Contributions on Online Learning in Stochastic Games

Lucas Baudin

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris-Dauphine

Contributions à l'apprentissage en ligne dans les jeux stochastiques

Soutenue par
Lucas BAUDIN
Le 28 septembre 2023

Ecole doctorale n° ED 543
Ecole doctorale SDOSE

Spécialité
Informatique

Dauphine | PSL UNIVERSITÉ PARIS

Composition du jury :

Mathieu, FAURE
Professeur, Aix-Marseille Université
Président

Johanne, COHEN
Directrice de recherche au CNRS, Université Paris-Saclay
Examinatrice

David, LESLIE
Professeur, Université de Lancaster
Rapporteur

Panayotis, MERTIKOPOULOS
Directeur de recherche au CNRS, Laboratoire d'informatique de Grenoble
Rapporteur

Rida, LARAKI
Directeur de recherche au CNRS, Université Paris-Dauphine
Directeur de thèse

Laurent, GOURVÈS
Directeur de recherche au CNRS, Université Paris-Dauphine
Directeur de thèse

Guillaume, VIGERAL
Maître de conférences, Université Paris-Dauphine
Directeur de thèse

Contents

Symbols

General

Notation	Description	Page List
$\boldsymbol{A}$	Pure action profiles, that is $\Pi_{i\in I} A^i$	16, 39–41
A^i	Set of actions of every player i	16, 40
A	Set of actions of the player in the case of one-player games (i.e., MDP)	41
$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$	Typical pure action profiles	40, 107
a^i, b^i, c^i	Typical pure action of player i	40
$\boldsymbol{a}_n$	Action profile played at time n	16, 40
a_n^i	Action played at time n by player i	40, 107
α_n	Update steps of either Q-values in Q-learning (and in this case denoted α as it is supposed constant) or the continuations in FP for stochastic games	6, 22, 55, 61, 71
$\mathrm{BR}^i(\boldsymbol{x}_n^{-i})$	Best-Response of player i against (mixed profile) $\boldsymbol{x}_n^{-i}$	18, 49, 72
δ	Discount factor, $\delta = 0$ when players are only interested in the immediate utility, $\delta = 1$ when there is no discounting	40, 64
$\mathcal{G}$	A stochastic game	41
$\mathcal{H}$	Set of all histories, that is $\bigcup_{n\geq 0}(\boldsymbol{A}\times S)^n \times S$ in stochastic games or $\bigcup_{n\geq 0} \boldsymbol{A}^n$ in repeated games	42
h_n	A typical history, i.e., state-action profile pairs of previous steps	42, 109
I	The set of players and its cardinal when it is not ambiguous	39–41
i, j	Typical players in I	40
n	Discrete time variable	69
P_s	Transition function $\boldsymbol{A} \to \Delta(S)$ of a stochastic game	41
$R_n^i(b^i, c^i)$	In the context of repeated games, internal regret of player i up to time n where action b^i is replaced by c^i	52
r_n^i	In the context of repeated games, external regret of player i up to time n	19, 50
S	Set of states	41, 108
s, s'	Typical states	41

Notation	Description	Page List
$SN(G)$	Best social payoff of Nash equilibria of game G	22, 61
$SO(G)$	Best social payoff in game G	22, 61
s_n	State of the system at step n	69
t	Continuous time variable	73
u^i	Utility function of player i	40
u^i_s	Stage utility function in state s in a stochastic game	41
$\boldsymbol{y}$	Typical mixed action profile, with y^i the mixed action of player i and $y^i(a^i)$ the probability for i to play a^i	40, 109

Fictitious Play and Best-Response Dynamics

Notation	Description	Page List
$\alpha(t)$	Update rate of the continuations in best-response dynamics	73–75
α_n	Update steps of either Q-values in Q-learning (and in this case denoted α as it is supposed constant) or the continuations in FP for stochastic games	55
β_-	Minimum update rate for a state in best-response dynamics in ergodic games	74, 88
$\beta_s(t)$	Update rate of state s in best-response dynamics, equivalent to the visiting frequency in discrete-time systems	74, 75
η	Regularization parameter in action choices in smooth procedures (the higher η, the smoother the procedure)	85, 86, 115
f^i_{s,v^i}	Utility function of the auxiliary Shapley game	71
h^i	Regularizer function (for instance the entropy)	26, 85, 86
$\mu_{s,n}$	Number of times that s was reached up to time n	70
$\hat{\mathrm{SBR}}^i_{s,v}(\boldsymbol{x}_s^{-i})$	Estimated smooth best-response of player i to empirical profile $\boldsymbol{x}_s^{-i}$ in state s with v for continuations	87
$\mathrm{SBR}^i_{s,v}(\boldsymbol{x}_s^{-i})$	Smooth best-response of player i to empirical profile $\boldsymbol{x}_s^{-i}$ in state s with v for continuations	86
σ_n	Sum of the update steps α_n	71
$v^i_{s,n}$, v^i_s	Continuation value of the auxiliary Shapley game starting from state s, depending on time or not	71
$v^i_s(t)$, v^i_s	Continuation value of the auxiliary Shapley game in continuous time	73
$x^i_s(t)$	Empirical action of player i seen as an element of $\Delta(A^i)$, in continuous-time dynamics	73
$x^i_{s,n}$	Empirical action of player i up to time n, that is the average of a^j seen as an element of $\Delta(A^i)$	70

Queueing Systems

Notation	Description	Page List
C	Penalty incurred by a player whose job arrives after the end of a period (denoted by C_{k_t} when it is time-dependent)	108
$c^i(\boldsymbol{a})$ or $c^i_{\boldsymbol{k}_t}(\boldsymbol{a})$	Cost of player i if players follow profile $\boldsymbol{a}$ with $\boldsymbol{k}_t$ left implicit when it is not ambiguous	108
k_t, k^i_t	Number of jobs of all players, of player i, at time t	108
k^i	Number of jobs of players i	108
$\boldsymbol{k}_t$	State of the queueing system at period t, i.e., a vector with the number of jobs of every player	108
$\tilde{k}^i$	Number of jobs of player i which arrive late	29, 116
k	Total number of jobs	108
L	Length of each period	108
$\lambda_{t,n}$	Number of times between 1 and t that k^i_t is equal to n	117
$p^i_{\boldsymbol{k}_t}(\boldsymbol{a})$	Probability for a job of i to be late assuming that the players action profile is $\boldsymbol{a}$	108
$\mathsf{SC}(\boldsymbol{y})$	Social cost of action profile $\boldsymbol{y}$, that is $\sum_{i\in I} c^i(\boldsymbol{y})$	109
τ^i_n	First time that player i has n jobs to dispatch	115
$W^i_{\tau^i_n,n}(b^i)$	Initial weight of action b^i of player i in level n in the multi-level EWA (MLEWA)	115
$w^i_t(b^i)$	Weight of action b^i of player i in the exponential weight algorithm (EWA) at time t	115
$w^i_{t,n}\big(b^i\big)$	Weight of action b^i of player i in level n in the MLEWA at time t	115

Remerciements

Je souhaite tout d'abord remercier Laurent Gourvès, Rida Laraki et Guillaume Vigeral d'avoir accepté d'encadrer cette thèse alors que j'étais novice en théorie des jeux. Leur soutien a été précieux tout au long de la thèse, de la définition du sujet à la relecture du manuscrit, en passant par les riches discussions scientifiques . Cette thèse a été réalisée en grande partie pendant la phase la plus aiguë de l'épidémie de Covid-19 ; leur disponibilité m'a permis de continuer à travailler malgré le contexte difficile.

Je remercie aussi Xavier Venel et Marco Scarsini de m'avoir invité à LUISS et d'avoir consacré tant de temps à notre projet de recherche. Juste après les confinements de l'épidémie de Covid-19, ce voyage fut l'occasion de considérer mes questions sous un jour nouveau. J'ai beaucoup appris en travaillant avec eux.

Enfin, je remercie toutes celles et ceux qui m'ont accompagné pendant cette thèse à Dauphine puis à l'ENSAE.

Chapitre 1

Résumé en français

1.1 Introduction

1.1.1 Apprentissage en ligne dans les jeux stochastiques

Cette thèse a l'ambition d'être une étape supplémentaire dans la compréhension de la dynamique des systèmes multi-agents dans lesquels les agents apprennent via des algorithmes. De manière formelle, il s'agit d'apprentissage en ligne (*online learning* en anglais) dans les jeux stochastiques, à l'intersection entre la théorie des jeux et l'informatique.

L'apprentissage en ligne est un champ des mathématiques et de l'informatique dans lequel on cherche à optimiser une fonction d'utilité ou de perte tout en interagissant avec l'environnement. On suppose habituellement que ces interactions prennent la forme d'une suite d'actions prises par un agent qui utilise un algorithme. À la différence du cadre plus général de l'apprentissage, toute l'information n'est pas disponible immédiatement. Elle est progressivement révélée – à chaque étape, un agent choisit une action et observe ensuite ce qu'elle lui rapporte. Parfois, l'agent ne connaît même pas ce qu'aurait rapporté une action différente et il lui faut alors l'estimer.

La théorie des jeux est un sous-champ des mathématiques appliquées et de l'économie théorique. Elle est consacrée à l'étude des interactions stratégiques entre plusieurs agents, appelés joueurs. Un jeu *stochastique* est un jeu, aussi qualifié de dynamique, dans lequel un environnement, dont dépendent les fonctions de paiement, est modélisé par une variable d'état dont l'évolution peut-être influencée par les joueurs.

Cette thèse étudie des procédures d'apprentissage, certaines originales et d'autres déjà connues, qui peuvent être utilisées par des agents qui interagissent dans un environnement modélisé par un jeu stochastique. Nous analysons les dynamiques résultant de ces systèmes, par exemple en prouvant que le comportement moyen des agents converge vers un équilibre quand l'interaction dure de plus en plus longtemps.

Formalisation Un jeu stochastique est caractérisé par ses fonctions de transition et d'utilité, notées respectivement P_s et u_s^i (paramétrées par l'état actuel s et le joueur i) et dont l'argument est l'ensemble des actions prises par les joueurs. Une séquence de jeu est indexée par le temps n et commence dans un état s_0. À chaque étape, le système est dans l'état s_n, chaque joueur i choisit une action a_n^i parmi son ensemble d'actions A^i, qui, avec l'état courant s_n, permet de déterminer le paiement (aussi appelé utilité) de chaque joueur et l'état suivant s_{n+1}. Ceci est détaillé dans le modèle 1 et avec plus de détails dans le chapitre 3.

On suppose que les joueurs souhaitent optimiser leur paiement de long terme, qui peut-être ou bien la moyenne de Cesàro des paiements, ou leur somme escomptée :

$$\sum_{n=0}^{\infty} \delta^n u_{s_n}^i(\boldsymbol{a}_n)$$

où $\delta \in (0,1)$ est le taux d'escompte. Dans cette thèse, on considèrera principalement le paiement escompté.

Modèle 1 Séquence de jeu dans un jeu stochastique

un état initial s_1 est choisi
for $n = 1, \ldots$ **do**
  chaque joueur i choisit une action $a_n^i \in A^i$
  un nouvel état s_{n+1} est tiré selon $P_{s_n}(\boldsymbol{a}_n)$
  chaque joueur i obtient $u_{s_n}^i(\boldsymbol{a}_n)$
end for

Nous nous intéressons au cas où tous les joueurs choisissent leurs actions en suivant un algorithme (potentiellement aléatoire), basé sur des observations antérieures. Celles-ci incluent, pour chaque joueur, au moins son propre gain et potentiellement d'autres éléments tels que les actions jouées par tous les joueurs. La définition des algorithmes dépend des informations dont les joueurs disposent, ce qui est donc un élément crucial du modèle. Avec les algorithmes des joueurs, le modèle 1 est un système dynamique, et nous allons étudier son comportement.

Exemples Considérons un réseau de producteurs d'énergie qui doivent faire face à la demande et ajuster leur production en conséquence. On suppose que ces agents sont indépendants, en particulier en ce qui concerne leur contrôle. Cependant, tous ces agents veulent éviter à tout prix une production inférieure à la demande (ce qui se traduirait par une pénalité dissuasive). Il s'agit d'un système multi-agents dans lequel les producteurs doivent coopérer et trouver une combinaison des niveaux de production qui maximise le profit total en tenant compte de l'état actuel (par exemple, on suppose que modifier la température et le niveau de production est coûteux). Lorsque le profit total est partagé à parts égales entre les producteurs, alors il s'agit d'un jeu stochastique à intérêt identique (c'est-à-dire que tous les joueurs ont la même fonction d'utilité).

En économie, le duopole de Bertrand est un modèle où deux entreprises sont en concurrence sur un marché avec éventuellement des chocs de demande. Ici le système ne présente pas d'état intrinsèque, mais les entreprises peuvent prendre des décisions basées sur l'histoire du jeu, par exemple en ajustant leur prix en fonction de la concurrence.

Ces deux exemples sont détaillés et formellement définis dans la section 3.4 du chapitre 3.

Un autre exemple étudié dans cette thèse (dans le chapitre 8) est celui des systèmes avec des files d'attente (ou goulots d'étranglement) : un certain nombre d'agents doivent envoyer des paquets (ou *jobs*) dans une file d'attente unique. Le temps est divisé en une suite de périodes de durées égales. Au début de chaque période, chaque joueur reçoit un paquet et doit choisir à quelle étape de la période il l'envoie : les paquets sont ensuite traités par la file d'attente dans leur ordre d'arrivée. Il s'agit d'une décision stratégique, car les joueurs veulent minimiser le temps d'attente du paquet (c'est-à-dire qu'ils veulent le garder le plus longtemps possible). Cependant, lorsque tous les paquets arrivent après une date limite (à la fin de la période), ils subissent une pénalité dissuasive. Lorsqu'on suppose que les paquets en retard restent dans le système à la période suivante, notre système est modélisé par des interactions répétées entre agents dans un environnement changeant, c'est-à-dire par un jeu stochastique.

Bien que cela sorte du cadre de cette thèse, on peut noter qu'en plus des systèmes multi-agents en informatique, les concepts de théorie des jeux peuvent être très bien appliqués en biologie évolutive, à la fois théoriquement (Weibull, 2004) et empiriquement avec, par exemple, des études sur les colonies d'araignées (Hammerstein et Riechert, 1988).

Concepts de solution La définition d'un jeu spécifie l'utilité (ou paiement) qu'un joueur obtient du fait de la combinaison des actions choisies par tous les joueurs. En informatique, l'utilité est généralement mesurée comme l'opposé d'un coût et cette fonction de coût peut par exemple représenter la précision d'une prédiction. En économie et en sciences politiques, la définition des fonctions d'utilité n'est pas triviale. De telles fonctions n'existent pas nécessairement et tenter d'en définir une peut conduire à des contradictions (Morrow, 1994). Cependant, sous l'hypothèse de préférences ordinales, il est possible de définir des fonctions d'utilité.

Une fois les fonctions d'utilité définies, ainsi que les fonctions de transition pour les jeux stochastiques, rien n'est encore dit sur le comportement du système, c'est-à-dire sur ce que font réellement les joueurs. C'est pourquoi les théoriciens des jeux ont défini ce que l'on appelle des *concepts de solution* ou *types*

de solution qui définissent le comportement du système dans le cadre d'un ensemble d'hypothèses, par exemple en supposant que les joueurs sont rationnels.

Les concepts de solution traditionnels sont les équilibres. Il s'agit des profils stratégiques stables, c'est-à-dire qu'aucun joueur n'est incité à dévier unilatéralement. Un concept de solution bien connu est l'équilibre de Nash (Nash, 1950). Ce concept a été à l'origine de nombreux travaux car il soulève néanmoins certaines questions : comment ces équilibres peuvent-ils être calculés ? Dans quels cas y a-t-il unicité ? Les joueurs coordonnent-ils réellement leurs actions pour jouer un équilibre ?

De nombreux théoriciens des jeux ont proposé d'autres concepts de solution (par exemple les équilibres corrélés de Aumann (1974)) ou des théories sous-jacentes telles que la théorie de la sélection des équilibres de Harsanyi et Selten, 1988.

On peut aussi considérer que les procédures d'apprentissage sont des concepts de solution en tant que tels : des hypothèses sont faites sur le comportement des agents (par exemple, ils utilisent un algorithme précis) et les systèmes peuvent alors être étudiés de manière dynamique. Une question intéressante est celle du lien entre l'apprentissage en tant que concept de solution et les concepts de solution traditionnels : les procédures d'apprentissage conduisent-elles à des équilibres ? C'est l'approche constructive poursuivie dans cette thèse.

Procédures d'apprentissage En informatique et en théorie des jeux, de nombreuses procédures d'apprentissage ont été proposées. En particulier, l'apprentissage par renforcement multi-agents (Busoniu et al., 2008) est le sous-domaine de l'informatique qui traite des procédures permettant de jouer dans des environnements avec plusieurs joueurs. De manière relativement contre-intuitive, ceci peut être considéré comme un cas particulier d'apprentissage par renforcement (voir Sutton et Barto (2018) pour une référence contemporaine) où un agent unique est confronté à un environnement changeant. En effet, dans le cas des systèmes multi-agents, du point de vue d'un joueur, l'environnement est composé des autres joueurs dont le comportement peut évoluer.

L'apprentissage par renforcement est un vaste domaine de l'informatique, avec des algorithmes bien connus tels que Q-learning (Watkins, 1989). Q-learning est défini à l'aide d'un tableau de Q-valeurs qui représentent la valeur attendue d'une paire d'actions d'état, c'est-à-dire le gain futur total $Q(s, a^i)$ que le joueur i peut espérer s'il joue l'action a^i dans l'état s. Les Q-valeurs sont mises à jour en suivant l'équation :

$$Q_{n+1}(s_n, a_n^i) = Q_n(s_n, a_n^i) + \alpha\delta(r_n^i + \delta \max_{b^i \in A^i} Q_n(s_{n+1}, b^i)) \tag{1.1}$$

où α est la vitesse d'apprentissage et les autres Q-valeurs sont inchangées. En utilisant ces Q-valeurs, le joueur peut choisir une action avec une Q-valeur maximale ou explorer d'autres actions, suivant le dilemme traditionnel exploration-exploitation.

En théorie des jeux, les procédures d'apprentissage sont presque aussi anciennes que la théorie des jeux elle-même (Brown, 1951 ; Robinson, 1951) et la littérature ne cesse de croître (Fudenberg et Levine, 1998 ; Cesa-Bianchi et Lugosi, 2006 ; Hart et Mas-Colell, 2013). L'une des procédures les plus anciennes est *Fictitious Play (FP)* (Brown, 1951 ; Robinson, 1951), dans laquelle on suppose que les joueurs jouent une réponse optimale aux actions moyennes empiriques passées des autres joueurs.

Comme l'explique Shoham et al. (2007), les objectifs de l'apprentissage par renforcement multi-agents sont multiples et parfois entrecroisés. Trois des objectifs décrits par Shoham et al. (2007) présentent un intérêt particulier dans le contexte de cette thèse. Tout d'abord, l'objectif computationnel consiste en la mise au point de procédures qui peuvent être utilisées pour calculer les équilibres. Ensuite, l'apprentissage étant lui-même un concept de solution, on peut chercher à décrire les systèmes dont les agents suivent des procédures. Enfin, il y a une dimension normative lorsqu'on étudie la manière dont les acteurs apprennent à atteindre un objectif spécifique, qui peut par exemple être la robustesse face aux attaques adverses.

Cette thèse est principalement consacrée au deuxième programme : nous étudions le comportement à long terme des systèmes dont le comportement des acteurs est défini par des procédures relativement simples.

1.2 Cette thèse

Cette thèse a trois directions principales : une étude expérimentale de Q-learning dans les jeux répétés et stochastiques ; la définition et la preuve de convergence de plusieurs procédures inspirées de Fictitious Play et Q-learning dans les jeux stochastiques ; et l'étude d'un algorithme sans regret (*no-regret*) dans un jeu avec des files d'attente.

La suite de ce chapitre est un résumé du reste de la thèse rédigé en anglais.

Définition et état de l'art La section 1.4 est un résumé des chapitres 3, et elle présente les définitions, notations et exemples utilisés dans cette thèse. Une liste de symboles au début de ce document ainsi qu'un glossaire à la fin complète les définitions. La section 1.5 reprend les points essentiels du chapitre 4, qui présente un état de l'art sur l'apprentissage dans les jeux répétés et stochastiques, un domaine en pleine expansion avec des liens avec l'optimisation en ligne.

Q-learning dans les jeux répétés et stochastiques La section 1.6 du résumé et le chapitre 5 montrent, à l'aide de plusieurs expériences, que même avec des règles d'apprentissage simples, le comportement des systèmes multi-agents est complexe. Plus précisément, nous étudions l'effet de la mémoire, qui peut avoir des impacts négatifs en fonction du jeu étudié, ainsi que la robustesse de résultats expérimentaux de collusion par rapport aux schémas d'exploration.

Fictitious Play et Smooth Fictitious Play La section 1.7 présente les chapitres 6 et 7, qui décrivent deux familles de procédures d'apprentissage dans les jeux stochastiques. Les procédures du chapitre 6 sont construites en utilisant des idées de Fictitious Play (FP) et de Q-learning. Le chapitre 7 étend ces procédures à la sélection d'actions lisse (*smooth*) lorsque le jeu est inconnu.

Files d'attente Le dernier chapitre (résumé dans la section 1.8) est consacré à l'étude d'un jeu particulier avec une file d'attente. Les joueurs doivent envoyer des paquets, appelés *jobs*, à travers une file d'attente.

Plus précisément, le temps est divisé en une suite de périodes de durée égale. Au début de chaque période, chaque joueur reçoit un paquet et doit choisir à quel moment de la période il l'envoie. L'utilité de chaque joueur est alors proportionnelle au temps pendant lequel il conserve le paquet : plus il l'envoie tard, meilleure est son utilité.

Cependant, les paquets sortent de la file d'attente dans l'ordre « premier arrivé, premier servi » et lorsqu'ils arrivent après une date limite fixe (c'est-à-dire après la fin de la période), les joueurs reçoivent une pénalité dissuasive. Par conséquent, l'étape au cours de laquelle un joueur choisit d'envoyer le paquet est stratégique.

Nous étudions ce jeu dans un cadre répété où les paquets qui n'ont pas été traités au cours d'une période sont encore présents au cours de la période suivante. Nous supposons que les joueurs utilisent un algorithme *no-regret* et sont intéressés par le comportement du système, en particulier par le fait de savoir si le nombre de travaux reste limité ou non. Sous des hypothèses appropriées sur le paramètre de pénalité, nous prouvons que le nombre de paquets est presque sûrement borné.

1.3 Publications

Plusieurs articles ont été rédigés au cours de ma thèse. Les deux premiers ont été rédigés avec l'un de mes directeurs de thèse, Rida Laraki, et le troisième a été écrit avec Marco Scarsini et Xavier Venel qui m'ont invité au LUISS (Rome) pendant deux mois pour étudier les systèmes de files d'attente stratégiques.

