSV10] J. Hörner, D. Rosenberg, E. Solan, and N. Vieille. On a markov game with one-sided information. Operations research, 58(4):1107–1115, 2010.

[Kal94] L.C.M. Kallenberg. Survey of linear programming for standard and nonstandard markovian control problems. part i: Theory. Mathematical Methods of Operations Research, 40(1):1–42, 1994.

[Koh74] E. Kohlberg. Repeated games with absorbing states. The Annals of Statistics, 2(4):724–738, 1974.

[Kuh53] H.W. Kuhn. Extensive games and the problem of information. Contributions to the Theory of Games, 2(28):193–216, 1953.

[KZ74] E. Kohlberg and S. Zamir. Repeated games of incomplete information: The symmetric case. The Annals of Statistics, 2(5):1040–1041, 1974.

[LL69] T.M. Liggett and S.A. Lippman. Short notes: Stochastic games with perfect information and time average payoff. Siam Review, 11(4):604–607, 1969.

[LS92] E. Lehrer and S. Sorin. A uniform Tauberian theorem in dynamic programming. Mathematics of Operations Research, pages 303–307, 1992.

[Mer87] J.-F. Mertens. Repeated games. In Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986), pages 1528–1577, Providence, RI, 1987. Amer. Math. Soc.

[MN81] J.-F. Mertens and A. Neyman. Stochastic games. Internat. J. Game Theory, 10(2):53–66, 1981.

[MNR09] J.F. Mertens, A. Neyman, and D. Rosenberg. Absorbing games with compact action spaces. Math Oper Res, 34:257–262, 2009.

[MS93] D. Monderer and S. Sorin. Asymptotic properties in dynamic programming. International Journal of Game Theory, 22(1):1–11, 1993.

[MS96] A. Maitra and W. Sudderth. Discrete gambling and stochastic games, volume 32. Springer Verlag, 1996.

[MSZ94] J.F. Mertens, S. Sorin, and S. Zamir. Repeated games. CORE Discussion Papers, 9420,9421,9422, 1994.

[MZ80] J.-F. Mertens and S. Zamir. Minmax and maxmin of repeated games with incomplete information. International Journal of Game Theory, 9(4):201–215, 1980.

[MZ85] J.-F. Mertens and S. Zamir. Formulation of bayesian analysis for games with incomplete information. International Journal of Game Theory, 14(1):1–29, 1985.

[MZ72] J.-F. Mertens and S. Zamir. The value of two-person zero-sum repeated games with lack of information on both sides. Internat. J. Game Theory, 1:39–64, 1971/72.

[Ney03] A. Neyman. Stochastic games and nonexpansive mappings. chapter 26 in a.neyman and s.sorin (eds). Stochastic games and Applications, pages 397–415, 2003.

[Ney08] A. Neyman. Existence of optimal strategies in markov games with incomplete information. International Journal of Game Theory, 37(4):581–596, 2008.

[NS98] A. Neyman and S. Sorin. Equilibria in repeated games of incomplete information: the general symmetric case. Internat. J. Game Theory, 27(2):201–210, 1998.

[OBM] Vigeral G. Oliu-Barton M. A uniform tauberian theorem in optimal control. http://arxiv.org/abs/1004.4174v1.

[QR11] M. Quincampoix and J. Renault. On the existence of a limit value in some nonexpansive optimal control problems. SIAM Journal on Control and Optimization, 49:2118, 2011.

[Ren06] J. Renault. The value of Markov chain games with lack of information on one side. Math. Oper. Res., 31(3):490–512, 2006.

[Ren11] J. Renault. Uniform value in dynamic programming. J. Eur. Math. Soc. (JEMS), 13(2):309–330, 2011.

[Ren12a] J. Renault. General long-term values in dynamic programming. mimeo, 2012.

[Ren12b] J. Renault. The value of repeated games with an informed controller. Mathematics of operations Research, 37:154–179, 2012.

[Rhe74] D. Rhenius. Incomplete information in markovian decision models. The Annals of Statistics, 2(6):1327–1334, 1974.

[Ros00] D. Rosenberg. Zero sum absorbing games with incomplete information on one side: asymptotic analysis. SIAM Journal on Control and Optimization, 39:208, 2000.

[RS01] D. Rosenberg and S. Sorin. An operator approach to zero-sum repeated games. Israel Journal of Mathematics, 121(1):221–246, 2001.

[RSV02] D. Rosenberg, E. Solan, and N. Vieille. Blackwell optimality in Markov decision processes with partial observation. Ann. Statist., 30(4):1178–1193, 2002.

[RSV03] D. Rosenberg, E. Solan, and N. Vieille. The maxmin value of stochastic games with imperfect monitoring. International Journal of Game Theory, 32(1):133–150, 2003.

[RSV04] D. Rosenberg, E. Solan, and N. Vieille. Stochastic games with a single controller and incomplete information. SIAM J. Control Optim., 43(1):86–110 (electronic), 2004.

[RV00] D. Rosenberg and N. Vieille. The maxmin of recursive games with incomplete information on one side. Mathematics of Operations Research, pages 23–35, 2000.

[Sch93] M. Schal. Average optimality in dynamic programming with general state space. Mathematics of Operations Research, pages 163–172, 1993.

[SF92] R. Sznajder and J.A. Filar. Some comments on a theorem of hardy and littlewood. Journal of optimization theory and applications, 75(1):201–208, 1992.

[Sha53] L.S. Shapley. Stochastic games. Proc. Nat. Acad. Sci. U. S. A., 39:1095–1100, 1953.