- Lucas Baudin et Rida Laraki. "Fictitious Play and Best-Response Dynamics in Identical-Interest and Zero-Sum Stochastic Games". In : *Proceedings of the 39th International Conference on Machine Learning*. T. 162. Proceedings of Machine Learning Research. PMLR, 17-23 juill. 2022, p. 1664-1690

  Dans cet article, nous étendons FP aux jeux stochastiques escomptés. Nous prouvons que les procédures qui en résultent convergent vers l'ensemble des équilibres de Nash stationnaires dans le cas des jeux stochastiques à intérêts identiques, et également dans le cas des jeux à somme nulle.

Ces résultats étendent les résultats de convergence similaires pour les jeux non stochastiques. Afin de prouver les résultats dans les jeux stochastiques en équipe et à somme nulle, nous analysons les dynamiques en temps continu correspondant à ces procédures en temps discret. Elles incluent comme cas particulier la dynamique de meilleure réponse introduite et étudiée par Leslie et al. (2020) dans le contexte des jeux stochastiques à somme nulle. Nous prouvons la convergence de cette dynamique vers des équilibres de Nash stationnaires dans des jeux stochastiques à intérêts identiques et dans les jeux stochastiques à somme nulle. Ensuite, l'approximation stochastique est utilisée afin d'obtenir des résultats de convergence en temps discret.

- Lucas Baudin et Rida Laraki. "Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions". In : *Advances in Neural Information Processing Systems*. 2022

  Cet article est la suite de l'article précédent dans lequel les joueurs estiment un modèle du jeu stochastique pendant la séquence de jeu dans le cas où le jeu n'est pas connu a priori. Pour ce faire, nous utilisons des fonctions de régularisation à l'intérieur des algorithmes du premier article. Nos nouvelles procédures peuvent être interprétées comme des extensions aux jeux stochastiques des procédures classiques d'apprentissage *smooth fictitious play* dans les jeux statiques (où les meilleures réponses des joueurs sont régularisées, grâce à une perturbation lisse de leurs fonctions de gain). La convergence vers des équilibres de Nash régularisés stationnaires est prouvée pour les mêmes classes de jeux dynamiques (jeux stochastiques à somme nulle et à intérêts identiques escomptés).

  Dans le cas d'un MDP (jeu stochastique à un joueur), nos procédures convergent globalement vers la politique stationnaire optimale du problème régularisé.

- Un article qui n'est pas encore publié : Lucas Baudin et al. *Strategic Behavior and No-Regret Learning in Queueing Systems*. Preprint, 2023. arXiv : `2302.03614 [cs.GT]`

  Cet article a été rédigé avec Marco Scarsini et Xavier Venel. Il a commencé comme une version dynamique de Rivera et al., 2018b.

  Nous étudions un modèle de file d'attente dynamique à temps discret où, à chaque période, les joueurs obtiennent un nouveau paquet et doivent envoyer tous leurs paquets dans une file d'attente. Les joueurs sont incités à envoyer leurs paquets le plus tard possible. Cependant, si un paquet ne sort pas de la file d'attente avant une date limite fixée, le propriétaire du paquet subit une pénalité et ce paquet est renvoyé au joueur et rejoint la file d'attente à la période suivante. Par conséquent, la stabilité, c'est-à-dire le fait que le nombre de paquets dans le système soit borné, n'est pas garantie. Nous montrons que si les joueurs sont stratégiques, le système est stable lorsque la pénalité est suffisamment élevée. De plus, si les joueurs utilisent un algorithme d'apprentissage dérivé d'un algorithme *no-regret* classique (*exponential weight*), alors le système est stable lorsque les pénalités sont supérieures à une borne qui dépend du nombre total de paquets dans le système.

1.4 Définitions et état de l'art en théorie des jeux

Depuis la publication de *Theory of Games and Economic Behavior* par John von Neumann et Oskar Morgenstern en 1944, la théorie des jeux s'est imposée dans les sciences économiques, et en particulier dans leur enseignement, en plus d'être un vaste champ des mathématiques appliquées. Plus récemment, beaucoup de recherche en informatique a été consacrée aux systèmes multi-agents qui, lorsqu'ils sont contrôlés par des algorithmes, peuvent correspondre à des jeux stochastiques. À la différence des agents économiques, les algorithmes obéissent à des règles précises et la théorie des jeux est d'autant plus adaptée pour formaliser ces problèmes.

1.4.1 Jeux sous forme normale

D'abord, nous définissons les jeux sous forme normale (ou sous forme stratégique (Fudenberg et Tirole, 1991)) de manière non-coopérative.

Définition 1.1 (Jeu)**.** Un jeu est un n-uplet $(I, \boldsymbol{A}, (u^i)_{i\in I})$ où :

- I est l'ensemble fini des joueurs

- $\boldsymbol{A}$ est l'ensemble des profils d'action, c'est-à-dire le produit des ensembles d'actions de chaque joueur $\Pi_{i\in I} A^i$
- $u^i : \boldsymbol{A} \to \mathbb{R}$ est la fonction d'utilité (ou de paiement) du joueur i.

Une action mixte est une distribution de probabilité sur les actions (qui sont appelées *pures*), c'est-à-dire les éléments du simplex $\Delta(A^i)$ pour un joueur i. On étend les fonctions d'utilité par n-linéarité aux profils d'actions mixtes, c'est-à-dire au domaine $\prod_{i\in I} \Delta(A^i)$.

On note en gras les profils d'action (les n-uplets avec une action par joueur), par exemple $\boldsymbol{a}, \boldsymbol{b}$ et on note a^i, b^i les actions d'un joueur i. On utilise $\boldsymbol{y}$ pour les profils mixtes et $(b^i, \boldsymbol{a}^{-i})$ désigne le profil d'action où i joue b^i et les autres joueurs jouent les actions de $\boldsymbol{a}$.

On considère deux classes de jeux en particulier, les jeux à deux joueurs et à somme nulle – où les joueurs sont dans une compétition complète – et les jeux à intérêts identiques – où les joueurs ont la même fonction d'utilité.

Définition 1.2 (Jeu à somme nulle)**.** Un jeu à somme nulle est un jeu à deux joueurs tel que $u^1 = -u^2$.

Définition 1.3 (Jeu à intérêts identiques)**.** Un jeu présente des intérêts identiques lorsque tous les joueurs partagent la même fonction d'utilité, c'est-à-dire qu'il existe $u : \boldsymbol{A} \to \mathbb{R}$ tel que $\forall i \in I, u^i = u$.

Remarque. *Dans le cas des jeux à intérêts identiques, on ne suppose pas que les fonctions d'utilité sont symétriques. Ainsi, ce n'est pas parce qu'ils ont les mêmes fonctions d'utilité que les joueurs sont interchangeables dans le jeu.*

1.4.2 Jeux répétés

On se place dans le cadre des jeux répétés, ce qui permet aux joueurs d'utiliser l'histoire, c'est-à-dire les profils d'actions passés, pour choisir leurs actions. Ainsi, ils peuvent conditionner leurs actions au comportement des autres joueurs et à l'effet de leurs propres actions. Cela peut modéliser une situation d'apprentissage. Plus tard, on généralise cette notion à des répétitions de jeux avec une variable d'état (des jeux stochastiques).

Définition 1.4 (Jeu répété)**.** Pour un jeu $G = (I, \boldsymbol{A}, (u^i)_{i\in I})$, on définit un jeu répété (ou itéré), indexé par $n \in \mathbb{N}$, qui se déroule comme suit : à chaque étape n, chaque joueur i choisit une action (de manière potentiellement randomisée) $a^i_n \in A^i$, qui est observée par les autres joueurs. On appelle stratégie un élément de l'ensemble :

$$\bigcup_{n=0}^{\infty} \boldsymbol{A}^n \to \Delta(A^i)$$

où $\boldsymbol{A}^n$ est la liste des profils d'actions jusqu'à l'étape n. Pour un taux d'escompte $\delta \in [0, 1)$, on considère le paiement escompté du joueur i :

$$\sum_{n=0}^{\infty} \delta^n u^i(\boldsymbol{a}_n)$$

1.4.3 Jeux stochastiques

Définis initialement par Shapley (1953), les jeux stochastiques sont des jeux avec une variable d'état. Le jeu se déroule sur une suite d'étape (comme pour les jeux répétés) et la variable d'état évolue suivant une fonction des actions des joueurs et de l'état actuel. Les fonctions d'utilité sont paramétrées par l'état.

Définition 1.5 (Jeu stochastique)**.** Un jeu stochastique $\mathcal{G}$ est un n-uplet $(I, A^i, S, (u^i_s)_{i\in I, s\in S}, (P_s)_{s\in S})$ où :

- S est l'ensemble des états
- $u^i_s : \boldsymbol{A} \to \mathbb{R}$ est la fonction d'utilité du joueur i dépendant de l'état
- $P_s : \boldsymbol{A} \to \Delta(S)$ est une fonction de transition entre les états

Dans toute cette thèse, on s'intéresse à l'utilité escomptée qu'obtient chaque joueur :

$$U^i_\delta = \sum_{n=0}^{\infty} \delta^n u^i_{s_n}(\boldsymbol{a}_n)$$

Jeu répété Ainsi, un jeu répété est un jeu stochastique avec un seul état.

Ergodicité On dit qu'un jeu stochastique est ergodique lorsque tous les états sont atteints avec une probabilité strictement positive après un horizon fixé, quel que soit l'état initial et les actions prises par les joueurs, voir définition 3.7 du chapitre 3.

Histoires et procédures Une stratégie du jeu stochastique (ou du jeu répété) est une fonction de l'histoire vers les actions mixtes, ce qui en fait formellement, pour un joueur i, des fonctions de la forme :

$$\bigcup_{n\geq 0}(S\times \boldsymbol{A})^n\times S\to\Delta(A^i)$$

Le domaine de ces fonctions est l'ensemble des histoires du jeu, noté $\mathcal{H}$.

Dans le cas d'un jeu répété, on définit pour un $N\in\mathbb{N}$ le jeu à mémoire bornée de longueur N comme le jeu stochastique avec pour ensemble d'états $\cup_{0\leq n\leq N}\boldsymbol{A}^n$ et les mêmes fonctions d'utilité que le jeu initial. À chaque étape du jeu, le nouvel état est obtenu en oubliant le premier élément du n-uplet et en ajoutant le profil d'action qui vient d'être joué.

Lorsque les stratégies sont calculables (au sens de la calculabilité en informatique) et sont conçues pour maximiser l'utilité du joueur qui les emploie, on les appelle aussi procédures d'apprentissage.

Équilibres La définition des jeux (stochastiques ou non) ne permet pas de déduire directement le comportement des systèmes modélisés. On y ajoute les « concepts de solution » qui sont des règles normatives décrivant quelles actions chaque joueur choisit. En particulier, les *équilibres* forment un ensemble de concepts de solutions définissant ce qu'est un profil d'actions stable, à commencer par le plus connu qui est celui de Nash (1950).

Définition 1.6 (Équilibre de Nash dans un jeu statique)**.** Un équilibre de Nash est un profil d'actions (potentiellement mixte) $\boldsymbol{y}$ tel qu'aucune déviation unilatérale n'est profitable :

$$\forall i\in I,\forall b^i\in A^i, u^i(\boldsymbol{y})\geq u^i\big(b^i,\boldsymbol{y}^{-i}\big) \tag{1.2}$$

Cette définition s'étend aux jeux stochastiques (voir définition 3.10), qui admettent nécessairement au moins un équilibre de Nash (Fink, 1964).

Aumann (1974) a défini un autre type d'équilibre, les équilibres corrélés, dans lesquels les joueurs choisissent une action aléatoirement, mais disposent d'un moyen de corréler leurs choix. Ainsi, ces équilibres sont des distributions de probabilités sur l'ensemble $\boldsymbol{A}$ qui sont stables au sens où, pour un joueur, remplacer une de ses actions par une autre ne lui permet pas d'augmenter son utilité espérée (voir définition 3.11 dans le chapitre 3). L'ensemble de ces équilibres contient l'ensemble des équilibres de Nash (un profil mixte est une distribution de probabilité sur $\boldsymbol{A}$ qui est le produit des distributions sur les actions de chaque joueur).

Un *coarse correlated equilibrium* est aussi une distribution de probabilités sur $\boldsymbol{A}$, mais il est stable pour les déviations où un joueur remplace *toutes* ses actions par une autre action (définition 3.12). Cette définition est plus faible que celle des équilibres corrélés et a fortiori des équilibres de Nash.

Exemples Le dilemme du prisonnier est un jeu avec 2 joueurs (colonne et ligne), sans état. Chaque joueur dispose de deux actions C (coopérer) et D (faire défection), avec la matrice d'utilité suivante :

	C	D
C	$(3,3)$	$(0,4)$
D	$(4,0)$	$(1,1)$

Il y a un unique équilibre de Nash (D,D). Dans le contexte des jeux répétés à horizon infini ou mémoire bornée, on dit que les joueurs coopèrent lorsqu'ils jouent (C,C) alors même qu'ils ont tous les deux un intérêt à dévier.

Une généralisation de ce jeu est la compétition de Bertrand, qui représente un duopole où deux producteurs fixent le prix auquel ils vendent un bien, détaillé formellement dans le jeu 2 du chapitre 3.

Enfin, nous considérons aussi un jeu à intérêts identiques où un ensemble de producteurs d'énergie doivent faire face à une demande ; chacun choisit un niveau de production. Ils partagent le profit de manière égale (ils ont la même fonction d'utilité), mais ont une pénalité dissuasive s'ils produisent moins que la demande. Le jeu est stochastique, car la température évolue et le coût de l'énergie dépend des niveaux de production précédent (il est coûteux de changer le niveau de production). Cela conduit au jeu 1 du chapitre 3 (page 44).

1.5 Procédures d'apprentissage

Dans cette section, on décrit brièvement la littérature sur l'apprentissage dont s'inspire cette thèse. La version longue en anglais est le chapitre 4.

L'apprentissage dit « en ligne » se déroule sur plusieurs étapes (éventuellement une infinité), au cours desquelles l'agent doit choisir une action et obtient une récompense qui a une certaine utilité. L'agent acquiert des informations avec le temps, par exemple sur l'état de l'environnement ou sur ce qu'ont rapporté ses actions.

Ce cadre très général s'adapte à une multitude de problèmes et pour cette thèse, on s'intéresse à deux cas particuliers qui ont été très étudiés : les jeux répétés et l'apprentissage par renforcement dans les systèmes multi-agents. Le premier a été très étudié par les théoriciens des jeux, alors que le second est investi par les informaticiens.

Dans les jeux répétés

L'apprentissage dans les jeux répétés a historiquement été étudié pour plusieurs raisons : d'abord pour calculer des équilibres (c'est le cas dans les travaux initiaux de Brown et Robinson au sujet de Fictitious Play), ensuite comme un concept de solution en tant que tel (Fudenberg et Levine, 1998) et enfin en informatique pour concevoir des algorithmes qui contrôlent des systèmes.

Formellement, l'apprentissage se déroule comme dans le modèle 2.

Modèle 2 Apprentissage dans un jeu répété

for $n = 1, \dots$ **do**
  for all $i \in I$ **do**
    i choisit une action $a_n^i \in A^i$
  end for
  for all $i \in I$ **do**
    i observe $\boldsymbol{a}_n^{-i}$
    i obtient $u^i(\boldsymbol{a}_n)$
  end for
end for

Fictitious Play est l'une des procédures les plus anciennes. Initialement définie pour les jeux à somme nulle par Brown (1951) et Robinson (1951), la procédure et sa convergence vers l'ensemble des équilibres de Nash a ensuite été étendue dans plusieurs classes de jeux comme les jeux de potentiels (Monderer et Shapley, 1996a) ou les jeux $2 \times n$ (Berger, 2005). Elle ne converge pas vers l'ensemble des équilibres dans tous les jeux, comme l'a montré Shapley (1964).

Avec cette procédure, les joueurs choisissent à chaque étape une action qui est une meilleure-réponse à la moyenne empirique des actions passées. Formellement, on définit cette moyenne empirique du joueur j de la manière suivante :

$$x_n^j := \frac{1}{n} \sum_{k=1}^{n} a_k^j.$$

À chaque étape n, un joueur i utilisant Fictitious Play choisit une action qui est une meilleure réponse (BR) :

$$a_n^i \in \mathrm{BR}^i\left(\boldsymbol{x}_n^{-i}\right) := \arg\max_{b^i \in A^i} u^i\left(b^i, \boldsymbol{x}_n^{-i}\right). \tag{FP}$$

Lorsqu'il y a plusieurs meilleures réponses, l'algorithme a en pratique besoin d'une règle de départage. Celle-ci n'est habituellement pas spécifiée lorsque FP est défini et les résultats de convergence sont prouvés indépendamment de cette règle.

On peut remplacer cette meilleure réponse par un choix plus lisse d'action (*smooth*). Cela permet d'une part d'explorer toutes les actions et d'estimer la matrice de paiement si elle n'est pas connue, et d'autre part de ne pas avoir de regret (et en conséquence d'être robuste à une famille d'attaque adversariale), comme défini ci-dessous. Formellement, pour une fonction h^i qui est $\mathcal{C}^1$ et *steep* (la dérivée diverge sur les bords du simplex) et un paramètre $\eta > 0$, la smoooth best-response est définie par :

$$\mathrm{SBR}^i\big(\boldsymbol{x}^{-i}\big) := \operatorname*{arg\,max}_{y^i \in A^i} u^i\big(y^i, \boldsymbol{x}^{-i}\big) + \eta h^i\big(y^i, \boldsymbol{x}^{-i}\big). \tag{1.3}$$

Regret et adversaires Une mesure intéressante de la fiabilité des procédures d'apprentissage est le *regret* (Hannan, 1957), qui est la différence entre les récompenses obtenues et la meilleure récompense possible si un joueur avait utilisé une action constante. Formellement, on définit le regret (externe) :

$$r_n^i = \max_{b^i \in A^i} \left(\sum_{k=0}^{n} u^i\big(b^i, \boldsymbol{a}_k^{-i}\big) - u^i(\boldsymbol{a}_k) \right) \tag{1.4}$$

On dit qu'une procédure est sans regret (*no-regret*) lorsque r_n^i est en $o(n)$. On peut montrer que sur certains jeux, Fictitious Play a du regret, ce qui n'est pas le cas de Smooth Fictitious Play avec une valeur de η judicieusement choisie. Benaïm et Faure (2013) ont défini une procédure inspiré de Smooth Fictitious Play mais avec un paramètre η changeant et tendant vers 0, qui est sans regret.

Des procédures d'apprentissage utilisent le regret pour calculer la prochaine action, comme *regret-matching* (Hart et Mas-Colell, 2000), qui est sans regret.

Intuitivement, le regret permet d'évaluer la robustesse de la procédure face à un adversaire : quoi que fassent les autres joueurs, les choix d'action faits par un joueur qui utilise une procédure sans regret sont au moins aussi bon qu'une action constante, autrement dit un joueur peut garantir qu'il obtient au moins ce que rapporte sa meilleure action (constante).

Le regret peut aussi être défini dans un jeu à un seul joueur (un processus de décision Markovien, éventuellement non stationnaire) et est un cas particulier de l'approchabilité (Blackwell, 1956a).

Apprentissage par renforcement

L'apprentissage par renforcement est un cadre très général, dans lequel un agent doit choisir une action depuis un ensemble A pendant un certain nombre (ou une infinité) d'étapes indexées par n. Le système est à chaque instant dans un état s_n (souvent observé par l'agent) et l'agent obtient une récompense à chaque étape. Il y a un « renforcement » car lorsqu'une action est jouée, la probabilité de la jouer dans le futur est d'autant plus renforcée que la récompense est élevée.

Modèle 3 Apprentissage par renforcement

initialisation dans l'état $s_1 \in S$
for $n = 1, \ldots$ **do**
  l'agent choisit une action $a_n \in A$
  il obtient $U_n \in \mathbb{R}$
  il observe $s_{n+1} \in S$
end for

Lorsque l'environnement est stationnaire, la variable aléatoire U_n et l'état suivant s_{n+1} suivent des lois qui dépendent de l'état s_n et de l'action a_n de l'agent, le système est un processus de décision markovien.

Q-learning Dans ce contexte, une procédure simple a été proposé par Watkins (1989) et permet de résoudre (c'est-à-dire trouver les actions optimales) les processus de décision markoviens. L'idée

principale est d'estimer la continuation obtenue en partant d'un état et d'une action, avec des Q-valeurs. À chaque étape, la Q-valeur correspondant à l'état actuel et l'action jouée est mise à jour :

$$Q(s_n, a_n) \leftarrow Q(s_n, a_n) + \alpha_n(U_n + \delta \max_{b \in A} Q(s_{n+1}, b) - Q(s_{n+1}, a_n)) \tag{1.5}$$

Les autres Q-valeurs ne sont pas modifiées. L'équation 1.5 n'a pas besoin de fonctions de transition entre les états, ce qui fait de Q-learning un algorithme simple.

À chaque étape, l'action est sélectionnée en tenant compte des Q-valeurs et en arbitrant entre exploration (tester les actions les moins expérimentées) et exploitation (utiliser une action qui maximise les Q-valeurs déjà calculées).

Plusieurs extensions de Q-learning ont ensuite été proposées, par exemple Double Q-learning (Hasselt, 2010) ou Q-learning avec du deep learning (Hasselt et al., 2016).

Watkins a prouvé que dans un processus de décision markovien, Q-learning converge vers les continuations de la solution optimale avec un certain nombre d'hypothèses, notamment les hypothèses habituelles suivantes :

$$\forall n, 0 < \alpha_n \leq 1$$
$$\sum_k^n \alpha_k = \infty$$
$$\sum_k^n \alpha_k^2 < \infty$$

Cependant, dans les systèmes multi-agents, du point de vue d'un joueur l'environnement n'est pas stationnaire, car les autres joueurs peuvent avoir un comportement variable (par exemple s'ils utilisent aussi Q-learning). Ceci a donné lieu à l'émergence du champ de l'apprentissage par renforcement multi-agents.

Apprentissage par renforcement dans les systèmes multi-agents

Les procédures d'apprentissage spécifiques aux systèmes multi-agents ont été très étudiées par les informaticiens, en général dans un cas où le modèle est inconnu (c'est-à-dire que les lois des récompenses et des fonctions de transitions ne sont pas connues a priori).

Lorsque le modèle est connu, on peut écrire le système sous la forme d'un jeu stochastique, ce qui donne une séquence de jeu décrite dans le modèle 4.

Modèle 4 Apprentissage dans un jeu stochastique

un état initial s_1 est choisi
for $n = 1, \ldots$ **do**
  for all $i \in I$ **do**
    i choisit une action $a_n^i \in A^i$
  end for
  s_{n+1} est tiré aléatoirement suivant $P_{s_n}(\boldsymbol{a}_n)$
  for all $i \in I$ **do**
    i observe s_{n+1} et $\boldsymbol{a}_n^{-i}$
    i obtient $u_{s_n}^i(\boldsymbol{a}_n)$
  end for
end for

Plusieurs procédures s'inspirent de Q-learning dans un contexte de jeu. Nash Q-learning (Hu et Wellman, 2003) est une procédure dans laquelle les continuations sont apprises comme dans Q-learning mais on suppose en plus que les joueurs ont accès à un oracle pour déterminer un équilibre de Nash et le jouer. Dans les jeux répétés, Leslie et Collins (2005) ont montré que sous certaines hypothèses sur le taux de mise à jour α_n, les systèmes multi-agents avec Q-learning peuvent être reliés à la dynamique de meilleure réponse lisse (smooth). Plus récemment, Sayin et al. (2022a) ont proposé une version

décentralisée de Q-learning avec deux échelles de temps pour apprendre à la fois la valeur d'un état et les Q-valeurs classiques.

D'autres algorithmes, comme WoLF (Bowling et Veloso, 2001) ou AWESOME (Conitzer et Sandholm, 2007) permettent d'apprendre dans des systèmes multi-agents, avec d'autres fonctionnements et en particulier des taux d'apprentissage qui dépendent du gain réalisé.

Enfin, Fictitious Play a été étendu à des jeux stochastiques. D'abord par Vrieze et Tijs (1982) dans le but de calculer la valeur du jeu. Plus récemment, Leslie et al. (2020) ont étudié la dynamique de meilleure réponse dans des jeux stochastiques avec deux échelles de temps (ce dont on s'inspire pour la section 1.7) et Sayin et al. (2022a) et Sayin et al. (2022b) ont proposé un algorithme en temps discret semblable à Fictitious Play mais avec là aussi deux échelles de temps. La procédure est différente de celle définie plus bas, voir la section 4.3.1 à la page 56 pour plus de détails.

1.6 Étude empirique de Q-learning avec une mémoire bornée

Avec la place croissante que prennent les algorithmes dans les prises de décisions, la question du comportement des systèmes multi-agents contrôlés par des algorithmes est devenue cruciale. C'est par exemple le cas pour les algorithmes qui fixent les prix, comme dans les stations services (Asker et al., 2021), ce qui interroge les pouvoirs publics sur la collusion algorithmique tacite qui pourrait se produire (OECD, 2017).

Lorsque tous les agents utilisent Q-learning, l'évolution des systèmes multi-agents n'est pas bien comprise (Wunder et al., 2010 ; Babes et al., 2009). Calvano et al. (2020) ont expérimentalement mis en évidence que dans un jeu de Bertrand, les joueurs apprennent à choisir des prix au-dessus de l'équilibre, ce qui ressemble à de la collusion (d'autant plus qu'ils montrent la présence d'un schéma de punition lorsqu'un joueur s'écarte des prix limites).

Dans le chapitre 5, résumé dans cette section, on réalise plusieurs expériences numériques sur des systèmes où tous les joueurs utilisent Q-learning avec les mêmes paramètres. On se place dans des jeux répétés, qui n'ont donc pas d'état intrinsèque mais des états représentent une mémoire bornée. On regarde d'abord l'influence de la mémoire sur la coopération dans le dilemme de prisonnier, puis celui de la méthode d'exploration de Q-learning dans le jeu de Bertrand dans un cadre similaire à celui de Calvano et al. (2020).

Coopération Dans un jeu répété, on dit que les joueurs coopèrent lorsque :

- d'une part ils jouent en moyenne un profil d'actions tel que l'utilité sociale (la somme de leurs utilités) est supérieure à celle des équilibres de Nash du jeu en une étape,
- et d'autre part que les utilités de chaque joueur sont individuellement rationnelles (c'est-à-dire supérieures à ce que les joueurs peuvent obtenir avec la meilleure action mixte indépendamment des autres joueurs).

Les profils de coopération ne sont en général pas stables au sens où il existe des déviations profitables dans le jeu statique. Ces profils peuvent cependant devenir soutenables lorsque le jeu est répété : les « Folk Theorems » (voir par exemple Fudenberg et Tirole, 1991, p. 150) établissent que certains profils d'utilités sont, lorsque le taux d'escompte δ est suffisamment élevé, soutenables avec une stratégie (non stationnaire) du jeu répété, qui est un équilibre du jeu répété. Les profils d'utilité qui sont dans ce cas sont tous ceux qui sont individuellement rationnels et dans l'enveloppe convexe des utilités des actions pures. Les stratégies de ces « Folk Theorems » sont construites autour d'un système de punition : lorsqu'un joueur i s'écarte du profil d'actions ciblé, une punition est délenchée par les autres joueurs, ce qui rend la déviation de i non-profitable. Ainsi, même si dans le jeu statique une déviation est profitable, elle ne l'est pas dans le jeu répété.

Une telle stratégie de punition peut être réalisée avec une mémoire bornée pour certaines valeurs des profils d'utilité visés. On montre à la page 60 que dans le cas du dilemme de prisonnier, une stratégie avec une mémoire de taille 2 et un taux d'escompte de $\delta = 0.65$ sont suffisants pour qu'aucune déviation ne soit profitable si la stratégie est de jouer C tant que l'autre joueur ne dévie pas, et D sinon.

Modèle On suppose que tous les joueurs utilisent Q-learning (WATKINS, 1989) dans un jeu répété. L'ensemble des états représente la mémoire bornée des joueurs (donc $S = \cup_{k=0}^{N} \boldsymbol{A}^k$ pour une mémoire N). Q-learning est formalisé de la façon suivante :

$$\begin{cases} Q^i_{n+1}(s, b^i) = \begin{cases} (1-\alpha) Q^i_n(s, b^i) + \alpha \left(u^i_s(a^i_n) + \delta \arg\max_{c^i \in A^i} Q^i_n(s_{n+1}, c^i) \right) & \text{si } s = s_n \text{ et } b^i = a^i_n \\ Q^i_n(s, b^i) & \text{sinon} \end{cases} \\ a^i_n \sim x^i_n\big(Q^i_n(s_n)\big) \end{cases}$$

$Q^i_n(s_n)$ est le vecteur $Q^i_n(s_n, \cdot)$ et $x^i_n : \mathbb{R}^{A^i} \to \Delta(A^i)$ est la fonction de sélection d'action dont l'argument est l'ensemble des Q-valeurs d'un état donné.

On va mesurer le profit en comparant l'utilité réalisée à la différence entre les utilités sociale du meilleur profil d'actions $SO(G)$ et celle du meilleur équilibre $SN(G)$ pour un jeu statique G. Cela donne la formule suivante :

$$\textsc{Prof}_n = \frac{\frac{1}{n}\left(\sum_{k=1}^{n} \sum_{i \in I} u^i(\boldsymbol{a}_k)\right) - SN(G)}{SO(G) - SN(G)}$$

Ainsi, un profit de 100% indique que les joueurs ont joué le meilleur profil, alors que 0% correspond au meilleur équilibre de Nash (et une valeur négative à un profil avec une utilité sociale encore inférieure).

Le but des expériences de ce chapitre est d'estimer $\mathbb{E}[\textsc{Prof}_n]$ pour un grand n à partir d'un état aléatoire. Pour cela, on exécute plusieurs simulations (en général 200) et on utilise la méthode du bootstrap pour évaluer la confiance qu'on a dans ces estimations, et on vérifie avec des valeurs différentes des paramètres comme expliqué dans le chapitre 5.

La mémoire dans le dilemme du prisonnier Pour ces premières expérimentations, on regarde le profit moyen qu'obtiennent les joueurs avec plusieurs taux d'escompte δ, plusieurs tailles de mémoire et deux versions du jeu : la première est le dilemme du prisonnier classique (voir page 17), et la deuxième est une version non-convexe avec la matrice d'utilité suivante (voir page 63 pour une représentation graphique) :

	C	D
C	$(3, 3)$	$(0, 10)$
D	$(10, 0)$	$(1, 1)$

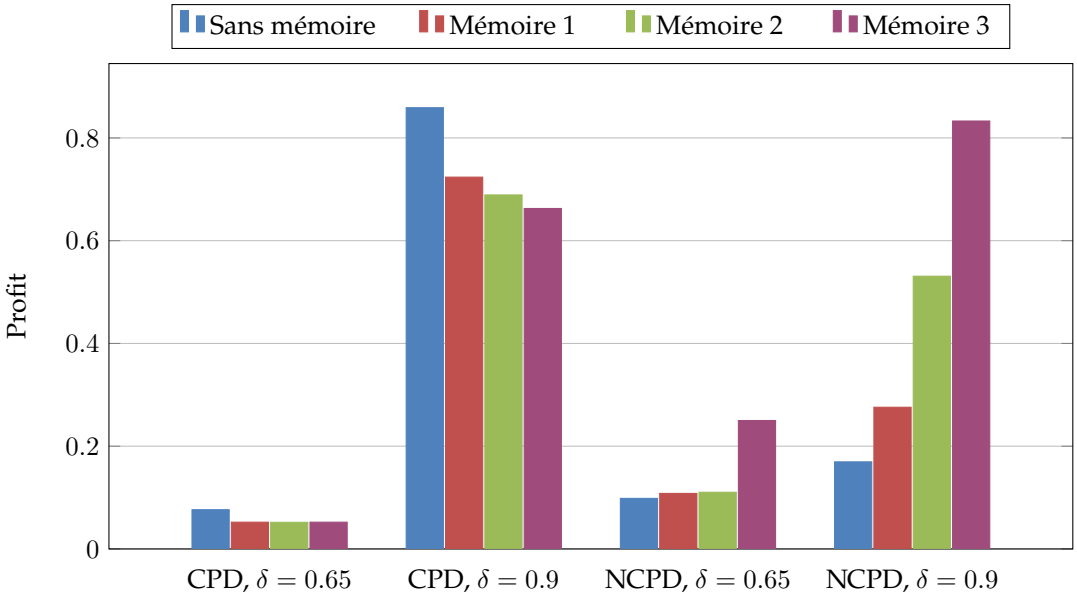
Le point important est que l'utilité sociale maximum est obtenue avec les profils (C, D) et (D, C). Par conséquent (comme les utilités doivent être individuellement rationnelles), ces utilités ne pourront être atteintes qu'avec une alternance entre (C, D) et (D, C).

Nous testons la coopération sur un horizon de $n = 2 \cdot 10^6$ étapes et on obtient les résultats de la figure 1.1. Étonnamment, le comportement est différent entre les deux versions du dilemme du prisonnier. Pour des mémoire de taille 1 ou 2, on observe peu de coopération dans la version non-convexe avec $\delta = 0.65$, ce qui peut s'expliquer par le fait que l'incitation à dévier est plus importante que dans la version convexe. On observe cependant davantage de coopération à partir d'une mémoire de taille 3. On constate une alternance entre les actions C et D et un schéma de punition dans les simulations.

Dans la version convexe (la version habituelle) du dilemme du prisonnier, un autre résultat surprenant est qu'une mémoire plus grande dégrade la coopération, et cela même si on augmente la durée des simulations proportionnellement au nombre d'états. Une interprétation de ce résultat est que la coopération émerge lorsque les joueurs arrivent à se synchroniser. Ainsi, avec des stratégies plus complexes du fait du nombre d'états plus grand, cette synchronisation entre les stratégies est plus difficile à obtenir et donc la coopération décroît.

La méthode d'exploration dans le duopole de Bertrand L'article de CALVANO et al. (2020) fait l'hypothèse que les joueurs utilisent Q-learning avec une exploration greedy dont le taux décroit exponentiellement, c'est-à-dire que :

- avec probabilité $\eta_n := \eta_0 \exp(-\beta n)$, une action est choisie uniformément aléatoirement, où $\eta_0 \in (0, 1]$ et $\beta > 0$



CPD est le dilemme du prisonnier standard et NCPD le dilemme du prisonnier non convexe.

FIGURE 1.1 : Profit moyen après $n = 2 \cdot 10^6$ étapes

- avec probabilité $1 - \eta_n$, une action qui maximise la Q-valeur est choisie :

$$a_n^i \in \arg\max_{b^i \in A^i} Q(s_n, b^i) \tag{1.6}$$

Le lemme de Borel-Cantelli implique que le nombre d'explorations est finie, ce qui paraît peu réaliste car au bout d'un certain temps, les joueurs n'explorent plus.

On réalise des expérimentations avec d'autres modes d'exploration :

- d'abord avec le mode le plus simple qui est η-greedy, où on joue une action aléatoirement avec probabilité $1 - \eta$ et uniformément aléatoire avec probabilité η (cette fois-ci η est constant).
- ensuite avec une sélection d'actions *smooth* (lisse) de la forme :

$$a_n^i \sim_{b^i \in A^i} \frac{\exp\left(\frac{1}{\eta_n} Q(s_n, b^i)\right)}{\sum_{c^i \in A^i} \exp\left(\frac{1}{\eta_n} Q(s_n, c^i)\right)} \tag{1.7}$$

On choisit $\eta_n = \beta\sqrt{n}$ comme dans Vanishingly Smooth Fictitious Play (BENAÏM et FAURE, 2013).

Dans le cas η-greedy, on obtient de la coopération mais avec des valeurs faibles du taux d'expérimentation η et du taux de mise à jour α (inférieur ou proche de 0,01, voir les figures du chapitre 5). Avec une exploration comme dans l'équation 1.7, on obtient de la coopération avec des valeurs de α dans un intervalle plus large, par exemple de 0,1 à 0,25.

Finalement, on peut en conclure que la coopération est possible avec d'autres modes d'exploration que la décroissance exponentielle. Il serait intéressant de compléter ces expériences avec d'autres modes d'exploration ou d'autres algorithmes, mais la comparaison n'est pas toujours facile car les paramètres de ces modes d'exploration ne sont pas les mêmes : si la comparaison des valeurs de α_n est pertinente entre le cas η-greedy et le cas β-exponentiel, on ne peut pas comparer η et β qui jouent des rôles différents.

Conclusion Ces deux expériences donnent un argument supplémentaire quant à la possibilité théorique de collusion algorithmique. Il serait possible de préciser ces résultats en étudiants d'autres algorithmes, d'autres modes d'exploration ou d'autres jeux, il est cependant établi qu'avec des algorithmes d'apprentissage, les systèmes constitués de jeux répétés ne convergent pas forcément vers un équilibre stationnaire lorsque des procédures classiques d'apprentissage par renforcement sont utilisées.

1.7 Fictitious Play et Smooth Fictitious Play dans les jeux stochastiques

Les chapitres 6 et 7 sont consacrés à l'extension de Fictitious Play (FP) et Smooth Fictitious Play (SFP) aux jeux stochastiques. Plusieurs procédures sont ainsi définies, et on prouve la convergence vers l'ensemble des équilibres de Nash pour les jeux stochastiques à somme nulle et ceux à intérêts identiques.

À la différence des jeux répétés, les jeux stochastiques disposent d'une variable d'état qui change pendant la séquence de jeu. Par conséquent, la fonction d'utilité peut aussi changer et donc un joueur doit prendre en compte l'état futur. Ainsi, Fictitious Play ne peut pas être utilisé directement faute de considérer l'utilité espérée au-delà de l'étape courante. On ne peut d'ailleurs pas non plus, en considérant que l'ensemble des stratégies stationnaires du jeu stochastique est l'espace d'action d'un jeu statique, utiliser FP dans une répétition du jeu stochastique : la fonction d'utilité du jeu statique en résultant n'est pas nécessairement concave (autrement dit, le paiement escompté n'est pas concave dans les stratégies stationnaires du jeu stochastique).

Nous allons donc définir une extension de FP pour les jeux stochastiques en ajoutant un ensemble de variables représentant les continuations, inspirés des Q-valeurs de Q-learning (section 1.7.1). Pour prouver la convergence de ces procédures en temps discret, on définit une dynamique de meilleure réponse dans les jeux stochastiques (section 1.7.2). Enfin, on définit une variation de FP avec des choix d'actions smooth, ce qui donne Smooth Fictitious Play (SFP) pour les jeux stochastiques (section 1.7.3).

1.7.1 Fictitious Play

Dans le chapitre 6, on propose des procédures pour des séquences de jeu *synchrones*, c'est-à-dire dans lesquels les joueurs donnent une action pour tous les états à chaque étape (il n'y a pas d'état courant, ou bien il est inconnu) en plus des procédures qu'on appelle *asynchrones*, qui correspondent aux séquences de jeu habituelles et présentées au-dessus. Dans ce résumé, on ne présente que le cas asynchrone pour en simplifier l'exposition.

On suppose dans cette section que le jeu est connu, c'est-à-dire que chaque joueur dispose des fonctions u_s^i, P_s et observe toutes les actions. On utilise les notations du modèle 4 à la page 20.

La procédure utilise deux ensembles de variables, définies à chaque étape n pour chaque joueur i et chaque état s : les actions empiriques $x_{s,n}^i$ et les estimations des continuations $v_{s,n}^i$.

Actions empiriques On commence par définir le nombre d'occurrences de l'action s :

$$\mu_{s,n} = \sharp\{k \mid 1 \leq k \leq n \wedge s_k = s\} \tag{1.8}$$

Cela permet de définir l'action empirique du joueur i dans l'état s comme un élément de $\Delta(A^i)$:

$$x_{s,n}^i := \frac{1}{\mu_{s,n}} \sum_{k=1}^{n} 1_{s_k=s} a_n^i \tag{1.9}$$

Estimation des continuations Pour définir les paiements de continuation, nous avons besoin du jeu auxiliaire de Shapley (Shapley, 1953). Étant donné un état s et un vecteur de continuation $v \in \mathbb{R}^{I \times S}$, ce jeu est défini comme un jeu en une étape avec A^i comme ensemble d'actions pour chaque joueur et les fonctions d'utilité définies comme suit :

$$f_{s,v^i}^i(\boldsymbol{a}) := (1-\delta) u_s^i(\boldsymbol{a}) + \delta \sum_{s' \in S} P_{s,s'}(\boldsymbol{a}) v_{s'}^i.$$

Chaque joueur i estime ses continuations dans un vecteur $v_{s,n}^i \in \mathbb{R}^S$, selon un taux de mise à jour α_n (et avec $\sigma_n := \sum_{k=1}^n \alpha_k$) :

$$v_{s,n+1}^i := \frac{1}{\sigma_n} \sum_{k=1}^{n} \alpha_n f_{s,v_k^i}^i \left(\boldsymbol{x}_{s,k}\right) \tag{1.10}$$

Sous forme incrémentale, et en rajoutant un facteur constant $\alpha^\star > 0$ pour contrôler la vitesse de convergence, cela donne :

$$v^i_{s,n+1} - v^i_{s,n} = \alpha^\star \frac{\alpha_n}{\sigma_n}(f^i_{s,v^i_n}(\boldsymbol{x}_{s,n}) - v^i_{s,n}). \tag{1.11}$$

Remarque. *Dans les jeux à somme nulle (respectivement à intérêts identiques), les valeurs de $v^i_{s,n}$ sont opposées (reps. identiques) pour tous les joueurs dès que les valeurs initiales sont les mêmes, on peut donc omettre l'exposant i.*

Choix des actions Chaque joueur choisit son action comme une meilleure réponse dans le jeu auxiliaire de Shapley :

$$a^i_n \in \mathrm{BR}^i_{s_n,v^i_{s,n}}\left(\boldsymbol{x}^{-i}_{s_n,n-1}\right) := \underset{b^i \in A^i}{\arg\max}\, f_{s_n,v_n}\left(b^i, \boldsymbol{x}^{-i}_{s_n,n-1}\right).$$

Convergence On appelle AFP le système qui résulte des définitions des actions empiriques, des continuations et des actions sélectionnées et on prouve les théorèmes suivants :

Théorème 1.1 (Convergence de FP dans les jeux à intérêts identiques)**.** *Sous l'hypothèse 6.1 (voir page 71), si tous les joueurs utilisent la procédure AFP, alors le système converge presque-sûrement vers l'ensemble des équilibres de Nash dans les jeux stochastiques escomptés à intérêts identiques et ergodiques.*

Théorème 1.2 (Convergence de FP dans les jeux à somme nulle)**.** *Sous l'hypothèse 6.1 (voir page 71) et si $\forall n \in \mathbb{N}, \alpha_n = 1$, alors il existe $A > 0$, indépendant du jeu, tel que si tous les joueurs utilisent AFP, alors le système converge presque-sûrement vers l'ensemble des $A\alpha^\star$-équilibres de Nash dans les jeux stochastiques escomptés à somme nulle et ergodiques.*

Pour prouver ces deux théorèmes, on va d'abord étudier la convergence d'un système analogue en temps continu et ensuite utiliser la théorie des approximations stochastiques (voir la section 4.4 du chapitre 4) pour relier le comportement limite des systèmes.

Remarque. *La théorie des approximations stochastiques (Benaim et al., 2005) ne fonctionne que lorsque le système continu est autonome, au sens où x^i_s et v_s sont des moyennes (éventuellement conditionnée à l'état actuel s_n (Perkins et Leslie, 2012)), ce qui implique que seul le cas $\alpha_n = \frac{1}{n}$ peut-être prouvé via un système continu.*

Pour les autres cas (dans le cas à intérêts identiques), nous avons écrit une preuve directement en temps discret dans le chapitre 7.

1.7.2 Dynamique de meilleure-réponse

La dynamique de meilleure réponse que nous définissons étend celle étudiée par Leslie et al. (2020) pour les jeux stochastiques à somme nulle. Le système asynchrone en temps continu correspondant à AFP est le suivant :

$$\begin{cases} \dot{v}_s(t) = \alpha(t)\left(f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s(t)\right) \\ \dot{x}^i_s(t) \in \beta_s(t)\left(\mathrm{BR}^i_{s,v_s(t)}\left(\boldsymbol{x}^{-i}_s(t)\right) - x^i_s(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{ABRD}$$

La fonction α contrôle la vitesse de mise à jour de v_s et β_s est la fréquence de mise à jour moyenne d'un état, c'est l'équivalent du $1_{s_n=s}$ en temps discret. Il y a une fréquence minimale $\beta_- > 0$ qui exprime le fait que le jeu est ergodique. Bien que le système soit plus général, intuitivement $\sum_{s\in S}\beta_s = 1$ et chaque β_s représente le nombre de fois que le système est dans s par unité de temps.

On définit des systèmes synchrones et complètement asynchrones analogues.