[Sin90] R. Sine. A nonlinear Perron-Frobenius theorem. Proc. Amer. Math. Soc., 109(2):331–336, 1990.

[Sio58] M. Sion. On general minimax theorems. Pacific J. Math., 8:171–176, 1958.

[Sor84] S. Sorin. "big match" with lack of information on one side (i). International Journal of Game Theory, 13(4):201–255, 1984.

[Sor85] S. Sorin. "big match" with lack of information on one side (ii). International Journal of Game Theory, 14(3):173–204, 1985.

[Sor02] S. Sorin. A first course on zero-sum repeated games, volume 37. Springer, 2002.

[Sor03] S. Sorin. The operator approach to zero-sum stochastic games. Stochastic Games and Applications, NATO Science Series C, Mathematical and Physical Sciences, 570:417–426, 2003.

[Sor05] S. Sorin. New approaches and recent advances in two-person zero-sum repeated games. Advances in Dynamic Games, pages 67–93, 2005.

[SY70] Y. Sawaragi and T. Yoshikawa. Discrete-time markovian decision processes with incomplete state observation. The Annals of Mathematical Statistics, 41(1):78–86, 1970.

[SZ85] S. Sorin and S. Zamir. A 2-person game with lack of information on 1 1/2 sides. Mathematics of operations research, pages 17–23, 1985.

[SZ91] S. Sorin and S. Zamir. " big match" with lack of information on one side (iii). Raghavan et al.[14], pages 101–112, 1991.

[TV92] F. Thuijsman and K. Vrieze. Note on recursive games. Lecture Notes in Economics and Mathematical Systems, pages 133–133, 1992.

[Vie00a] N. Vieille. Two player stochastic games. I. A reduction. Israel J. Math., 119:55–91, 2000.

[Vie00b] N. Vieille. Two-player stochastic games II: The case of recursive games. Israel Journal of Mathematics, 119(1):93–126, 2000.

[Vil03] C. Villani. Topics in optimal transportation, volume 58. Amer Mathematical Society, 2003.

[VN28] J. Von Neumann. Zur theorie der gesellschaftsspiele. Mathematische Annalen, 100(1):295–320, 1928.

[VN32] J. Von Neumann. Proof of the quasi-ergodic hypothesis. Proceedings of the National Academy of sciences of the United States of America, 18(1):70, 1932.

Résumé

Dans cette thèse, nous nous intéressons à un modèle général de jeux répétés à deux joueurs et à somme nulle et en particulier au problème de l'existence de la valeur uniforme. Un jeu répété a une valeur uniforme s'il existe un paiement que les deux joueurs peuvent garantir, dans tous les jeux commençant aujourd'hui et suffisamment longs, indépendamment de la longueur du jeu.

Dans un premier chapitre, on étudie les cas d'un seul joueur, appelé processus de décision Markovien partiellement observable, et des jeux où un joueur est parfaitement informé et contrôle la transition. Il est connu que ces jeux admettent une valeur uniforme. En introduisant une nouvelle distance sur les probabilités sur le simplexe de $\mathbb{R}^m$, on montre l'existence d'une notion plus forte où les joueurs garantissent le même paiement sur n'importe quel intervalle de temps suffisamment long et non pas uniquement sur ceux commençant aujourd'hui.

Dans les deux chapitres suivants, on montre l'existence de la valeur uniforme dans deux cas particuliers de jeux répétés : les jeux commutatifs dans le noir, où les joueurs n'observent pas l'état mais l'état est indépendant de l'ordre dans lequel les actions sont jouées, et les jeux avec un contrôleur plus informé, où un joueur est plus informé que l'autre joueur et contrôle l'évolution de l'état.

Dans le dernier chapitre, on étudie le lien entre la convergence uniforme des valeurs des jeux en n étapes et le comportement asymptotique des stratégies optimales dans ces jeux en n étapes. Pour chaque n, on considère le paiement garanti pendant ln étapes avec $0 < l < 1$ par les stratégies optimales pour n étapes et le comportement asymptotique lorsque n tend vers l'infini.

Mots-clés : Théorie des Jeux, Valeur uniforme, Processus de Décision Markoviens Partiellement Observable, Jeux répétés, Jeux stochastiques.

Abstract

In this dissertation, we consider a general model of two-player zero-sum repeated game and particularly the problem of the existence of a uniform value. A repeated game has a uniform value if both players can guarantee the same payoff in all games beginning today and sufficiently long, independently of the length of the game.

In a first chapter, we focus on the cases of one player, called Partial Observation Markov Decision Processes, and of Repeated Games where one player is perfectly informed and controls the transitions. It is known that these games have a uniform value. By introducing a new metric on the probabilities over a simplex in $\mathbb{R}^m$, we show the existence of a stronger notion, where the players guarantee the same payoff on all sufficiently long intervals of stages and not uniquely on the one starting today.

In the next two chapters, we show the existence of the uniform value in two special models of repeated games : commutative repeated games in the dark, where the players do not observe the state variable, but the state is independent of the order the actions are played, and repeated games with a more informed controller, where one player controls the transition and has more information than the second player.

In the last chapter, we study the link between the uniform convergence of the value of the n-stage games and the asymptotic behavior of the sequence of optimal strategies in the n-stage game. For each n, we consider n-stage optimal strategies and the payoff they are guaranteeing during the ln first stages with $0 < l < 1$. We study the asymptotic of this payoff when n goes to infinity.

Keywords : Game Theory, Uniform value, Partial Observation Markov Decision Processes, Repeated Games, Stochastic Games.