Convergence On prouve la convergence de ABRD dans les jeux stochastiques à intérêts identiques et les jeux stochastiques à somme nulle, ce qui donne les théorèmes suivants :

Théorème 1.3 (Convergence de ABRD dans les jeux stochastiques à intérêts identiques)**.**
Soit $\{v_s, \beta_s, x^i_s\}_{s\in S, i\in I}$ une solution de ABRD. Sous l'hypothèse 6.2 (page 74), il existe $\boldsymbol{v}_\infty \in \mathbb{R}^{|S|}$ tel que :

(i) *pour tout s, $f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) \xrightarrow[t\to\infty]{} v_{\infty,s}$ et $v_s(t) \xrightarrow[t\to\infty]{} v_{\infty,s}$*

(ii) $\boldsymbol{v}_\infty$ *est l'utilité d'un équilibre de Nash stationnaire*

(iii) $\{\boldsymbol{x}_s(t)\}_{s\in S}$ *converge vers l'ensemble des équilibres de Nash avec utilité* $\boldsymbol{v}_\infty$

Théorème 1.4 (Convergence de ABRD dans les jeux stochastiques à somme nulle)**.**
Soit $\{v_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ *une solution de ABRD. Il existe une constante* $A > 0$ *(qui ne dépend que du jeu et de* δ*) telle que pour tout* $\alpha^\star > \lim_{t\to\infty} \alpha(t)$ *et avec l'hypothèse 6.2 (page 74) :*

(i) *pour tout* s, $\limsup_{t\to\infty} |f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s(t)| \leq A\alpha^\star$

(ii) $\{x_s(t)\}_{s\in S}$ *converge vers l'ensemble des* $A\alpha^\star$*-équilibres stationnaires quand* $t \to \infty$.

En utilisant des théorèmes d'approximations stochastiques (dont un théorème de Perkins et Leslie (2012) que nous avons généralisé), on prouve les théorèmes de la section précédente.

1.7.3 Smooth Fictitious Play

Dans le chapitre 7, on définit une procédure analogue à FP du chapitre 6 mais on utilise une sélection d'action *smooth* (ou lisse). Cela permet en particulier que toutes les actions soient jouées avec une probabilité positive et par conséquent d'explorer le modèle. Ainsi, une première procédure est définie en supposant que le monde est connu, mais une deuxième estime le modèle au fur et à mesure de la séquence de jeu.

Pour cela, on utilise des fonctions de régularisation, ce qui est une technique largement répandue en informatique, par exemple avec « follow the perturbed leader » (Cesa-Bianchi et Lugosi, 2006) ou Smooth (ou Stochastic) Fictitious Play (Fudenberg et Levine, 1995). Ces fonctions de régularisations peuvent être interprétées ou bien comme une « main tremblante » (c'est-à-dire une forme d'incertitude sur l'action jouée) ou bien comme une incitation à jouer aléatoirement en donnant une récompense additionnelle aux joueurs qui n'emploient pas d'action pure. Formellement, un joueur i cherche à maximiser une fonction d'utilité perturbée égale à $u_s^i + \eta h^i$ avec η le paramètre de lissage et h^i la fonction de régularisation qui doit satisfaire l'hypothèse suivante :

Hypothèse 1.1.

$$h^i : \Delta(A^i) \to \mathbb{R}^+, \textit{ strictement concave en } x^i,\ C^1 \textit{ dans l'intérieur,}$$
$$\lim_{x^i \to \partial\Delta(A^i)} \|\nabla_{x^i} h^i\| = +\infty \textit{ et } \eta > 0$$

La fonction de régularisation est supposée être la même pour tous les joueurs. Elle peut par exemple être égale à l'entropie de Shannon, c'est-à-dire à :

$$h(y) = -\sum_{j\in I} \sum_{a^j \in A^j} y^j(a^j) \log\big(y^j(a^j)\big).$$

On peut ensuite définir la fonction de meilleure réponse régularisée, qui n'est cette fois-ci pas un ensemble d'actions pures, mais une action mixte qu'on voit comme une distribution sur les actions :

$$\mathrm{SBR}^i_{s,v^i}\big(\boldsymbol{x}_s^{-i}\big) := \operatorname*{arg\,max}_{y^i\in\Delta(A^i)} f^i_{s,v^i}\big(y^i, \boldsymbol{x}_s^{-i}\big) + \eta h\big(y^i, \boldsymbol{x}_s^{-i}\big). \tag{1.12}$$

Dans le cas de l'entropie de Shannon, la fonction de meilleure réponse régularisée est la fonction logit.

On obtient le système suivant :

$$\begin{cases} \forall s \in S, v^i_{s,n+1} - v^i_{s,n} = \dfrac{\alpha^\star}{n+1}\left(f^i_{s,v^i_n}\big(\boldsymbol{x}_{s,n}\big) + \eta h\big(x^i_{s,n}\big) - v^i_{s,n}\right) \\ a^i_{n+1} \sim \mathrm{SBR}^i_{s_{n+1},v_{n+1}}\left(\boldsymbol{x}^{-i}_{s_{n+1},n}\right) \\ \forall s \in S, x^i_{s,n+1} - x^i_{s,n} = \dfrac{1_{s=s_{n+1}}}{\mu_{s,n+1}}\left(a^i_{n+1} - x^i_{s,n}\right) \end{cases} \tag{SFP}$$

Fonctions de transition inconnues et paiements perturbés On suppose maintenant que les fonctions d'utilité du jeu u_s^i ne sont pas connues des joueurs, tout comme les fonctions de transition. On suppose de plus que le paiement qu'ils reçoivent est perturbé, c'est-à-dire qu'à chaque étape n, le joueur i obtient une récompense aléatoire R_n^i qui suit une loi déterminée par les actions $\boldsymbol{a}_n$ et l'état courant s_n, dont l'espérance est $u_{s_n}^i(\boldsymbol{a}_n)$ et la variance bornée conditionnellement à l'histoire :

$$\mathbb{E}\left[R_n^i|\mathcal{F}_{n-1}\right] = u_{s_n}^i(\boldsymbol{a}_n) \tag{1.13}$$

où $\mathcal{F}_{n-1}$ est la σ-algèbre qui contient toutes les informations jusqu'à l'étape $n-1$.

Ensuite, on peut estimer les fonctions de transition et d'utilité du jeu stochastique (équation MFP.1 de la page 87) et on obtient un système MFP qui est identique à SFP sauf pour la fonction du jeu auxiliaire de Shapley qui est remplacée par la fonction estimée.

Puis on en déduit les théorèmes suivants, analogues aux théorèmes de la section précédente. La convergence est prouvée vers les équilibres de Nash *régularisés*, c'est-à-dire ceux du jeu stochastique dont la fonction de paiement est $u_s^i + \eta h$ (Takahashi, 1964).

Théorème 1.5 (Convergence dans les jeux stochastiques à intérêts identiques)**.** *Dans un jeu stochastique, ergodique et à intérêts identiques, si tous les joueurs suivent la procédure SFP (ou MFP), pour tous les états s, leurs actions empiriques $\boldsymbol{x}_{s,n}$, convergent presque sûrement vers l'ensemble des équilibres de Nash régularisés et le vecteur des continuations $v_{s,n}^i$ converge vers la continuation optimale de l'ensemble des équilibres de la limite.*

Théorème 1.6 (Convergence dans les jeux stochastiques ergodiques à somme nulle)**.** *Dans un jeu stochastique, ergodique et à somme nulle, si tous les joueurs suivent la procédure SFP (ou MFP) avec les mêmes valeurs initiales, pour tous les états s, leurs actions empiriques $\boldsymbol{x}_{s,n}$, convergent presque sûrement vers l'ensemble des $D\alpha^\star$-équilibres de Nash régularisés (où $D > 0$ est une constante ne dépendant que de G) et le vecteur des continuations $v_{s,n}^i$ converge vers les continuations correspondantes.*

De la même manière que pour FP, on définit dans la suite du chapitre une dynamique de meilleure réponse qui correspond à SFP. On prouve des théorèmes analogues en temps continu, avant d'utiliser un théorème d'approximation stochastique pour en déduire les théorèmes 1.5 et 1.6.

1.7.4 Conclusion sur FP et SFP dans les jeux stochastiques

En s'appuyant sur des idées de Leslie et al. (2020) et Sayin et al. (2022a), nous avons proposé dans le chapitre 6, une extension de FP aux jeux stochastiques. Nous avons ensuite défini dans le chapitre 7 une généralisation de SFP. Nous avons montré que, lorsqu'ils sont utilisés conjointement, nos algorithmes convergent vers des équilibres (régularisés) stationnaires des jeux stochastiques dans le cas des jeux à intérêts identiques et à somme nulle.

Un certain nombre de questions restent sans réponse, que ce soit avec nos procédures ou d'autres algorithmes récents d'apprentissage décentralisé dans les jeux stochastiques.

Premièrement, la théorie des approximations stochastiques ne donne aucune indication sur la vitesse de convergence des algorithmes correspondants en temps discret. De plus, même en temps continu, cette question reste ouverte, y compris pour les jeux potentiels (Harris, 1998). Par conséquent, on peut se demander, tant dans le cadre des jeux répétés ou stochastiques, quelle garantie FP et SFP ont-elles concernant le taux de convergence ?

Deuxièmement, et en s'inspirant de la convergence de Vanishingly Smooth FP vers les équilibres dans les jeux répétés classiques (voir Benaïm et Faure, 2013 ; Hadikhanloo et al., 2021), il serait intéressant d'étudier une version vanishingly smooth de notre procédure SFP.

Enfin, l'obtention d'une convergence *last-iterate* au lieu d'une convergence moyenne temporelle est une autre direction intéressante. Cependant, on sait qu'il n'y a pas nécessairement ce type de convergence avec SFP dans le cas non stochastique : comme le montre Giannou et al. (2021), dans tout jeu répété, SFP ne converge pas vers un profil d'équilibre à moins qu'il n'existe un équilibre strict.

1.8 Une procédure de non-regret avec une file d'attente

Le chapitre 8 est consacré à l'étude d'un modèle de file d'attente en temps discret, qui se déroule sur une suite de périodes. Au début de chaque période, les joueurs reçoivent un paquet et doivent choisir

le moment de la période auquel ils envoient leur paquet dans la file d'attente. La file traite les paquets suivant le principe du premier arrivé, premier servi, avec un paquet traité par unité de temps.

Le choix des joueurs est stratégique : si le paquet arrive après la fin de la période, alors on inflige une pénalité élevée au joueur (dans le but qu'elle soit dissuasive), mais les joueurs sont aussi incités à envoyer les paquets le plus tard possible.

De plus, on suppose que les paquets qui ne sont pas arrivés à temps sont de nouveau présents dans la période suivante. Par conséquent, certains joueurs peuvent avoir plus d'un paquet mais doivent toujours les envoyer ensemble. Ainsi, se pose la question de la stabilité du modèle : sous quelle condition le nombre de paquets reste-t-il borné ?

On cherche à répondre à cette question pour deux concepts de solutions différents : d'abord en supposant que les joueurs sont myopes et jouent des équilibres (section 1.8.3) et dans le cas où ils utilisent une procédure de non-regret (section 1.8.4). On commence par une brève revue de littérature (section 1.8.1) et la description du modèle (section 1.8.2).

1.8.1 Littérature

L'analyse des comportements stratégiques dans les files d'attente remonte à l'article de Naor (1969), qui a étudié les files d'attente avec une règle de type premier arrivé-premier servi. Dans son système, l'utilité des agents est une récompense qu'ils obtiennent lorsqu'ils sont servis, moins un coût d'attente proportionnel au temps d'attente dans la file. Les équilibres de Nash ne sont pas efficaces socialement car les agents ne prennent pas en compte les externalités.

La littérature sur les files stratégiques a ensuite explosé, voir (Hassin et Haviv, 2003) pour une revue de la littérature.

Dans notre cas, nous nous intéressons à un système où les agents stratégiques peuvent décider quand rejoindre la file, à la manière de l'article initial de Glazer et Hassin (1983).

Nous proposons dans le chapitre 8 une version dynamique du modèle de l'article de Rivera et al. (2018b), qui est une version discrète du modèle de goulot d'étranglement de Vickrey (1969).

1.8.2 Modèle

Soit I l'ensemble des joueurs de cardinal supérieur ou égal à 3. On suppose que le temps est divisé en périodes de durées égales à $L \geq 2$. La variable t désigne une période générique[1].

Au début d'une période $t \geq 0$, chaque joueur reçoit un paquet et choisit (indépendamment) une action $a_t^i \in \{0, \dots, L-1\}$. Le choix de ces actions peut dépendre de l'histoire (comme dans le modèle 4).

Lorsqu'un paquet n'est pas sorti de la file à la fin de la période, il est conservé par le joueur pour la période suivante. L'état du système au temps t est un vecteur $\boldsymbol{k}_t = (k_t^i)_{i \in I}$, avec k_t^i le nombre de paquets d'un joueur au début de la période t. L'espace d'état est donc $S := \mathbb{N}^I$ et le nombre de paquets au temps t est :

$$k_t := \sum_{i \in I} k_t^i. \tag{1.14}$$

Les paquets en retard se voient infliger une pénalité C_{k_t} qui dépend du nombre total de paquets dans le système. À la fin de la période t, le joueur i paye un coût :

$$c^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t) = k_t^i(L - a_t^i) + C_{k_t}\mathbb{E}\left[\text{nombre de paquets de } i \text{ qui sont en retard en } t\right]. \tag{1.15}$$

On suppose que $L \geq I$. Sans cette hypothèse, le nombre de paquets dans le système est nécessairement non borné, car la file peut traiter moins de paquets en une période que le nombre de nouveaux paquets.

1.8.3 Joueurs myopes et stratégiques

Dans cette section, on suppose que les joueurs sont stratégiques mais myopes. Ils jouent à chaque instant un équilibre. On commence par étudier la structure des équilibres de Nash et corrélés dans les cas $L \geq k_t$ et $L < k_t$.

[1] Il y a donc un changement de notation : alors que dans les sections précédentes t désignait la variable de temps continu, cela désigne maintenant une variable de temps discret.

Théorème 1.7 (Structure des équilibres de Nash si $L \geq k_t$)**.** *Si $L \geq k_t$, $C_{k_t} > k_t^2$, et $\boldsymbol{y}_t$ est un équilibre de Nash, alors :*

(i) *pour tous les joueurs j et actions $a_t^j < L - k_t$, $y_t^j(a_t^j) = 0$;*

(ii) *il existe un joueur i tel que*

(a) $y_t^i(L - k_t) > 0$,

(b) *pour tout $a_t^i > L - k_t + 1$, $y_t^i(a_t^i) = 0$;*

(iii) *pour tout joueur $j \neq i$, $y_t^j(L - k_t + 1) > 0$;*

(iv) *pour tout joueur j, si $y_t^j(L - k_t) > 0$, alors $y_t^j = y_t^i$;*

(v) $k_t^2 - k_t + 1 \leq \mathsf{SC}(\boldsymbol{y}_t) \leq k_t^2$.

Un corollaire de ce théorème est que même en supposant que les joueurs jouent un équilibre de Nash à chaque étape, il est possible que des paquets soient en retard. Par conséquent, on étudie aussi la structure des équilibres coarse-corrélés (qui comprennent les équilibres de Nash) dans le cas $L < k_t$ avec le théorème ci-dessous :

Théorème 1.8. *Si $k_t > L$ et $C_{k_t} > k_t^2 L$, alors il existe un unique équilibre coarse-corrélé, qui est le profil pur où tous les joueurs ont pour action 0.*

Ce dernier théorème garantit la stabilité du système tant que les joueurs jouent un équilibre, qu'il soit corrélé, coarse-corrélé ou de Nash. En effet, on va avoir une alternance entre deux phases : celle où le système est dans le cas $L \geq k_t$ et celle où $L < k_t$. Dans la première phase, qui est celle du théorème 1.7, le nombre de paquets peut augmenter (voir le chapitre 8 pour une explication détaillée) jusqu'à atteindre ou dépasser L. Dans cette seconde phase, on est dans le cas du théorème 1.8 et le nombre de paquets redescend.

1.8.4 Comportement avec une procédure de non-regret

On va maintenant étudier le cas où les joueurs utilisent un algorithme de type Exponential Weight (EWA) pour apprendre. On définit une version d'Exponential Weight à plusieurs niveaux pour que les joueurs apprennent selon le nombre de leur paquets, il s'agit de l'algorithme 1.

Algorithme 1 EWA multi-niveaux

$\forall b^i \in A^i$, initialisation de $w_1^i(b^i)$
$\forall i \in I, k^i \leftarrow 1$ ▷ niveau 1
for $t \geq 1$ **do**
  for $i \in I$ **do**
    sélection d'une action $a_t^i \sim x_{k^i}^i$ ▷ proportionnellement à $w_{k^i}^i$
  end for
  for $i \in I$ **do**
    $\forall b^i \in A^i, w_{k^i}^i(b^i) \leftarrow w_{k^i}^i(b^i) \exp(-\eta c_{\boldsymbol{k}_t}^i(b^i, \boldsymbol{a}_t^{-i}))$ ▷ nombre de paquets en retard + 1
    $k^i \leftarrow \tilde{k}^i + 1$
    if $w_{k^i}^i$ n'est pas définit **then**
      $\forall b^i \in A^i$, initialisation de $w_{k^i}^i(b^i)$ ▷ niveau k^i
    end if
  end for
end for

L'idée est de maintenir une liste de poids $w_{t,n}^i(b^i)$ pour chaque joueur i, niveau n (qui représente le nombre de paquets du joueur i) et action $\boldsymbol{b}$. Ces poids sont multipliés par l'utilité de l'action $\boldsymbol{b}$ lorsque le nombre de paquets courants k_t^i est égal à n.

Cette procédure est inspirée d'une procédure sans-regret car avec un paramètre η, EWA a un regret proportionnel à ηt (et donc, classiquement, avec une valeur de η judicieusement choisie, EWA est sans-regret).

On prouve qu'avec l'algorithme 1, le système est stable, ce qui donne le théorème suivant :

Théorème 1.9 (Stabilité d'EWA multi-niveaux utilisé conjointement dans le cas $L < I$)**.** *Si $I < L$ et $C_k > 4kL$ pour tout k_t et si tous les joueurs utilisent EWA multi-niveaux, alors le système est stable, c'est-à-dire que le nombre de paquets est borné presque-sûrement.*

1.8.5 Conclusion sur les systèmes de file d'attente

Nous avons étudié un jeu avec une file d'attente où les joueurs qui ne font pas passer leurs paquets avant une date limite reçoivent une pénalité dissuasive. Plus précisément, nous avons étudié le comportement d'un tel système dans deux cas : premièrement, lorsque les joueurs sont stratégiques, ce qui signifie qu'ils jouent conjointement un équilibre (de Nash, corrélé ou coarse corrélé). Deuxièmement, nous supposons que tous les joueurs utilisent la même procédure inspirée de exponential weight algorithm (EWA), connue pour être sans regret. Contrairement à l'EWA standard, notre multi-level EWA (MLEWA) est sur plusieurs états, appelés niveaux. Dans ce cas, un niveau est le nombre de paquets détenus par un joueur.

Nous avons montré que dans les deux cas, si la pénalité est suffisamment dissuasive, alors le système est stable, c'est-à-dire que le nombre de paquets reste borné est limité. Par conséquent, on peut considérer qu'il s'agit d'un problème de *mechanism design* dans lequel le concepteur du système doit mettre en place des pénalités suffisamment élevées.

Notre modèle pourrait être étendu dans plusieurs directions. Tout d'abord, les joueurs pourraient utiliser des algorithmes différents de MLEWA, peut-être des algorithmes à plusieurs niveaux basés sur d'autres procédures - par exemple FP. La preuve de stabilité pourrait être abstraite de MLEWA pour comprendre quelles sont les propriétés nécessaires pour qu'une procédure converge.

Deuxièmement, on pourrait généraliser le jeu à plusieurs files d'attente, comme dans Gaitonde et Tardos (2020b). Le choix de la file deviendrait un aspect stratégique supplémentaire.

Troisièmement, les valeurs limites de la pénalité pourraient être affinées. En particulier, nous ne savons pas s'il existe des cas intermédiaires entre le scénario stable et le scénario divergeant. Est-il possible que le système soit infiniment récurrent mais non stable ?

1.9 Conclusion

Cette thèse comporte trois contributions, toutes liées à l'apprentissage dans les jeux stochastiques ou répétés. Tout d'abord, nous avons réalisé une étude empirique de jeux répétés dans lesquels plusieurs joueurs utilisent des algorithmes de type Q-learning. Ensuite, nous avons étendu des procédures connues, Fictitious Play (FP) et Smooth Fictitious Play (SFP), aux jeux stochastiques à somme nulle et à intérêts identiques. Nous avons montré que plusieurs de ces procédures convergeaient. Le dernier chapitre de cette thèse est consacré à l'étude d'un jeu stochastique avec une file d'attente lorsque les joueurs utilisent un algorithme inspiré d'une procédure no-regret.

Tout au long de cette thèse, nous avons identifié des problèmes ouverts, communs à plusieurs chapitres, qui sont autant de pistes pour de futures recherches.

- Tout d'abord, la convergence des systèmes multi-agents simples, tels que les jeux répétés avec Q-learning, comme expérimenté dans le chapitre 5. Bien qu'il y ait eu des tentatives pour étudier empiriquement ce système (Babes et al., 2009 ; Wunder et al., 2010), il n'y a pas de garantie formelle (Sutton et Barto, 2018), même dans des classes simples de jeux.

  Dans les chapitres 6 et 7, nous avons prouvé la convergence de plusieurs classes d'algorithmes dans les jeux stochastiques à intérêts identiques et à somme nulle. La plupart de ces algorithmes utilisent deux échelles de temps (pour les continuations et les actions empiriques). Cependant, la théorie des approximations stochastiques (comme décrite dans (Benaim et al., 2005)) ne peut être utilisée directement dans les systèmes non autonomes en temps. Par conséquent, la correspondance entre les systèmes à temps continu à deux échelles de temps arbitraires et les systèmes à temps discret n'est pas facile à établir. Par ailleurs, pour des applications et en particulier des algorithmes de contrôle, avoir deux échelles de temps semble particulièrement complexe, tout comme le fait d'avoir beaucoup de variables différentes. Par conséquent, établir la convergence de systèmes avec une seule échelle de temps et/ou un ensemble de variables plus réduit serait conceptuellement

plus simple (la théorie des approximations stochastiques s'appliquerait plus facilement) et plus applicable à des problèmes de contrôle sur des systèmes réels.

- Ensuite, la définition du regret dans un environnement changeant lorsque la transition est endogène. En effet, la définition standard du regret convient au cas où l'environnement est inconscient, c'est-à-dire qu'il peut y avoir une variable d'état, mais que les joueurs ne l'influencent pas.

  En revanche, si nous supposons que la variable d'état est (au moins partiellement) contrôlée par les joueurs, la théorie du regret n'est pas adaptée : un changement d'action entraîne un changement de la récompense instantanée (ce qui est pris en compte par le regret), mais aussi de l'état, ce dont le regret ne tient pas compte.

  Il n'existe pas de définition largement acceptée du regret dans les jeux stochastiques à transitions endogènes. Wei et al. (2017) définit le regret comme la différence entre la valeur du jeu stochastique et le gain obtenu. Gaitonde et Tardos (2020b) utilisent la différence entre le gain obtenu et la somme des gains qui auraient été obtenus si le joueur avait changé ses actions, mais avec les mêmes états, c'est-à-dire si les transitions sont exogènes.

  En effet, une difficulté majeure est d'estimer ce qui se serait passé au temps t si l'état actuel était différent en raison de choix d'actions antérieurs différents. Une autre difficulté est le choix de la fonction de paiement.

  À un niveau plus abstrait, cela répondrait à la question suivante : en supposant que nous ayons un certain nombre d'experts qui recommandent des actions dans un certain nombre d'états, quels experts devraient être sélectionnés afin de maximiser le paiement escompté à long terme ?

  En ce qui concerne cette thèse, il serait intéressant de prouver des propriétés de ce type pour SFP. En effet, le fait que la procédure soit aléatoire devrait la rendre plus robuste aux adversaires, ce qui est généralement exprimé comme étant sans regret.

Chapter 2

Introduction

2.1 Online Learning in Stochastic Games

This thesis aims to be a step forward in the understanding of the dynamics of multiagent systems with learning agents. This is formalized as *online learning* in *stochastic games*, at the intersection of game theory and computer science.

Online learning is a field in mathematics and computer science that examines how to optimize a utility or loss function while interacting in an environment. It is typically supposed that interactions are a sequence of actions taken by an agent whose behavior is specified in algorithmic terms. Compared to the more general field of machine learning, information is only available to agents progressively, as feedback of their actions and the problem may be changing with time. Game theory is a subfield of applied mathematics and theoretical economics. It is dedicated to the study of strategic interactions of multiple agents. A *stochastic* game is a (dynamic) game which in addition to interacting agents, features an environment whose evolution may be influenced by agents.

This thesis studies new and existing learning procedures which can be used by agents to interact in a changing environment modeled as a stochastic game, and analyses the resulting dynamics, for instance via the time average behavior of the players.

Problem Setting Our model is formalized as follows: a stochastic game is characterized by transition and utility functions (denoted by u_s^i for player i and state s) whose arguments are the current state of the game and actions taken by several agents-also called players. A sequence of play is indexed by time n and unfolds as follows. Initially, the system is in state s_0, and at every step n, every player i chooses an action a_n^i from her action set A^i, which, with the current state s_n, defines the payoff of every player and the next state s_{n+1}. This is specified in Model 1 and in more details in Chapter 3.

Model 1 Playing a Stochastic Game

an initial state s_0 is chosen
for $n = 0, \dots$ **do**
  every player i chooses an action $a_n^i \in A^i$
  a new state s_{n+1} is drawn
  every player i gets $u_{s_n}^i(\boldsymbol{a}_n)$
end for

Players are supposed to be interested in optimizing their long term payoff, which may either be the Cesaro-average of all payoffs, or the discounted sum of payoffs

$$\sum_{n=0}^{\infty} \delta^n u_{s_n}^i(\boldsymbol{a}_n)$$

where $\delta \in (0, 1)$ is the discount factor.

We are interested in the case where all players choose their actions following a (potentially randomized) algorithm, based on previous observations (which include, for each player, its own payoff and

may or may not include other pieces of information such as actions played by all players). When these algorithms are specified, Model 1 is a dynamical system whose behavior can be studied.

Examples Consider a network of energy producers that have to face demand and adjust production accordingly, under the assumption that they are controlled independently but want to avoid at all cost a production lower than the demand (or, formally, incur a deterrent penalty). This is a multiagent system where producers have to cooperate and find a combination that minimizes day-to-day reward following the current state (the temperature and previous production level if changing them is costly).

In economics, a well known illustration is the Bertrand competition game, where several firms have to compete on a market, possibly facing demand shocks. Both examples are detailed and formally defined in Section 3.4.

Another example studied in this thesis is that of queuing systems or bottleneck: a number of agents have to send a number of jobs through a single queue. Time is divided in a sequence of periods of equal duration. At the start of every period, each player gets a job and has to choose at which step of the period it is sent. This step is strategic since they want to minimize the waiting time of the job (i.e., they want to keep it as long as possible). However, when their jobs arrive after a fixed deadline (i.e., the end of the period), they incur a deterrent penalty. If we suppose that late jobs stay in the system at the next period, then this is modeled with repeated interactions between agents with a changing environment, i.e., it is a stochastic game.

While outside the scope of this thesis, it is worth noting that along with multiagent systems in computer science, evolutionary biology is another field where concepts of game theory can be applied very well, both theoretically (Weibull, 2004) or empirically with for instance studies on spider colony Hammerstein and Riechert (1988).

Solution Concepts The definition of a particular (stochastic) game specifies how much a player benefits from the combination of actions taken by all players. In computer science, utility is usually measured as the opposite of a cost, and a cost function represents the accuracy of a prediction. In economics and political science, the definition of utility functions is far from a trivial task. Such functions may not necessarily exist and trying to define one may lead to contradictions (Morrow, 1994). However, under the assumption of ordinal preferences, one can define utility functions.

Once utility functions are defined (and additional transition functions for stochastic games), nothing is said yet about the *behavior* of the system, i.e., what players actually do. This is why game theorists have defined so-called *solution concepts* that aim to define what is an actual result of the system under a set of assumptions, for instance supposing that players are rational.

The traditional solution concepts are equilibria, which are strategy profiles that are stable, i.e., no player has incentives to unilaterally deviate. A well-known solution concept is the Nash equilibrium (Nash, 1950). However, this was the beginning of much work as it raises more questions that it solves: how can these equilibria be computed? How many of them are they? Do players really coordinate their actions to play an equilibrium ?

Numerous mathematicians proposed alternative solution concepts (for instance the correlated equilibria of Aumann (1974)) or underlying theories such as the theory of equilibrium selection by Harsanyi and Selten, 1988.

Another fruitful line of work has been the definition of learning procedures as solution concepts: assumptions are made on the behavior of agents (for instance, they play with an algorithm) and then systems can be studied dynamically. An interesting question is the link between learning as a solution concept and traditional solution concepts: do learning procedures lead to equilibria? This is the constructive approach pursued in this thesis.

Learning Procedures In both computer science and game theory, multiple learning procedures have been proposed. In particular, Multiagent Reinforcement Learning (MARL) (Busoniu et al., 2008) is the subfield of computer science that deals with procedures that provide actions to play in environments with several players. Somewhat counter-intuitively, it may be considered as a special case of Reinforcement Learning (see Sutton and Barto (2018) for a contemporary reference) where a single agent faces a changing environment. Indeed, in MARL, the environment is composed of other players, which is an exploitable piece of information.

Reinforcement Learning is a broad field in computer science and features well-known algorithms such as Q-learning (Watkins, 1989). Q-learning is defined using a table of Q-values that represent the expected value of a state action pair, that is the total future payoff $Q(s, a^i)$ that player i can expect if he plays action a^i in state s. Q-values are updated following the simple scheme

$$Q_{n+1}(s_n, a^i_n) = Q_n(s_n, a^i_n) + \alpha(r^i_n + \delta \max_{b^i \in A^i} Q_n(s_{n+1}, b^i) - Q_n(s_n, a^i_n)) \tag{2.1}$$

where α is a learning rate, and other Q-values are unchanged. Using these Q-values, the player may choose an action with a maximal Q-value or explore other actions, following the traditional exploration-exploitation dilemma.

In game theory, learning procedures are almost as old as game theory itself (Brown, 1951; Robinson, 1951) with an ever-growing literature (Fudenberg and Levine, 1998; Cesa-Bianchi and Lugosi, 2006; Hart and Mas-Colell, 2013). One of the oldest procedures is Fictitious Play (Brown, 1951; Robinson, 1951). It supposes that players play a best-response to the past empirical average actions of other players.

As explained by Shoham et al. (2007), there are multiple research agendas for MARL, sometimes entangled. Three of these agendas are of particular interest in the context of this thesis. First, the computational agenda includes procedures that can be used to compute equilibria. Second, systems whose agents follow a set of learning rules can be described, i.e., learning is a solution concept itself. Third, the prescriptive agenda deals with how players *should* learn to achieve a specified goal, for instance be robust to adversarial attack.

This thesis is mostly devoted to the second agenda: we investigate the long term behavior of systems whose players' behavior is defined with relatively simple procedures.

2.2 This Thesis

This thesis comprises three main works: an experimental study of Q-learning in repeated and stochastic games, the definition and the proof of convergence of several procedures inspired from Fictitious Play in stochastic games, and the study of a no-regret algorithm in a game with queues.

Background Chapter 3 introduces game-theoretic definitions, notations and examples used throughout this document. There is also a symbol list in the beginning of the document and a glossary at the end of the manuscript. Chapter 4 describes the literature on learning in repeated and stochastic games, which is a growing field with links to online optimization. The mathematical framework of stochastic approximations is also explained, as we made an extension of this framework in order to prove the convergence of our procedures.

Q-learning in repeated and stochastic games Chapter 5 shows with several experiments that even with simple learning rules, the behavior of multiagent systems is complex. More precisely, it shows that memory may have negative impacts depending on the game under study, and that most experimental results are robust to changes in exploration schemes.

Fictitious Play and Smooth Fictitious Play Chapters 6 and 7 respectively present two families of learning procedures in stochastic games. Procedures of Chapter 6 are built using ideas from both Fictitious Play (FP) and Q-learning. Chapter 7 extends these procedures to smooth action selection and in an unknown game.

These chapters pursue different agendas. First, they are descriptive in the sense that they study the convergence of simple learning rules inspired from well-known ones that are Fictitious Play and Q-learning. Although proving convergence of such systems in general cases is probably a dead end due to mathematical tractability, it is interesting to understand which assumptions play a key role in the convergence. Second, they provide procedures that could be starting points to design algorithms that control real-world systems. Third, these procedures, although not designed to be efficient for this task, may be used to compute equilibria of stochastic games, especially the synchronous versions of our procedures.

Queuing Systems The last chapter is dedicated to the study of a particular game with a queue. Players have to send some packets, named *jobs* through a queue.

More precisely, time is divided in a sequence of periods of equal duration. At the start of every period, each player gets a job and has to choose at which step of the period it is sent. Then, the utility of each player is proportional to the time during which they keep the packet: the later they send it, the better the utility.

However, packets exit the queue in a first-in first-out order and when they arrive after a fixed deadline (i.e., the end of the period), players get a deterrent penalty. Therefore, the step during which a player chooses to send the packet is strategic.

We study this game in a repeated setting where jobs that were not processed in a period are present in the next one. We suppose that players use a no-regret algorithm and are interested in how the system behave, in particular whether the number of jobs stays bounded or not. Under appropriate conditions on the penalty parameter, the number of jobs is proven to be almost surely bounded.

2.3 Publications

Several papers were written during my thesis. The two first ones were written with one of my PhD supervisors Rida Laraki and a third one was written with Marco Scarsini and Xavier Venel who invited me at LUISS (Roma) during two months to study strategic queuing systems.

- Lucas Baudin and Rida Laraki. "Fictitious Play and Best-Response Dynamics in Identical-Interest and Zero-Sum Stochastic Games". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 17–23, 2022, pp. 1664–1690

  In this paper, we extend FP to discounted stochastic game with a family of procedures. We prove that they converge to the set of stationary Nash equilibria for identical-interest and zero-sum discounted stochastic games. This extends similar convergence results for non-stochastic games. In order to prove results in team and zero-sum stochastic games, we analyse the continuous-time counterpart these discrete time procedures. They include as a particular case the best-response dynamic introduced and studied by Leslie et al. (2020) in the context of zero-sum stochastic games. We prove the convergence of this dynamic to stationary Nash equilibria in identical-interests and zero-sum discounted stochastic games. Then, the stochastic approximation is used in order to obtain convergence results in discrete time.

- Lucas Baudin and Rida Laraki. "Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions". In: *Advances in Neural Information Processing Systems*. 2022

  This paper is a sequel to the preceding one where agents estimate a model of the stochastic game during the sequence of play, i.e., the model is not known a priori. This is done using regularizer functions inside algorithms of the first paper. Our novel procedures can be interpreted as extensions to stochastic games of the classical smooth fictitious play learning procedures in static games (where players best-responses are regularized, thanks to a smooth perturbation of their payoff functions). The convergence to stationary regularized Nash equilibria is proven for the same classes of dynamic games (zero-sum and identical-interests discounted stochastic games). Our paper also introduces the continuous smooth best-response dynamics counterparts.

  In the case of a MDP (a one-player stochastic game), our procedures globally converge to the optimal stationary policy of the regularized problem.

- Not yet published: Lucas Baudin et al. *Strategic Behavior and No-Regret Learning in Queueing Systems*. Preprint, 2023. arXiv: `2302.03614 [cs.GT]`

  This paper was written with Marco Scarsini and Xavier Venel. It started as a dynamic version of Rivera et al., 2018b.

  We study a dynamic discrete-time queuing model where at every period players get a new job and must send all their jobs to a queue that has a limited capacity. Players have an incentive to send their jobs as late as possible; however if a job does not exit the queue by a fixed deadline, then the owner of the job incurs a penalty and this job is sent back to the player and joins the queue at

the next period. Therefore, stability, i.e., the boundedness of the number of jobs in the system, is not guaranteed. We show that if players are myopically strategic, then the system is stable when the penalty is high enough. Moreover, if players use a learning algorithm derived from a typical no-regret algorithm (exponential weight), then the system is stable when penalties are greater than a bound that depends on the total number of jobs in the system

Chapter 3

Background on Games

In this chapter, we introduce classical definitions and notations (they are also referenced in the symbol list on page 4). Section 3.4 is dedicated to running examples.

The founding book of game theory is unquestionably *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern, published in 1944. This framework is nowadays widely established in economics but originates from mathematics of the second world war (Erickson, 2015). After 1944, game was extensively funded for at least two reasons. First, after the importance of mathematics during the war, it seemed sensible to continue research in areas relevant to the Cold War. Game theory, with its application to planning and prediction, in particular for nuclear deterrence, was inescapable. Second, the pool of well-respected, wartime mathematicians was substantial. Moreover, game theory was studied in applied mathematics department of universities while military-funded think tanks, such as the RAND Corporation, took shape. Numerous game theorists carried out research during their time at this company. For instance, Loyd S. Shapley and George W. Brown defined respectively stochastic games and fictitious play (at the same time as Julia Robinson) while working at RAND. The combination of these theories is central in this thesis.

Consistently with the original work of von Neumann and Morgenstern, game theory became central in economics. In particular in teaching of economics, vocabulary from game theory is widespread in most curriculum, when they do not comprise a course dedicated to game theory. This attracted its fair share of criticism (Amadae, 2015), as it led to grant an unreasonable importance to agents rationality, even though numerous economists strived to qualify this assumption. Interestingly, algorithmic-led devices may perfectly fit into the framework of game theory, as they are reliable, mostly deterministic, and not disrupted by common thought.

Stochastic games were introduced by Shapley (1953) quickly after von Neumann and Morgenstern's book. They add to the theory of games the concept of *states*, which model the environment. Actions taken by agents have influence on the evolution of the state variable. They may entirely determine the new state, have a partial influence or none at all. Therefore, players face an additional strategic tradeoff: they may take into account that they can orient the state towards one that yields higher payoffs, in spite of a lower contemporary reward. This arbitration depends on the evaluation of future payoffs. For instance, they may be discounted, meaning that instantaneous payoffs are preferred up to a factor (and this recursively) or all payoffs may be averaged uniformly.

In computer science, these systems are also known as multiagent systems and a number of engineering applications fit into this framework, even though the stochastic game formalism is not necessarily used.

3.1 Normal-Form Games

First, we define normal-form (also called strategic form by, for instance, Fudenberg and Tirole (1991)) games in a non-cooperative manner.

Definition 3.1 (Game)**.** A game G is a tuple $(I, \boldsymbol{A}, (u^i)_{i\in I})$ where

- I is the finite set of players,
- $\boldsymbol{A}$ is the set of action profiles, that is a product of action sets $\prod_{i\in I} A^i$ where A^i is the finite set of actions of player i,
- and $u^i : \boldsymbol{A} \to \mathbb{R}$ is the utility (or payoff) function of player i.

A mixed action is a probability distribution over (pure) actions, i.e., a member of $\Delta(A^i)$ for player i. Functions on pure action profiles such as utility functions are typically extended by n-linearity to mixed action profiles, that is to functions with domain $\Pi_{i\in I}\Delta(A^i)$.

Throughout this thesis (and even if the context should be clear enough), bold face letters typically denote action profiles, such as $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ for pure action profiles, while standard letters a^i, b^i, c^i denote individual pure actions of player i. Mixed action profiles are typically denoted by $\boldsymbol{y}$. As usual, $(b^i, \boldsymbol{a}^{-i})$ denotes a pure action profile where action of i is b^i and action of any other player $j \neq i$ is a^j.

Below are two simple classes of games that will be particularly studied in this thesis. Zero-sum games are two player, fully competitive games where an increase in one player's payoff implies a decrease of the other player's payoff. On the contrary, identical-interest games are fully cooperative games where all players share the same utility functions.

Definition 3.2 (Zero-Sum Game)**.** A zero-sum game is a two-player game such that $u^1 = -u^2$.

Definition 3.3 (Identical-Interest Game)**.** A game has identical interests when all players share the same utility function, i.e., there exists $u : \boldsymbol{A} \to \mathbb{R}$ such that $\forall i \in I, u^i = u$.

Identical-interest games may be generalized to potential games (Monderer and Shapley, 1996b) where a player i's payoff variation has the same variation as a potential function when the only changing action is the one of i. Formally, a game is an exact potential game if there exists a function $P : \boldsymbol{A} \to \mathbb{R}$ such that for all i, $\boldsymbol{a}$, b^i,

$$u^i(\boldsymbol{a}) - u^i\big(b^i, \boldsymbol{a}^{-i}\big) = P(\boldsymbol{a}) - P(b^i, \boldsymbol{a}^{-i}). \tag{3.1}$$

More generally, non-exact potential games are games such that the postivity of the left handside of Eq. (3.1) implies the positivity of its right handside.

3.1.1 Repeated Games

In order to learn, games must be repeated so that players can build on history to choose strategies. Therefore, the notion of repeated games is central in this thesis and is later generalized to games with states (i.e., stochastic games).

Definition 3.4 (Repeated Game)**.** Given a game $G = (I, \boldsymbol{A}, (u^i)_{i\in I})$, we define a so-called *repeated* (or iterated) game indexed by $n \in \mathbb{N}$ whose play unfolds as follows: at every step n, every player i chooses a (potentially random) action $a^i_n \in A^i$, which is observed by all players. Therefore, every player has a strategy that depends on the *history* of the game. Formally, a strategy of player i is an element of

$$\bigcup_{n=0}^{\infty} \boldsymbol{A}^n \to \Delta(A^i)$$

where A^i is the action set of player i and $\boldsymbol{A}^n$ is the list of action profiles up to stage n. The discounted payoff of player i is the real value

$$\sum_{n=0}^{\infty} \delta^n u^i(\boldsymbol{a}_n)$$

where $\delta \in [0, 1)$ is the discount factor.

Remark. *A repeated game has itself the structure of a game, its action set being the set of strategies and its utility being the discounted utility. Strategies of Definition 3.4 are sometimes called* behavioral strategies *by some authors to disambiguate.*

3.1.2 Stochastic Games

Introduced by Shapley (1953), stochastic games are games equipped with a state structure. A sequence of play unfolds on a number of steps (similarly to repeated games) and starts with an initial state, which may then change, potentially influenced by actions taken by players. As a consequence, players may face a trade-off between instantaneous rewards and future rewards because they may be able to influence the state variable favorably or not.

We start with the definition of one-player stochastic games, called Markov Decision Processes (Bellman, 1957).

Definition 3.5 (Markov Decision Process)**.** A (stationary) Markov Decision Process is a tuple

$$(A, S, u_s, (P_s)_{s\in S})$$

where

- A is the set of actions,
- S is the (finite, unless specified otherwise) set of states,
- $u_s : A \to \mathbb{R}$ is the state-dependent utility function,
- and $P_s : A \to \Delta(S)$ is the transition function.

This classical definition assumes that Markov Decision Processes (MDPs) are stationary in the sense that utility and transition functions do not depend on time but only on the state. In practice, this may be an optimistic assumption as the environment may be non-stationary or even adversarial.

In fact, a stationary MDP is a special case of stochastic games with a single player:

Definition 3.6 (Stochastic Game)**.** A stochastic game $\mathcal{G}$ is a tuple $(I, \boldsymbol{A}, S, (u_s^i)_{i\in I, s\in S}, (P_s)_{s\in S})$ where

- S is the set of states,
- $u_s^i : \boldsymbol{A} \to \mathbb{R}$ is the state-dependent utility function of player i,
- and $P_s : \boldsymbol{A} \to \Delta(S)$ is the transition function.

On the contrary, from the perspective of one player, a stochastic game can be considered as a non-stationary MDP where the environment is defined by the strategies of other players which may change over time.

Remark. *We suppose that players have the same action sets in all states. Since action sets and states are supposed to be finite, this is not an important restriction as most results in this thesis could be extended to different action sets per state.*

Sequence of play Similarly to repeated games, a play in a stochastic game unfolds on steps $n \in \mathbb{N}$, with players choosing action profile $\boldsymbol{a}_n$ at step n. It starts from an arbitrary state s_0, and then a new state s_{n+1} is selected randomly following distribution $P_{s_n}(\boldsymbol{a}_n)$. Player i obtains the discounted payoff

$$U^i_{\delta,s_0} := \sum_{n=0}^{\infty} \delta^n u^i_{s_n}(\boldsymbol{a}_n).$$

Equipped with this payoff function, we say that the stochastic game is a discounted stochastic game (DSG). Other payoff aggregations are possible, for instance uniform payoffs where one considers the limit of

$$U^i_{n,s_0} := \frac{1}{n+1} \sum_{k=0}^{n} u^i_{s_k}(\boldsymbol{a}_k)$$

when n goes to infinity. This aggregation has the interesting property that it does not depend on the initial state when the game is ergodic (see Definition 3.7). However, such an aggregation is not used in this thesis.

Unless otherwise noted, we suppose that all players observe all actions, i.e., information is complete.

Repeated games A repeated game can be seen as a stochastic game with a single state.

Ergodicity A central notion for a state is that of *recurrence*, which is defined as the property that the state is visited infinitely often.

Definition 3.7 (Ergodicity)**.** A stochastic game is ergodic if for all states $s, s' \in S$, there exists $N > 0$ such that for any sequence of play of length N starting from state s, then s' is visited during the play with positive probability.

3.1.3 Histories

The concept of history in repeated games is generalized to stochastic games with the following definition.

Definition 3.8 (History)**.** In a repeated game, the set of histories is the set $\bigcup_{n\geq 0} \boldsymbol{A}^n$. In a stochastic game, it is the set of past states and actions augmented with the current state, so it is

$$\bigcup_{n\geq 0} (S \times \boldsymbol{A})^n \times S.$$

In either case, it is denoted by $\mathcal{H}$ with typical element h_n.

Bounded Memory Another case of stochastic games is that of repeated games *with bounded memory*. Formally, starting from a (static) game, the N-bounded memory repeated game is defined with its (finite) state space equal to

$$S = \bigcup_{0\leq n\leq N} \boldsymbol{A}^n$$

and utility functions equal to the static utility functions (therefore they do not depend on the state). At every step, the new state is defined by forgetting the first element of the tuple and appending the profile of actions played.

3.2 Equilibria

Definitions of games and stochastic games say little as of the behavior of systems they are supposed to model. "Solution concepts" are normative ways that specify which action each player chooses. In particular, *equilibria* are a class of solution concepts that comprise a number of "stable" action profiles (which, in the case of multiple equilibria, leads to the theory of equilibrium selection (Harsanyi and Selten, 1988)). There are multiple definitions of equilibria, and we review some of them in this subsection.

The definition and the existence of the most famous concept of equilibria were introduced and proved by Nash (1950).

Definition 3.9 (Nash Equilibrium in (one-shot) Games)**.** A Nash equilibrium is a (potentially mixed) action profile $\boldsymbol{y}$ such that no unilateral deviation is profitable, that is to say

$$\forall i \in I, \forall b^i \in A^i, u^i(\boldsymbol{y}) \geq u^i\big(b^i, \boldsymbol{y}^{-i}\big). \tag{3.2}$$

Nash equilibria in discounted repeated games were characterized by Fudenberg and Maskin (1986). This definition can be extended to stochastic games (Shapley, 1953; Fink, 1964):

Definition 3.10 (Nash equilibrium in stochastic games with an initial state)**.** For an initial state $s_0 \in S$, a mixed profile $\boldsymbol{y} : S \to \Pi_{i\in I}\Delta(A^i)$ is a stationary Nash equilibrium when there is no profitable deviation in the discounted payoff.

Fink (1964) showed that in the case of discounted stochastic games, there are mixed profiles that are Nash equilibria for any initial state. Such profiles are called Nash equilibria of a stochastic game.

Nash equilibria suppose that players choose their potentially random actions independently. On the contrary, other solution concepts hinge on the assumption that player's actions may be correlated. We expose two such solution concepts for one-shot games below.

Correlated distributions are elements of set $\Delta(\boldsymbol{A})$. This corresponds to a probability distribution over profiles, so players can play a correlated distribution only when they have access to a correlation device. Payoff functions are linearly extended to correlated distributions.

Aumann (1974) showed the existence of *correlated* equilibria which are a special case of correlated distributions.

Let $\boldsymbol{z} \in \Delta(\boldsymbol{A})$ be a correlated distribution. For a player i and pure actions b^i and c^i, we write $\boldsymbol{z}_{b^i \to c^i}$ for the distribution such that

$$\boldsymbol{z}_{b^i \to c^i}(\boldsymbol{a}) = \begin{cases} \boldsymbol{z}(\boldsymbol{a}) + \boldsymbol{z}(b^i, \boldsymbol{a}^{-i}) & \text{if } a^i = c^i \\ 0 & \text{if } a^i = b^i \\ \boldsymbol{z}(\boldsymbol{a}) & \text{otherwise.} \end{cases}$$

Then, $\boldsymbol{z}_{b^i \to c^i}$ is also a correlated distribution, which is informally equal to distribution $\boldsymbol{z}$ where action a^i has been replaced by action c^i.

Definition 3.11 (Correlated Equilibrium in (one-shot) Games)**.** A correlated distribution $\boldsymbol{z} \in \Delta(\boldsymbol{A})$ is a correlated equilibrium when for any player i and any actions b^i and c^i,

$$u^i(\boldsymbol{z}) \geq u^i(\boldsymbol{z}_{b^i \to c^i}). \tag{3.3}$$

As we will see later, when a game is repeated, a natural correlation device is the history of the sequence of play.

A Nash equilibrium is a correlated equilibrium. However, the set of correlated equilibria is simpler in the sense that Eq. (3.3) is a linear inequality on $\boldsymbol{z}$ seen as a vector of dimension $|\boldsymbol{A}|$. As a consequence, correlated equilibria are solutions of a set of linear inequalities and the set of correlated equilibria is a polytope. On the contrary, Eq. (3.2) is not linear in the $\sum_{i \in I} |A^i|$ coefficients of $\boldsymbol{y}$ and Nash equilibria are more challenging to compute.

An even larger set is that of coarse correlated equilibria which contains all correlated distributions such that no unilateral pure deviation is profitable.

Definition 3.12 (Coarse Correlated Equilibrium in (one-shot) Games)**.** A mixed distribution $\boldsymbol{z} \in \Delta(\boldsymbol{A})$ is a coarse correlated equilibrium when for any player i and any action b^i,

$$u^i(\boldsymbol{z}) \geq u^i\big(b^i, \boldsymbol{z}^{-i}\big). \tag{3.4}$$

where $\boldsymbol{z}^{-i}$ is the marginal distribution where player i's actions have been marginalized out, i.e.,

$$\forall \boldsymbol{a} \in \boldsymbol{A}, \boldsymbol{z}^{-i}\big(\boldsymbol{a}^{-i}\big) = \sum_{c^i \in A^i} \boldsymbol{z}\big(c^i, \boldsymbol{a}^{-i}\big).$$

As Eq. (3.4) is weaker than Eq. (3.3), which is itself weaker than Eq. (3.2), we have the following well-known sequence of inclusion

$$\text{Nash Equilibria} \subseteq \text{Correlated Equilibria} \subseteq \text{Coarse Correlated Equilibria}.$$

3.3 Procedures

Procedures are a refined class of strategies, defined in stochastic and repeated games, which can actually be computed. In this thesis, we study and define new procedures in stochastic games.

Definition 3.13 (Procedure)**.** We define procedures for player i as (computable) functions $\mathcal{H} \to \Delta(A^i)$ which, given a history, choose a (mixed) action for player i.

This definition is similar to that of strategies but must be taken with a computational perspective: in this thesis, procedures are specified in pseudocode and are implemented for simulation purposes. Therefore, attention is paid to the complexity of such procedures–usually it is constant or linearly dependent on the number of states and actions.

Example (Fictitious Play)**.** *Introduced by Brown (1951) and Robinson (1951), this is one of the oldest non-trivial procedures. It specifies that the player plays a best-response to the empirical past actions of other players. Therefore, it is optimal against players that use constant actions. See Section 4.2.1 for details.*

3.4 Running Examples

3.4.1 Power-Gen

We define an applied example inspired from the Power Unit Commitment[1] game of the *International Planning Competition*.

A network of power plants must meet a demand $d(\theta) \in \mathbb{N}$ determined by a changing temperature $\theta \in \{\theta_{\min}, \dots, \theta_{\max}\}$. The demand is expressed in discrete units. Each plant i is parameterized by a minimum m^i and maximum M^i number of produced power units, a cost c^i_{change} of changing production (proportional to the change) and the cost c^i_{unit} per power unit. A deterrent penalty C is incurred when the network of plants does not meet the demand.

Every plant i chooses at time n its production $a^i_n \in \{m^i, \dots, M^i\}$.

Income is proportional to power units used by customers, at a price r. Profit is equally shared, i.e., all players have the same reward function, but decision-making is decentralized. Formally, the stochastic game is defined as follows.

Game 1 (Power Unit Commitment)**.**

- Players I are the plants
- For all plants i, $A^i = \{m^i, \dots, M^i\}$
- A state comprises the current temperature and last stage production levels $\boldsymbol{p}$, so states are $S = \{\theta_{\min}, \dots, \theta_{\max}\} \times \prod_{i \in I} A^i$
- The transition function between states is an unknown Markov chain for the temperature and production levels are determined by actions played at the previous step.
- The reward function is

$$u_{(\theta,\boldsymbol{p})}(\boldsymbol{a}) = \min\left\{d(\theta), \sum_{i \in I} a^i\right\} r - \left(\sum_{i \in I} c^i_{\text{change}} |p^i - a^i| + c^i_{\text{unit}} a^i\right) - 1_{\sum_{i \in I} a^i < d(\theta)} C$$

3.4.2 Bertrand Competition and Prisoner Dilemma

Bertrand competition describes a model with a number of firms which simultaneously give prices for a same good. Demand on this market then determines how many products every firm sells with the firm having the lowest price selling the largest quantity.

Rewards depend on the producing cost but importantly, they are neither zero-sum nor identical. The equilibrium is prices in perfect competition. However, if firms all pre-agree on a high price and somehow commit to it, then they may get higher payoffs.

Formally, Bertrand competition can be described as a (non-stochastic) game, it is detailed below with two firms, following Calvano et al. (2020) definitions.

Game 2 (Smoothed Bertrand Competition)**.**

- Players $I = \{1, 2\}$ are the two firms.
- Firms can choose discrete prices between $\underline{a}$ and $\overline{a}$, so $A^i = \{\underline{a}, \dots, \overline{a}\}$.
- Utilities are

$$u^i(\boldsymbol{a}) = (a^i - c^i) q^i(\boldsymbol{a}) \tag{3.5}$$

where c^i is the unit producing cost for firm i (which we generally suppose to be equal to 1)

[1] https://ataitler.github.io/IPPC2023/powergen.html

and $q^i(\boldsymbol{a})$ is the demand of goods produced by firm i, defined as

$$q^i(\boldsymbol{a}) = \frac{\exp\left(-\frac{a^i}{\mu}\right)}{\sum_{j\in I}\exp\left(-\frac{a^j}{\mu}\right)+1}.$$

It is well known that there exists a symmetric equilibrium, and we suppose that the corresponding price is between $\underline{a}$ and $\overline{a}$. Another well-known result is that there is a higher price that maximizes the sum of the rewards but which does not constitute a stable profile in the sense of Nash: both firms have incentives to drop their price so as they increase their market share. When firms have a mechanism to commit to this price, then there is a collusion and the market is inefficient.

The Prisoner Dilemma is a well-known game which can be seen as a simplified version of Smoothed Bertrand Competition with only two actions, $\underline{a}$ and $\overline{a}$, that we rename "cooperate" and "defection".

Game 3 (Prisoner Dilemma)**.**

- There are two players, the row player and the column player.
- The game is not stochastic.
- Utilities are defined by matrix

	C	D
C	$(3,3)$	$(0,4)$
D	$(4,0)$	$(1,1)$

There is a single Nash equilibrium, which is (D,D). In the context of repeated games, with infinite or bounded memory, we say that players *cooperate* when they play (C,C), although they both have an incentive to deviate.

Chapter 4

Background on Learning Procedures

In this chapter, we describe the literature related to learning in stochastic games, starting from the most general framework which is online learning (Section 4.1). Then, we review a first special case, that is the study of game dynamics in game theory (Section 4.2) and a second one which is reinforcement learning in computer science (Section 4.3). This chapter ends with a section detailing the theory of stochastic approximation (Section 4.4) which is particularly suitable for studying the dynamics of such systems.

4.1 Online Learning

In this section, we define a general setting, that of online learning, which encompasses both learning in repeated games and stateful learning. Various procedures of these two frameworks were an inspiration to design two procedures studied during my thesis (see Chapters 6 and 7).

An agent faces a problem said to be online when this problem takes place in a sequence of steps, usually indexed by $\mathbb{N}$ and all information is not available in the beginning. Instead, it is given as feedbacks after every action taken in every step. More specifically, online optimization is a setting when the agent endeavors to find a minimum point of either an unknown loss function (and information is, for instance, gradients of this function on specific points), or the cumulated loss of a sequence of functions.

Formally, it is a repeated decision problem where an action a_n must be selected at every step n. Feedback $g_n(a_n) \in F$ is then provided to the agent and a reward (or loss) $f_n(a_n) \in \mathbb{R}$ is attributed to the agent. For the sake of consistency with game theory, we formulate online optimization using utilities instead of loss, leading to the formalism of Model 2 inspired from Shalev-Shwartz (2011).

Model 2 Online Optimization

for $n = 1, \dots$ **do**
  choose an action $a_n \in A$
  observe $g_n(a_n) \in F$
  get $f_n(a_n) \in \mathbb{R}$ (usually observed)
end for

Multiple refinements of Model 2 have been proposed and there exists a gigantic literature on online learning. The next two sections of this chapter expose the state of the art in learning procedures in two special cases: *learning in repeated games* and *multiagent reinforcement learning* that are respectively detailed in Section 4.2 and Section 4.3. Before going into these details, we start with other well-known special cases of Model 2.

Multi-Armed Bandits If there is no additional feedback to the reward and rewards are randomly, independently selected according to a distribution depending only on a_n, then this problem is known as *multi-armed bandits* originally introduced by Robbins (1952), typically with a fixed action set.

Online Optimization with Gradient Feedback A classical case for the feedback is to be the gradient at points a_n of a fixed function. Then, usual gradient descent (in this case ascent) can be used to compute an optimum.

Uncoupled Learning in Games If the agent is evolving in an environment where its payoff is partially shaped by other agents, then this is a game. Furthermore, if we suppose that the agent is aware of its payoff function and other player actions (this is the feedback at every step) but not of other player payoff functions, then this is the *uncoupled* learning setting in games, as defined by Hart and Mas-Colell (2003a).

Convex Optimization Similarly to Shalev-Shwartz (2011), the online convex optimization is recovered when g_n is a concave function and $f_n = g_n$, the action set being a convex set. In particular this applies to functions of the form $g_n = \langle b_n, \cdot \rangle$ with b_n a vector.

Example (Follow the Leader)**.** *In the online convex optimization setting where the complete function f_n is observed, the well-established Follow the Leader (FTL) procedure specifies that the action taken must maximize the cumulated past functions, that is*

$$a_n \in \arg\max_{b \in A} \sum_{k=1}^{n} f_k(b). \tag{FTL}$$

A simple extension of such a procedure is Follow the Regularized Leader (FTRL), which given a regularizer function $h : \Delta(A) \to \mathbb{R}$, is defined as

$$a_{n+1} \sim \arg\max_{y \in \Delta(A)} \sum_{k=1}^{n} f_k(y) - h(y). \tag{FTRL}$$

The regularizer is typically a smooth, strictly convex function which is steep on the boundary of the simplex, therefore imposing the $\arg\max$ *to be a completely mixed action and a_n to be chosen randomly. Because of this randomness, FTRL is more robust to adversarial play than FTL. Such an adversarial play can be framed as a game and is the focus of the next section.*

4.2 Learning in Repeated Games

As explained by Mertikopoulos (2019), online optimization has deep links with learning in games. Procedures are called learning procedures when they are designed to maximize the reward of players based on previous steps in the context of repeated games.

The modern motivation of such procedures in computer science is control: based on the past and on a model of the environment, a player (*i.e.,* a device with some autonomy) must choose an action adapted to its goals. Therefore, a central question is the efficiency of such systems, in particular when there is little coordination possible between agents in a decentralized setting.

However, this is not the full historical justification of these procedures. Perhaps the most famous procedure, Fictitious Play, was originally designed as a way to *solve* the game, that is to find its equilibria (Erickson, 2015, p. 108). Indeed, game theory and in particular computations in game theory took off during the second world war to analyse military worth. Brown, having participated in military research during the war, had a contract with the RAND Corporation when he published its original paper on Fictitious Play and his goal was the computation of equilibria of specific games. This is in stark contrart with contemporary questions in game theory and computer science that aim to design or understand systems that learn.While Fictitious Play was originally specified in a turn-based style (Berger, 2007), it is nowadays usually defined in a setting where both players play at the same time, without knowing what other players are doing for the same step.

Formally, procedures are defined in the setting of Model 3: time is indexed by $n \in \mathbb{N}$ and at every step n, all players have to choose an action, observe other player actions and get the corresponding payoff.

The rest of this subsection is as follows: Section 4.2.1 introduces Fictitious Play, which is the basis of some of the procedures defined in this thesis, Section 4.2.2 explains which properties are studied in systems where players use learning procedures, and Sections 4.2.3 and 4.2.4 review other procedures in the adversarial and non-adversarial cases.

Model 3 Learning in a Repeated Game

for $n = 1, \dots$ **do**
  for all $i \in I$ **do**
    i chooses an action $a_n^i \in A^i$
  end for
  for all $i \in I$ **do**
    i observes $\boldsymbol{a}_n^{-i}$
    i gets $u^i(\boldsymbol{a}_n)$
  end for
end for

4.2.1 Fictitious Play

One of the oldest procedures is Fictitious Play (FP) introduced for zero-sum repeated games. It specifies that every player plays a best response to a prior (a mixed strategy profile) which is equal to the empirical past actions of the opponents. It was initially proposed by Brown (1951) and Robinson (1951) who proved that when the stage game is a zero-sum game and both players use FP, the empirical distribution of actions converges to the set of Nash equilibria of the stage game. We consider in the following the widespread variation of the original FP, where players update their beliefs and play simultaneously (Berger, 2007).

Definition Formally, the procedure defined for player i uses the empirical past actions of every other player j at time n, defined as

$$x_n^j := \frac{1}{n} \sum_{k=1}^{n} a_k^j. \tag{4.1}$$

where actions are embedded in a vector space, i.e., in the simplex $\Delta(A^i)$. Then, at every time n, player i chooses a best-response action

$$a_{n+1}^i \in \mathrm{BR}^i\left(\boldsymbol{x}_n^{-i}\right) := \arg\max_{b^i \in A^i} u^i\left(b^i, \boldsymbol{x}_n^{-i}\right). \tag{FP}$$

If there are multiple best-responses, the tie-breaking rule is usually left unspecified, and convergence theorems apply for all rules. Moreover, written as in Eq. (4.1), $\boldsymbol{x}_n$ can only be expressed in the basis of actions with rational coefficients, so the space of games where multiple best-responses are possible has a measure of zero.

Notice that when there are only two players, then this is a rewriting of Follow the Leader (FTL) in the context of repeated games: indeed, the normalization by $\frac{1}{n}$ in Eq. (4.1) does not change the value of the best-response. Therefore, in repeated linear two-person games, the action selection of FP is the same as FTL. When there are more than two players, then actions taken by FTL is a best-response against the joint empirical probability distribution of past actions, while actions taken by FP are best-response against the product of margin distributions of past actions for every player.

Convergence When all players use the FP procedure and the stage game (i.e., the static game which is repeated) is zero-sum, then the joint belief converges to the set of Nash equilibria. This was the original result of Robinson and was later extended to several classes of games, for instance potential games (Monderer and Shapley, 1996a) or $2 \times n$ games (Berger, 2005).

However, FP do not converge to Nash equilibria in all finite games, for instance in a 3×3 game detailed by Shapley (1964). This is consistent with general impossibility results, see below and Hart and Mas-Colell (2003b), Hofbauer and Sigmund (1998), and DeMichelis and Germano (2000).

Extensions FP is mostly deterministic (except in cases where there is a tie for the best-response), which implies that the procedure cannot be efficient against an opponent which knows that it is employed. This motivates a well-known extension, namely Smooth Fictitious Play (SFP) (also called Stochastic Fictitious Play) which is defined using a smooth best-response instead of the classical best-response (Fudenberg and Levine, 1995; Fudenberg and Levine, 1998). Technically, smooth best-response are defined with

regularizer functions (continuous, concave functions with infinite gradient on the boundary) h^i and a smoothness parameter η as

$$\mathrm{SBR}^i(\boldsymbol{x}^{-i}) := \arg\max_{y^i \in A^i} u^i(y^i, \boldsymbol{x}^{-i}) + \eta h^i(y^i, \boldsymbol{x}^{-i}). \tag{4.2}$$

Assumptions on h^i imply that the smooth best-response is single-valued but not a pure action and as such can be seen as a distribution over pure actions. Therefore, a player using SFP has to draw a pure action according to this distribution.

If all players use SFP, then the beliefs converge towards a regularized Nash equilibrium in zero-sum and identical-interest games (Hofbauer, 2001; Hofbauer and Hopkins, 2005) and the procedure has no η-regret, meaning that it is better than any constant action chosen in hindsight up to a factor η (see Section 4.2.3 below). Benaïm and Faure (2013) proposed another procedure called Vanishingly-Smooth Fictitious Play, where the smoothness parameter goes to 0 and which has no regret.

SFP is similar to FTRL but for an additional normalization: in FTRL, if the regularizer does not change with time, then it becomes progressively negligible (but still is sufficient to avoid ties for instance), whereas in the context of (non-vanishingly) SFP, there is a minimum (strictly greater than 0) probability for an action to be played.

Marden et al. (2009) proposed a variation of fictitious play with correlation, i.e., empirical strategies are considered jointly and not as a product of marginals. Of course, this difference is only relevant when there are more than two players–and a part of the literature is dedicated to two-player games and in particular zero-sum games.

4.2.2 Systems with Learning Procedures

The definition of every player behavior leads to a dynamical system whose properties (and in particular convergence) can be studied. A natural question is the degree of specification that we suppose for every player. A first interesting case is the one where all players act according to a predefined rule, which can be the same. In the case of FP, this setting is the one in which most convergence results have been established. This is a solution concept on its own for repeated games, especially if the procedure is supposed to imitate the behavior of real-world players. As such, understanding the limiting behavior of these systems contributes to the debate of what is an appropriate solution concept for stage games (Vega-Redondo, 2003, pp. 36-37).

An interesting result was proved by Hart and Mas-Colell for the so-called *uncoupled dynamics*, meaning systems where the player behaviors do not depend on the utility functions of other players (but may depend on its own utility function). In this repeated game setting, it is shown that it is not possible for such dynamics to converge to Nash equilibria in all classes of games. Therefore, it is necessary to either weaken the uncoupling hypothesis (which would not be natural as the utility function of a particular player is supposed to reflect its preferences) or restrict studies to particular classes of games or convergence to bigger sets. The convergence of FP in various classes of games (see above) is an illustration of this constraint, and the convergence of some procedures to the set of correlated equilibria (Hart and Mas-Colell, 2000; Hart and Mas-Colell, 2013) shows that this is indeed specific to Nash equilibria.

On the other end of the spectrum lies the case where a single player uses a specific procedure, and others are unspecified. Then the stakes are the robustness of the procedures to adversarial players.

4.2.3 Adversarial Learning

Regret An interesting benchmark for learning procedures is whether they have a negligible *regret* or not, that is if a constant action could have been better in average than using the procedure.

Formally, the (external) regret vector $\boldsymbol{r}_n^i$ of player i is defined as

$$r_n^i(b^i) = \sum_{k=0}^{n} u^i(b^i, \boldsymbol{a}_k^{-i}) - u^i(\boldsymbol{a}_k) \tag{4.3}$$

and the (external) regret is the norm of this vector

$$r_n^i = \max_{b^i \in A^i} \left(\sum_{k=0}^{n} u^i(b^i, \boldsymbol{a}_k^{-i}) - u^i(\boldsymbol{a}_k) \right) = \max_{b^i \in A^i} r_n^i(b^i). \tag{4.4}$$

It was introduced by Hannan (1957).

A no-regret procedure is a procedure whose usage by a player i implies a regret negligible with respect to time, i.e., $r_n^i = o(n)$ for any sequence of actions taken by the other players. Informally, this implies that no constant action would have performed better knowing the future sequence of actions of other players. Therefore, the interpretation of regret is somewhat limited to the case where the environment, that is other players, does not react to changes in player i's actions.

Even sensible procedures can lead to regret. For instance, Fudenberg and Levine (1998) give an example similar to the following one with the (ϵ-)Matching-Pennies game where FP leads to regret[1].

Game 4 (Matching-Pennies)**.** Matching-Pennies is a well-known game with two players $1, 2$ and two actions per player $\top^1, \top^2, \bot^1, \bot^2$. Player 1 gets 1 if both players play the same action and -1 otherwise, while player 2 gets the opposite, leading to the following utility matrix:

$$\begin{bmatrix} 1,-1 & -1,1 \\ -1,1 & 1,-1 \end{bmatrix} \tag{Matching-Pennies}$$

We are going to use a perturbed version of the previous game with $\epsilon > 0$ so that there is no tie in FP:

$$\begin{bmatrix} 1,-1 & -1-\epsilon,1+\epsilon \\ -1,1 & 1,-1 \end{bmatrix} \tag{ϵ-Matching-Pennies}$$

Example (Fictitious Play has regret)**.** *We now study the following system, where players play repeatedly ϵ-Matching-Pennies with the following procedures:*

- *Player i uses FP.*
- *Player j predicts the action of player i and plays a best-response, that is if $u^i\big(a^i, x_n^j\big) > u^i\big(b^i, x_n^j\big)$, then j plays b^j (a best-response to a^i), otherwise player j plays a^j.*

In this setting, utility of either action of player i with respect to empirical actions of j is

$$u^i\big(a^i, x_n^j\big) = x_n^j(a^j) - (1+\epsilon)x_n^j(b^j) = (2+\epsilon)x_n^j(a^j) - (1+\epsilon) \tag{4.5}$$

and

$$u^i\big(b^i, x_n^j\big) = -x_n^j(a^j) + x_n^j(b^j) = -2x_n^j(a^j) + 1. \tag{4.6}$$

Therefore, using Eq. (4.5) *and Eq.* (4.6)*, we deduce that $u^i\big(a^i, x_n^j\big) > u^i\big(b^i, x_n^j\big)$ if and only if*

$$(4+\epsilon)x_n^j(a^j) > 2+\epsilon \tag{4.7}$$

Then it is straightforward to see that if ϵ is a fraction of $\sqrt{2}$, then inequality of Eq. (4.7) *can not be an equality because $x_n^j(a^j)$ is necessarily a rational number. Therefore, we have either $u^i\big(a^i, x_n^j\big) > u^i\big(b^i, x_n^j\big)$ or $u^i\big(a^i, x_n^j\big) < u^i\big(b^i, x_n^j\big)$, so the prediction of player j always maximizes its payoff.*

As a consequence, the average reward of player j will be between 1 and $1+\epsilon$ and the reward of player i will be opposite. However, it is straightforward to see that $x_n^j(a^j)$ is going to converge towards $\tilde{x} := \frac{2+\epsilon}{4+\epsilon}$ since when $x_n^j(a^j) > \tilde{x}$, player i plays a^i and j plays b^j, so $x_n^j(a^j)$ decreases (and similarly in the other case, with decreasing steps).

Therefore, it is clear that, player i has regret since using Eq. (4.5)*, playing constantly a^i would yield, in the limit, an average reward of*

$$(2+\epsilon)\left(\frac{2+\epsilon}{4+\epsilon}\right) - (1+\epsilon) \tag{4.8}$$

which is close to 0 when ϵ is small, while the average reward using FP is close to -1.

[1]Their example uses Matching-Pennies but for the example to work, players must have priors or tie-breaking rules must be defined. We slightly change the game so that it works in the most basic setting of Fictitious Play.

Internal Regret Foster and Vohra (1997) refined the notion of regret with a new notion, that of *internal* regret with respect to two actions b^i and c^i aiming at measuring how the player would have fared if he had played c^i every time he had played b^i. Formally, it is defined as

$$R_n^i(b^i, c^i) = \sum_{k=0}^{n} 1_{a_k^i = b^i} u^i\left(c^i, \boldsymbol{a}_k^{-i}\right) - u^i(\boldsymbol{a}_k). \tag{4.9}$$

Foster and Vohra show that there exist procedures such that $R_n^i(b^i, c^i)$ is $o(n)$ for all b^i, c^i, called no-internal-regret procedures.

A consequence of Eq. (4.9) is that when all players use no-internal-regret procedures, then the empirical joint actions converge to the set of correlated equilibria (Hart and Mas-Colell, 2000; Foster and Vohra, 1997)[2]. When all players use no-(external)-regret procedures, then the distribution of joint actions converges to the set of (coarse) correlated equilibria.

Regret-Based Procedures Hart and Mas-Colell (2000) proposed a procedure called "regret-matching" where player i chooses the next action proportionally to past-regret of this action, formally

$$a_{n+1}^i = \begin{cases} b^i & \text{w.p. } \dfrac{1}{\mu} R_n^i(a_n^i, c^i) \\ a_n^i & \text{w.p. } 1 - \displaystyle\sum_{b^i \in A^i} \dfrac{1}{\mu} R_n^i(a_n^i, c^i) \end{cases} \tag{4.10}$$

This procedure is shown to have no internal regret. Hart and Mas-Colell explain that a modified version of their procedure using external regret would be *universally consistent*, which is the term used by Fudenberg and Levine (1998) to denote no-regret procedures. A variation of this procedure studied in the same paper is defined by an action selection proportional to the external regret and no bias toward the previously played action, leading to

$$a_{n+1}^i = b^i \text{ with probability proportional to } r_n^i\left(b^i\right) \tag{4.11}$$

This last procedure is extended to a class of strategies (Hart and Mas-Colell, 2001) where the probability of selecting an action is not necessarily proportional to the internal regret but still only depends on it. Formally, a Λ stationary regret-based strategy is defined as

$$a_{n+1}^i = b^i \text{ with probability proportional to } \Lambda(\boldsymbol{r}_n^i) \tag{4.12}$$

under the assumption that Λ is the differential of a continuously differential function P on $\mathbb{R}^{A^i}$ and that for all vector $\mathbf{x} \in \mathbb{R}^{A^i}$,

$$\Lambda(\mathbf{x}) \cdot \mathbf{x} > 0 \tag{4.13}$$

Then, procedure of Eq. (4.11) is a special case of Eq. (4.13) with $\Lambda(\mathbf{x}) = \mathbf{x}$ and $P(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. In this extended case, the procedure is shown to be no-regret.

Interestingly, Hart and Mas-Colell note that FP can also be written in the form of Eq. (4.13) with $P(\mathbf{x}) = \max_{b^i} x(b^i)$. However, it is not differentiable and as a consequence their theorem does not apply.

Approachability Introduced by Blackwell (1956a), approachability was originally presented as an analogous theorem to the minmax theorem (which is one of the first and most famous theorems of zero-sum games (Neumann, 1928)) for games with vector payoffs. This framework gives guarantees as of the possibility for a player to force its average reward vector to reach a given set. To link this theory with regret, it is possible to define a game with vector payoffs that is exactly the regret of every action. Then, finding a no-regret procedure is finding a procedure that forces the vector payoff to approach set $\mathbb{R}_-^{|A^i|}$. Links between regret and approachability were first noticed by Blackwell (1956b) and described by multiple authors since then, including Perchet (2014).

[2]Hart and Mas-Colell publication is posterior but Foster and Vohra actually cite an unpublished manuscript of Hart and Mas-Colell.

Calibration Calibration was introduced by Foster and Vohra (1997) and Foster (1999) and led to multiple articles, see Olszewski (2015) for a recent survey. The central idea of calibration is the design of *forecast functions* that are accurate in the sense that they reflect in average the sequence of play of the opponent.

Formally, a forecast is a function that associates a mixed action of the other player to every history, leading to a sequence of forecast f_n. To ease the exposition, we suppose here that the set of forecast is finite, so a forecast procedure is a (computable) function $\mathcal{H} \to X$ where $X \subseteq \Delta(A^2)$ is finite.

At time n, player 1 forecasts that player 2 will play f_n and can use this information to choose a best response of the forecast. Therefore, a typical set for X is the set of distributions y^2 such that $\forall b^2, y^2(b^2) = \frac{l}{m}$ with $l, m \in \mathbb{N}^\star$ and bounded.

A forecast sequence is said to be calibrated with respect to the sequences of actions played by player 2 if for all b^2,

$$\lim_{n\to\infty} \sum_{y^2 \in X} |\rho(y^2, b^2, n) - y^2(b^2)| \frac{N(y^2, n)}{n} = 0 \tag{4.14}$$

where

$$N(y^2, n) = |\{k \leq n+1 \mid f_k = y^2\}|$$
$$\text{and } \rho(y^2, b^2, t) = \frac{|\{k \leq n \mid f_k = y^i \wedge a_n^2 = b^2\}|}{N(y^2, n)}.$$

Intuitively, Eq. (4.14) holds when the forecast y^2 is on average correct on the times it is made. For instance, for an action $\top$ of player 2, it must happen 25% of times on all times that a 25% forecast is made on action $\top$, i.e., conditionally on $f_n(\top) = 25\%$, the empirical action must be 25% $\top$.

Then, Foster and Vohra (1997) and Foster (1999) prove that calibrated learning rules make the average play converge to the set of calibrated equilibria, and moreover that there exist calibrated learning rules that are calibrated for a player in spite of any behavior of other players.

Multiple extensions of calibration have been introduced, including (Kakade and Foster, 2008) where the play converges to the set of convex combination of Nash equilibria.

Hypothesis Testing Another framework used to do learning in repeated games (with, to the best of our knowledge, few practical implementations) is *hypothesis testing*, introduced by Foster and Young (2003). A player supposes that its opponents in a repeated game has a bounded memory of length m and formulates hypothesis on this player behavior. When the hypothesis is rejected (in the classical statistical sense), then a new one is tested and so on. It is shown that when all players use such a mechanism, then they employ strategies that are close to a subgame perfect equilibrium *in the repeated game* with bounded memory (at least ϵ-close to an ϵ-equilibria during a proportion $1 - \epsilon$ of the time).

4.2.4 Non-Adversarial Dynamics

Another line of work in procedures to play repeated games is focused on the behavior of the systems and not that of particular players. For instance, in (Young, 1993), the author explains how simple (and not designed to be optimal) learning rules can direct the system towards socially accepted convention. The author even insists in the conclusion that agents do not learn:

> *Society can "learn" even when its members do not.*

Formally, the model is designed to explain social conventions and Young assumes that there is a large population of agents and that at every step, m players are chosen randomly from a partition $C_1, \ldots, C_m$. Then, these agents play a game using a procedure called adaptive play.

However, it looks like this is not important from a mathematical point of view and we can look at the analysis assuming that there are n agents using the same procedure, the so-called adaptive play.

Then, if the game is weakly acyclic (from every state, i.e., history of length m, there is a path in the graph of best responses to a pure Nash equilibrium), then this procedure leads almost surely to the set of Nash equilibria. A theory of stochastically stable state is then developed to ensure that even in the presence of randomization, so-called stochastically stable states are most often played.

Other authors have proposed a variety of algorithms, including Arieli and Babichenko (2012) who showed that their procedure spends most of its time close to the Pareto frontier.

Algorithms with Mixed Objectives Then, some algorithms have mixed guarantees, for instance Conitzer and Sandholm (2007) proposed an algorithm that is both efficient against stationary opponents and converges to an equilibrium in self-play (when all players use the same procedure). This is possible by supposing that players have a mechanism to agree on a common equilibrium (circumventing the whole theory of equilibrium selection (Harsanyi and Selten, 1988)).

Learning in Time-Varying Games Another concept between repeated and stochastic games is *time-varying* games (Besbes et al., 2019b; Besbes et al., 2015a; Duvocelle et al., 2022a; Anagnostides et al., 2023a) which are repeated games with a changing matrix. The difference with stochastic states is that the changing budget is not time autonomous and is typically of the order of n^r with $r < 1$, meaning that compared to time, the changing bugdet grows slower. Therefore, games are evolving more and more slowly. This is used to derive regret-bounds that depends on r. On the contrary, in stochastic games, the probability to go from one state to another given an action profile played is constant.

4.3 Reinforcement Learning

Reinforcement Learning denotes a general setting, unfolding during a (possibly infinite) number of steps $n \in \mathbb{N}$, during which an agent has to take an action from set A, the system is in a state $s_n \in S$ and the agent gets a reward at each step. The reward and the new state are at least partially depending on action a_n^i, therefore the agent may face a trade-off between optimizing instantaneous reward and long term reward. This setting is different from other forms of machine learning in at least three ways. First, there is no supervision in the sense that the reward that the agent obtains does not depict complete information, in particular there is no a priori correct action. Second, it is online: as above all information is not available in the beginning, but additionally the performance of the current run of the algorithm is important. Third, it supposes that the problem is rather general, i.e., the goal is to design procedures that do not depend on a particular state structure, however it is commonly assumed that the state space is finite or at least countable.

Reinforcement learning (RL) took off in the eighties, with the well-know algorithm introduced by Watkins (1989) (see Kaelbling et al. (1996) for an early survey) to the impressive, unsupervised contemporary algorithm AlphaGo (Silver et al., 2017). Sutton and Barto (2018) give a more recent overview of the field.

Model 4 Reinforcement Learning

start at $s_1 \in S$
for $n = 1, \dots$ **do**
  choose an action $a_n \in A$
  get $U_n \in \mathbb{R}$
  observe $s_{n+1} \in S$
end for

If the environment is stationary, i.e., $\forall n, U_n = u_{s_n}(a_n)$ and there exists $\forall s, P_s : A \to \Delta(S)$ such that $s_{n+1} \sim P_{s_n}(a_n)$, then the system is a Markov Decision Process (MDP).

Values U_n may also be noisy rewards, due to incomplete information about the reward or the state. A typical setting for U_n is to be equal to $u_{s_n}(a_n) + R_n$ where R_n is some independent Gaussian noise.

Q-learning Watkins (1989) introduced Q-learning to solve Markov Decision Process with provable guarantees when the environment is stationary. The central idea is to estimate the so-called Q-values for every state-action pairs that represents the continuation payoff using the following scheme:

$$Q(s_n, a_n) \leftarrow Q(s_n, a_n) + \alpha_n(U_n + \delta \max_{b \in A} Q(s_{n+1}, b) - Q(s_{n+1}, a_n)) \tag{4.15}$$

with Q-values for other state-action pairs are left unchanged. This simple rule does not need any information about the transitions: they are implicitly sampled since the expectancy of s_{n+1} is $P_{s_n}(a_n)$ in the MDP case.

In this procedure, Q-values are updated towards a target value $U_n + \delta \max_{b \in A} Q(s_{n+1}, b)$. This is a special case of the general reinforcement learning procedure TD(λ) (Sutton, 1988) where learned continuations are updated towards a λ-discounted sum of rewards and expected continuations (so, in the case of Q-learning, $\lambda = 0$). Tesauro (1992) famously applied TD(λ) to learn weights of a neural network in the backgammon game in self-play.

Since the introduction of Q-learning, numerous extensions have been introduced, for instance double Q-learning (Hasselt (2010)) or Q-learning with deep learning (Hasselt et al. (2016)). However, in multiagent systems, from the one player's point of view the environment is not stationary because it comprises other players that are also using some non-stationary learning rule. Therefore, there is no guarantee and indeed it may not converge to the set of Nash equilibria (Wunder et al., 2010; Calvano et al., 2020). Chapter 6 explains how to overcome this difficulty with a different updating rule for estimating the continuation payoffs. This updating rule is closer to Expected Sarsa (Sutton and Barto, 2018) where the continuation values are moved towards the *expectation* of future payoffs.

Watkins proved that under standard stochastic approximation hypothesis

$$\begin{aligned} &\forall n, 0 < \alpha_n \leq 1, \\ &\sum_{k}^{n} \alpha_k = \infty, \\ \text{and } &\sum_{k}^{n} \alpha_k^2 < \infty \end{aligned}$$

and if U_n had an expectancy depending only on the state and on the action taken, then Q-learning converges towards the optimal strategy. These results were extended in various directions but the assumption that the environment is stationary remains crucial.

However, when there are several players, from the perspective of a single agent, the environment may not seem stationary as other agents adapt their strategies. Therefore, this raises the important question of *Multiagent* Reinforcement Learning.

4.3.1 Multiagent Reinforcement Learning (MARL)

Combining the game theory framework and stateful reinforcement learning leads to Model 5, which is called Multiagent Reinforcement Learning (MARL). We are going to study procedures that are globally efficient and converge to equilibria of the stochastic game.

Remark. *Another line of work, including algorithms such as COMA (Foerster et al. (2018)) are designed in order to lead to local optima. This is of particular interest when the state or action space can not be explored sufficiently, for instance when it is high-dimensional or when iterations are costly. However, in this thesis we are focused on* global *optima.*

Model 5 Learning in a Stochastic Game

an initial state s_1 is chosen
for $n = 1, \ldots$ **do**
  for all $i \in I$ **do**
    i chooses an action $a_n^i \in A^i$
  end for
  s_{n+1} is drawn according to $P_{s_n}(\boldsymbol{a}_n)$
  for all $i \in I$ **do**
    i observes s_{n+1} and $\boldsymbol{a}_n^{-i}$
    i gets $u_{s_n}^i(\boldsymbol{a}_n)$
  end for
end for

This setting was extensively studied by computer scientists, usually without knowing model parameters a priori. An interesting reflection of Shoham et al. (2007) in the paper "If Multi-Agent Learning Is the

Answer, What Is the Question?" puts into perspective this field and emphasizes that there are different research agendas for MARL, some inspired from control theory to others dedicated to contemplating the system dynamics. Busoniu et al. (2008) give an extensive survey of MARL.

Below, we describe other extensions of Q-learning to multiagent systems.

Minimax Q-learning Littman (1994) generalized Q-learning to zero-sum stochastic games. The $\max$ operator of Eq. (4.15) becomes a $\min\max$ operator which can be solved via linear programming. Even though this adds complexity to the algorithm, it takes advantages of the game structure and in particular of the opposite interests of the two players in zero-sum stochastic games.

Nash Q-learning Hu and Wellman (2003) proposed another way to learn Nash equilibria based on Q-learning, supposing that players can solve the auxiliary game parameterized by the Q-values and have access to a Nash equilibrium of the auxiliary game parameterized by the Q-values. Similarly to minimax Q-learning, this implies that a potentially costly computation is done at every step of the loop.

Q-learning in Repeated games When there is only one state, this is the classical framework in which the players repeatedly play a fixed normal form game. Leslie and Collins (2005) study Q-learning in this context. With a Boltzmann action selection and an update rate depending on the probability to play a given action, that is to say

$$\alpha_n = \alpha \frac{1_{a^i = a_n}}{\mathbb{P}(a_n = a^i)}, \tag{4.16}$$

the authors show that a system where all players use joint Q-learning can be related to the smooth best-response dynamics. Therefore, convergence in classical classes of game can be derived.

Decentralized Q-learning Sayin et al. (2021) introduced a decentralized version of Q-learning using a two-timescale stochastic approximation: players estimate at the same time a Q-function and a set of values for every state (which corresponds to $\max_{a'} Q(s', a')$ of (4.15)).

WoLF Bowling and Veloso (2001) designed the WoLF algorithm that retains the key ingredients of Q-learning but have a variable learning rate α that depends on whether the algorithm is losing or winning. When it loses, the learning rate is high so as it adapts quickly, otherwise it is slow so that other players have time to adapt. This algorithm was proven to converge in multiplayer but non-stochastic games in simple settings (with restriction such as the number of actions). In the same line of work, Conitzer and Sandholm (2007) proposed an ad-hoc algorithm designed to either adapt to players that employ the same algorithm or are stationary.

Fictitious Play Algorithms in Stochastic Games Vrieze and Tijs (1982) studied fictitious play in the context of repeated games with a converging sequences of approximations of the payoff matrix. However, similarly to the original fictitious play of Brown and Robinson, it is designed as a way to compute the value of a zero-sum game, and it is not a behavioral strategy (i.e., there is no current state, the goal is to compute the minmax of a series of matrices).

More recently, Leslie et al. (2020) proposed a best-response dynamics in continuous time which converges to stationary Nash equilibria in zero-sum discounted stochastic games. Sayin et al. (2022b) introduced a fictitious play algorithm for zero-sum stochastic games in discrete time with a continuation updating rule close to Q-learning in a model-based or model-free setting: both players update a Q-table with an entry per state-action pair whereas our continuation payoff are only indexed by the state. Compared to algorithms introduced in Chapter 6, algorithm of Sayin et al. (2022b) does not estimate the model explicitly, which may or may not be an advantage depending on the context. Furthermore, their action selection is not smooth–the selected action must maximize the Q-values with an ϵ-greedy strategy in the model-free case.

Sayin et al. (2022a) recently introduced another Fictitious Play with provable convergence in identical-interest stochastic games with a single controller.

Convergence Results One of the challenges of MARL is the design of procedures that not-only satisfy some benchmarks (for instance no-regret) on average but also have guarantees on the convergence when played jointly. In particular, a class of procedures called "optimistic" were shown to have last-iterate convergence (Daskalakis and Panageas, 2020; Syrgkanis et al., 2015) properties. These procedures follow a general scheme (gradient descent/ascent for Syrgkanis et al., 2015 and multiplicative weight updates for Daskalakis and Panageas, 2020) but increase the weight of the latest feedback. Both papers study the case of non-stochastic games supposing that all players use this same procedure.

4.3.2 Collusion

In the context of pricing algorithms, several recent papers raise the interesting question of implicit collusion by algorithms, that is collusion without explicit agreements between sellers of a good. Calvano et al. (2020) conducted experiments with Q-learning in a Bertrand competition game. This was the starting point of multiple articles in theoretical economics studying the competition aspects of algorithmic pricing (Asker et al., 2021; Brown and MacKay, 2021). Assad et al. (2020) present evidence from real data from the German retail gasoline market, estimating at the same time when sellers start using algorithms (via structural changes in the pricing) and the impact of these pricing algorithms on the overall competition.

4.4 Stochastic Approximations

The theory of stochastic approximations has long been used to study asymptotic behavior of discrete-time systems using their continuous-time counterparts (Benaim et al., 2005; Konda and Borkar, 1999). In this framework, one typically assumes that there is a set-valued function $F : \mathbb{R}^K \rightrightarrows \mathbb{R}^K$, a sequence of decreasing, positive update steps $\{\gamma_n\} \in \mathbb{R}^{\mathbb{N}}$ and Y_{n+1} a noise difference random variable. Then the two following systems may be related:

$$\frac{dy}{dt} \in F(y) \tag{4.17}$$

$$y_{n+1} - y_n - \gamma_{n+1} Y_{n+1} \in \gamma_{n+1} F(y_n) \tag{4.18}$$

Update steps γ_n are supposed to decrease towards 0 at a moderate pace, in this thesis we suppose that

$$\sum_{n=0}^{\infty} \gamma_n = \infty \text{ and } \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

Note that Eqs. (4.17) and (4.18) are respectively differential *inclusion* and recurrence *inclusion*, meaning that the (point) derivative is not completely specified. Any function or sequence that satisfies such relations is then called a solution of this system. The existence of solutions of (4.17) is guaranteed as long as F is a Marchaud map (Benaim et al., 2005), defined as:

Definition 4.1 (Marchaud map)**.** $F : \mathbb{R}^K \rightrightarrows \mathbb{R}^K$ is a Marchaud map if:

(i) F is a closed set-valued map, i.e., $\{(y, z) \in \mathbb{R}^K \times \mathbb{R}^K \mid z \in F(y)\}$ is closed.

(ii) for all $y \in \mathbb{R}^K$, $F(y)$ is a non-empty, compact, convex subset of $\mathbb{R}^K$

(iii) there exists $c > 0$ such that $\sup_{y \in \mathbb{R}^K z \in F(y)} ||z|| \leq c(1 + ||y||)$

Stochastic approximations theorems typically assert that the limit set of a solution of (4.18) is *internally chain transitive* for differential inclusion (4.17), meaning that two points of the limit set must be linked by a number of chained solutions of (4.17):

Definition 4.2 (Internally chain transitive)**.** A set A is internally chain transitive (ICT) for a differential inclusion $\frac{dy}{dt} \in F(y)$ if it is compact and if for all $y, y' \in A$, $\epsilon > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $y_1, \dots y_n$ to the differential inclusion and real numbers $t_1, t_2, \dots, t_n$ greater than T such that:

(i) $\forall i \in \{1, \dots, n\}, y_i(s) \in A$ for $0 \leq s \leq t_i$

(ii) $\forall i \in \{1, \dots, n-1\}, \|y_i(t_i) - y_{i+1}(0)\| \leq \epsilon$

(iii) $\|y_1(0) - y\| \leq \epsilon$ and $\|y_n(t_n) - y'\| \leq \epsilon$

More precisely, Theorem 4.3 of (Benaim et al., 2005) states that limits sets of (4.18) are included in internally chain transitive (ICT) sets of (4.17).

Robbins and Monro (1951) and Kiefer and Wolfowitz (1952) originally introduced systems of the form of (4.18). Then, Kushner and Clark (1978) proposed a link with continuous time systems such as (4.17), establishing that behaviors were identical in the case of the existence of fixed points for (4.17). Benaim (1996) showed that this could be generalized to internally chain transitive sets. Importantly, Benaim et al. (2005) extended both systems to the case of multiple derivatives, i.e., differential inclusions. Borkar (1997) introduced systems with two timescales, that is there are two equations similar to (4.18) but with different update steps, one being slower than the other one.

In another direction, Konda and Borkar (1999) and Borkar (1998) generalized the framework of stochastic approximations Benaim et al. (2005) to an *asynchronous* setting, meaning that different parts of the vector y_n are updated at every step. Under an ergodicity condition, the discrete time system can be shown to be related to an asynchronous continuous time system with an update term strictly positive for every part of the vector. This is a mathematically convenient way to use the ergodicity hypothesis.

More recently, Perkins and Leslie (2012) simplified assumptions needed by Borkar's original work and combined this theory with that of stochastic approximations of differential inclusions by Benaim et al.

Chapter 5

Behavior of Q-Learning in Games with Bounded Recall

This chapter originates from a project that began with Pierre Boudart during his master internship. I then established the methodology and carried out experiments during the last year of my PhD.

This small chapter aims to show that even simple procedures lead to complex and hard to predict dynamics in multiagent systems. More precisely, we start by experimentally showing that even though Q-learning has been shown to converge in MDPs, joint Q-learning has a peculiar behavior in bounded memory repeated games. In the family of Prisoner Dilemma games, there may be cooperation, but it depends on the size of memory. Memory *increases* or *decreases* cooperation depending on the game under study, which is a surprising finding.

Experimental research in multiagent systems has recently gained momentum in economics, in particular with an article by Calvano et al. (2020). They show that in a simple Bertrand game, players using Q-learning algorithms are prone to cooperate, which, in economics may be equated to collusion. Therefore, an important question is that of convergence by chance or true convergence. Indeed, their system supposes that players explore only a finite number of times, so the system is deterministic starting from a particular time and may be trapped in an otherwise unstable attractor. Therefore, our second set of experiments aims to extend experiments of Calvano et al. (2020) and in particular to show that the exploration scheme of Q-learning does not change the results in terms of cooperation.

After a paragraph on cooperation, punishment schemes and the Folk Theorem, a first section describes the setting in which experiments are carried out. Then, results of experiments regarding memory and exploration scheme are presented.

Cooperation and Punishment We say that players *cooperate* when (i) they play an action profile whose social utility is higher than the best social utility of Nash equilibria and (ii) individual utilities in cooperative profiles are also higher than utilities in the Nash equilibria. For instance, in the Prisoner Dilemma game, the (C, C) profile is a cooperative profile because both players get 3, so the social utility is 6 while the unique Nash equilibrium is (D, D) with social utility 2. Moreover, individual utilities are 3 in (C, C) and 1 in (D, D).

However, a cooperative profile is non-sustainable in a one-shot game in the sense that at least one player has an incentive to deviate, since it is not an equilibrium. The well-known "Folk Theorem" (Fudenberg and Tirole, 1991, p. 150) states that action profiles may be sustainable in a repeated setting as long as their payoff is individually rational for every player and the discount factor is high enough (i.e., if they are patient enough). More precisely, the Folk Theorem indicates that there exist (non-stationary) equilibria in the repeated game where a punishment scheme makes it possible to sustain such action profiles. For instance, consider the strategy where for every history h,

$$\sigma_h^i = \begin{cases} C & \text{when all actions played by the other player are } C \\ D & \text{otherwise} \end{cases} \tag{5.1}$$

where player i plays D as soon as the other player has not played C in previous steps. Then, if both players use such a strategy, they start with C and play C forever. If we now suppose that player 1 deviates to a strategy ω^1, i.e., player 1 plays D while player 2 never played D. Then, for the rest of steps, player 2 is going to play D, so player 1's utility will be 4 on the step where it deviated and between 0 and 1 afterwards. Without loss of generality, we suppose that the deviation happens at the first step, and this leads to the following utility for player 1

$$u^1\big(\omega^1, \sigma^2\big) \leq 4 + \delta \cdot 1 + \delta^2 \cdot 1 + \ldots = 4 + \frac{\delta}{1-\delta} \tag{5.2}$$

whereas the utility before deviation is

$$u^1(\sigma) = 3 + \delta \cdot 3 + \delta^2 \cdot 3 + \ldots = \frac{3}{1-\delta}. \tag{5.3}$$

Equations (5.2) and (5.3) implies that the deviation is not profitable when

$$4 + \frac{\delta}{1-\delta} < \frac{3}{1-\delta} \iff 4 - 3\delta < 3 \iff \frac{1}{3} < \delta. \tag{5.4}$$

In other words, if the discount factor is high enough, it is possible to ensure cooperation with a joint punishment scheme.

It is worth noting that a similar strategy to that of Eq. (5.1) is the strategy where player i only does a punishment for a bounded number of steps, denoted by m. In this case, Eq. (5.2) is replaced by

$$u^1\big(\omega^1, \sigma^2\big) \leq 4 + \underbrace{\delta \cdot 1 + \ldots + \delta^m \cdot 1}_{\text{punishment}} + \delta^{m+1} \cdot 4 + \ldots = 4 + \delta\frac{1-\delta^m}{1-\delta} + \delta^{m+1}\frac{4}{1-\delta}. \tag{5.5}$$

Eq. (5.3) is valid, and the deviation is not profitable[1] as soon as

$$\begin{aligned} & 4 + \delta\frac{1-\delta^m}{1-\delta} + \delta^{m+1}\frac{4}{1-\delta} < \frac{3}{1-\delta} \\ \iff & 4(1-\delta) + \delta(1-\delta^m) + 4\delta^{m+1} < 3 \\ \iff & \frac{1}{3} < \delta - \delta^{m+1}. \end{aligned} \tag{5.6}$$

This latest equation can be rewritten as

$$m > \frac{\log\left(1 - \frac{1}{3\delta}\right)}{\log \delta} \tag{5.7}$$

as long as $\delta > \frac{1}{3}$. Equation (5.7) shows that when δ is high (player 1 grants much interest in future utilities), then m must be large, since the right hand side goes to ∞ when δ goes to 1. For $\delta = 0.65$, $m = 2$ is sufficient.

Other punishment schemes may require a larger memory size, as will be clear in the Non-Convex Prisoner Dilemma described below. More generally, the Folk Theorem ensures that with a high enough (but bounded) memory size and a high enough discount factor, any rational utility if the convex polytope is feasible with a punishment scheme.

5.1 Model

Joint Q-learning In these numerical experiments, all players use the same procedure, online Q-learning (Watkins, 1989), specified as

$$\begin{cases} Q^i_{n+1}(s, b^i) = \begin{cases} (1-\alpha)Q^i_n(s, b^i) + \alpha\left(u^i_s(\boldsymbol{a}_n) + \delta \arg\max\limits_{c^i \in A^i} Q^i_n(s_{n+1}, c^i)\right) & \text{if } s = s_n \text{ and } b^i = a^i_n \\ Q^i_n(s, b^i) & \text{otherwise} \end{cases} \\ a^i_n \sim x^i_n\big(Q^i_n(s_n)\big) \end{cases}$$

[1] This is a sufficient but not necessary condition, it depends on the specification of ω^1.

where $Q_n^i(s_n)$ is the vector $Q_n^i(s_n, \cdot)$, $x_n^i : \mathbb{R}^{A^i} \to \Delta(A^i)$ is the action selection function and $\alpha > 0$ is the constant update rate.

We informally assume that a lot of weight is given to actions with the highest Q-value, i.e., actions c^i that maximize $Q_n^i(s_n, c^i)$. Other actions are explored with a probability specified by $x_n^i\big(Q_n^i(s_n)\big)$ according to an exploration scheme, which might be

- η-greedy, meaning that
  - with probability $1 - \eta$, one of the best actions is randomly selected (i.e., an action in $\arg\max_{b^i} Q_n^i(s_n, b^i)$)
  - with probability η, an action is randomly selected among A^i
- β-exponential, similar to the previous scheme but the probability to explore exponentially decreases with time (this is the scheme used in (Calvano et al., 2020)), so
  - with probability $1 - \exp(-\beta n)$, one of the best actions is randomly selected
  - with probability $\exp(-\beta n)$, an action is randomly selected among A^i

  An important aspect of the β-exponential is that it is not stationary, the probability to explore is decreasing with time. Moreover, it is straightforward to establish using the Borel-Cantelli Lemma that only a finite number of exploration actually occurs. Therefore, even if this is the online version of Q-learning, it actually does not explore after an expectedly finite time.

Experimental comparison is carried out in Section 5.3 in order to investigate whether collusion evidenced by Calvano et al., 2020 is linked to a particular exploration scheme.

Measure and Confidence To evaluate whether players cooperate (or, from another point of view, whether the system fails to converge to a Nash equilibrium payoff), we compute the average profit with respect to the socially-best Nash equilibrium payoff and the socially-optimal payoff (which are, in particular, Pareto-optimal). Formally, for a game G, we define

$$SN(G) = \max_{\boldsymbol{y} \text{ is a Nash eq.}} \sum_{i \in I} u^i(\boldsymbol{y})$$
$$SO(G) = \max_{\boldsymbol{y} \in \Pi_{i \in I} A^i} \sum_{i \in I} u^i(\boldsymbol{y}).$$

Then, for a sequence of play, the profit is defined as

$$\textsc{Prof}_n = \frac{\frac{1}{n}\left(\sum_{k=1}^{n} \sum_{i \in I} u^i(\boldsymbol{a}_k)\right) - SN(G)}{SO(G) - SN(G)} \tag{5.8}$$

Example. *The Prisoner Dilemma game (seen previously) was defined with payoff matrix*

	C	D
C	$(3,3)$	$(0,4)$
D	$(4,0)$	$(1,1)$

Therefore, the unique Nash equilibrium is (D, D) and its social payoff is 2. Moreover, a Pareto profile is (C, C) and the best social payoff is 6, so

$$SN(\textit{Prisoner Dilemma}) = 2$$
$$SO(\textit{Prisoner Dilemma}) = 6$$

In the experiments that follow, we are looking to estimate $\mathbb{E}[\textsc{Prof}_n]$ for a large n, starting from a random state and random Q-values[2]. We do so by running several simulations and then estimating a

[2]We estimated the maximum/minimum Q-values possible and then chose randomly a matrix at every new simulations–resulting in different initial Q-values for the same set of parameters

confidence interval of our estimator via a bootstrap reverse percentile method (with, unless specified, at least 200 initial samples and bootstrap confidence interval at 95%).

The value of the horizon n is chosen so as results are already meaningful for $n/2$. For the first experiment, we ensure that profit has converged numerically (i.e., that temporal variations for a set of parameters are negligible compared to variations between sets of parameters) and that there is no overlap of confidence intervals for different memory sizes.

For the second experiment, with limited computing resources, a trade-off had to be made between the precision of our grid of parameters and the number of times each experiment was run.

While it is impossible to test every possible combinations of parameters, we have tried to check the robustness of our simulations with various combinations of δ, η, α, as specified below. From a computation time perspective, it is delicate to change more than two parameters at once with a lot of precision: in the second experiment, we experimented with different values[3] of α and with a parameter of the exploration scheme (which expresses how random is this exploration scheme). Therefore, while it could be interesting to experiment with additional values of δ, it appeared that two values of δ for the first experiment and a single value of δ for the second experiment already gave enough insight.

Games Simulations are run in Smoothed Bertrand Competition, Prisoner Dilemma and a variation of Prisoner Dilemma whose description follows. It has increased payoffs for the defecting player in (C, D) or (D, C), leading to the following definition.

Game 5 (Non-Convex Prisoner Dilemma)**.** This is defined similarly as Prisoner Dilemma but with matrix

	C	D
C	$(3, 3)$	$(0, 10)$
D	$(10, 0)$	$(1, 1)$

Figure 5.1 shows both game payoff functions. An interesting property of Non-Convex Prisoner Dilemma is that in the repeated game, (C, C) does not give the highest social payoffs. Indeed, a well-known property of repeated games is that feasible average payoffs can be any convex combination of vector payoffs. Therefore, if both players want to cooperate, they are better off with a sequence of (D, C) and (C, D) which may be encoded as a strategy with a memory of size 1.

However, following the Folk Theorem outlined above, a punishment needs an extra round to be encoded, so a (non-stationary) equilibrium strategy achieving a social utility of 5 needs at least a memory of size 2, depending on the discount factor.

Implementation Source code of our simulations is available in the supplementary material of this thesis: `https://www.lamsade.dauphine.fr/~lbaudin/manuscript/`. Most of the code is annotated with types, it is built using a framework for game theory developed during my PhD, numerical library `numpy` and scientific library `scipy` for the bootstrap method. Parts of the Python code is transpiled to C via cython.

5.2 Memory in the Prisoner Dilemma

This section is dedicated to the study of memory in the form of bounded recall in Prisoner Dilemma games (convex and non-convex). This was done before, for instance by Wunder et al. (2010) or Kianercy and Galstyan (2012), but to the best of our knowledge, this influence of memory was never tested.

Without memory, the state space is a singleton $\{s\}$, while it has 4 elements with a memory of size 1 and 4^m elements with a memory of size m. Therefore, if S_m denotes the state space of (Non-Convex)

[3]It is necessary for the horizon n, the update rate value α and the minimum number of visits to a state to be linked: the value of α must be of the order of magnitude of the maximum changes in Q-values (which is itself of the order of $\frac{\|u^i\|}{1-\delta}$) divided by the horizon n and the minimum number of visits to a state. This latest quantity is in most cases determined by η.

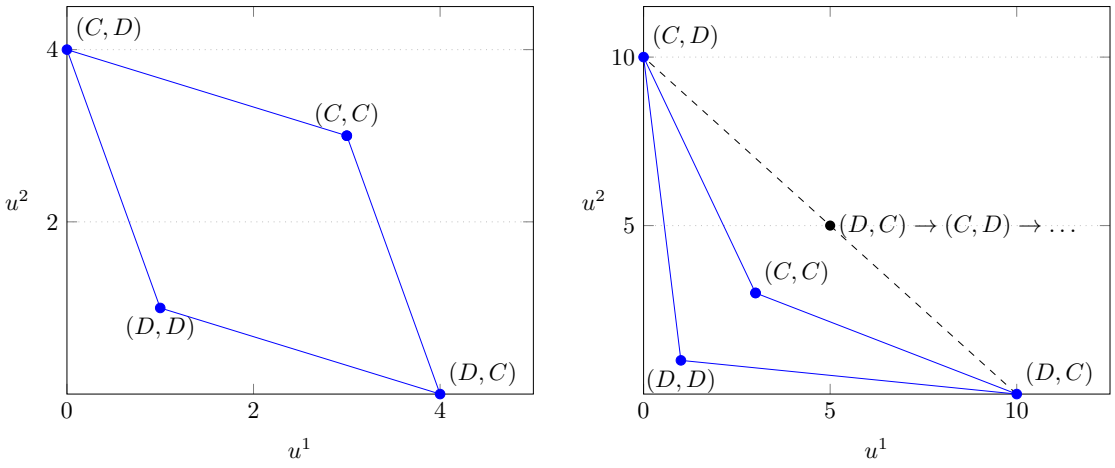


Figure 5.1: Payoff functions of Prisoner Dilemma (left) and Non-Convex Prisoner Dilemma (right)

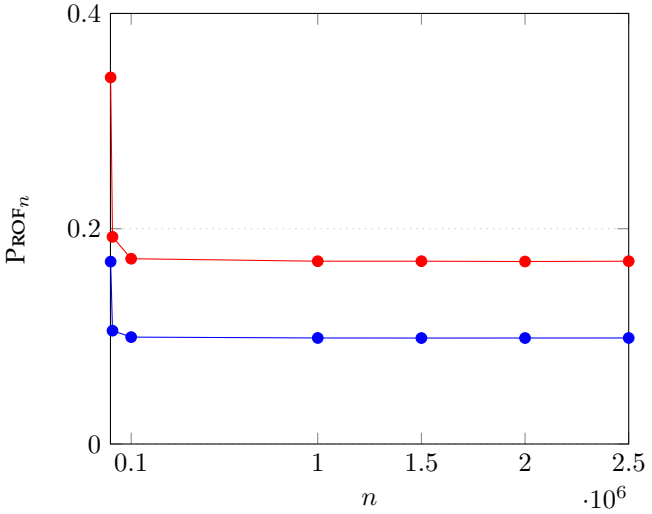


Figure 5.2: Profit for $\delta = 0.65$ (blue) and $\delta = 0.9$ (red) for several durations in Prisoner Dilemma without memory

Prisoner Dilemma repeated with memory of size m seen as a stochastic game, then

$$S_0 = \{s\}$$
$$S_1 = \{(C, C), (C, D), (D, C), (C, C)\}$$
$$S_m = S_1^m.$$

Parameters The following experiments use joint Q-learning with a varying memory size and an η-greedy exploration scheme, with learning parameter $\alpha = 0.1$ and $\eta = 0.1$.

Duration In order to compare experiments with different memory sizes, it turns out that $n = 2 \cdot 10^6$ is a sufficient duration. Indeed, results with different discount factors and durations are depicted on Fig. 5.2 for experiments without memory in both games. Profit values for $n = 10^6$, $n = 1.5 \cdot 10^6$ and $n = 2 \cdot 10^6$ are very close and confidence intervals do not intersect between experiments with different memory sizes. Different durations for different memory size are discussed at the end of this section.

Discount Factors As expected, the higher δ is, the higher the cooperation. In the following, results are shown for discount factors $\delta = 0.9$ and $\delta = 0.65$ but other discount factors (for instance 0.8) show the same behavior.

Cooperation Results Results of our experiments are shown in Fig. 5.3. The usual convex version of Prisoner Dilemma exhibits cooperation with $\delta = 0.9$. This is consistent with the work of Wunder et al. (2010), where they experiment with a variant of Q-learning called Infinitesimal Q-learning without memory.

Surprisingly, this does not hold for Non-Convex Prisoner Dilemma with the same discount factor even if the value of the (C, C) profile is the same. Without memory, the profit in the $\delta = 0.9$ case is estimated to be 0.17, which according to Eq. (5.8) means that a there is an average social payoff equal to 3.36, which is significantly lower than the average social payoff 5.4 in the standard Prisoner Dilemma case (still without memory). This suggests that the stronger incentive to deviate (in the Non-Convex Prisoner Dilemma case) impedes most cooperation without memory.

Regarding the role of memory : increasing memory improves profit in the non-convex case but not in the standard, convex case, as shown in Fig. 5.3. Note that the difference of behavior between these two games is not entirely suprising because the best social payoff are obtained with different strategies: in the non-convex case, to achieve a 0.8 profit, players have to alternate between (C, D) and (D, C), whereas to achieve a profit greater than 0.5 in the convex case, they have to stick to a (C, C) profile.

However, comparing simulations with different memory sizes may require different number of steps: as the number of states is equal to 2^k where k is the memory size, it is reasonable to multiply the duration by the size of the memory. For instance, a simulation with n steps without memory may be compared with a simulation of size $4n$ with memory 2. In addition to Fig. 5.3, we also ran the simulations with these additional durations. Although profit may be slightly different (which is not surprising as average values are slow to converge), profit is still decreasing when we increase the duration of the simulations by the same factor is the state increase (so for instance $n = 16 \cdot 10^6$, see the attached notebooks). Nonetheless, comparing the limiting average of such systems with different state space remains a challenge, especially when the state space grows exponentially with the size of the memory.

Actions Played In the non-convex case, in order to achieve a high cooperation measured as a $\textsc{Prof}_n$ value close to 1, players have to play a combination of profiles (C, D) and (D, C). In our simulations, we consistently observe that the profit is equally shared between players and that they frequently alternate, in the sense that, for instance on a run with a memory of size 3,

$$\begin{gathered}\mathbb{P}(a_n^1 = C \wedge a_{n+1}^1 = C) = 0.06 \\ \mathbb{P}(a_n^1 = D \wedge a_{n+1}^1 = C) + P(a_n^1 = C \wedge a_{n+1}^1 = D) = 0.76 \\ \mathbb{P}(a_n^1 = D \wedge a_{n+1}^1 = D) = 0.17.\end{gathered}$$

That is to say, player 1 spends most of its time alternating between actions D and C.

Furthermore, even if the punishment scheme is not deterministic, we notice that player 2 plays relatively more often D where player 1 plays two subsequent actions C than when player 1 plays a single C,

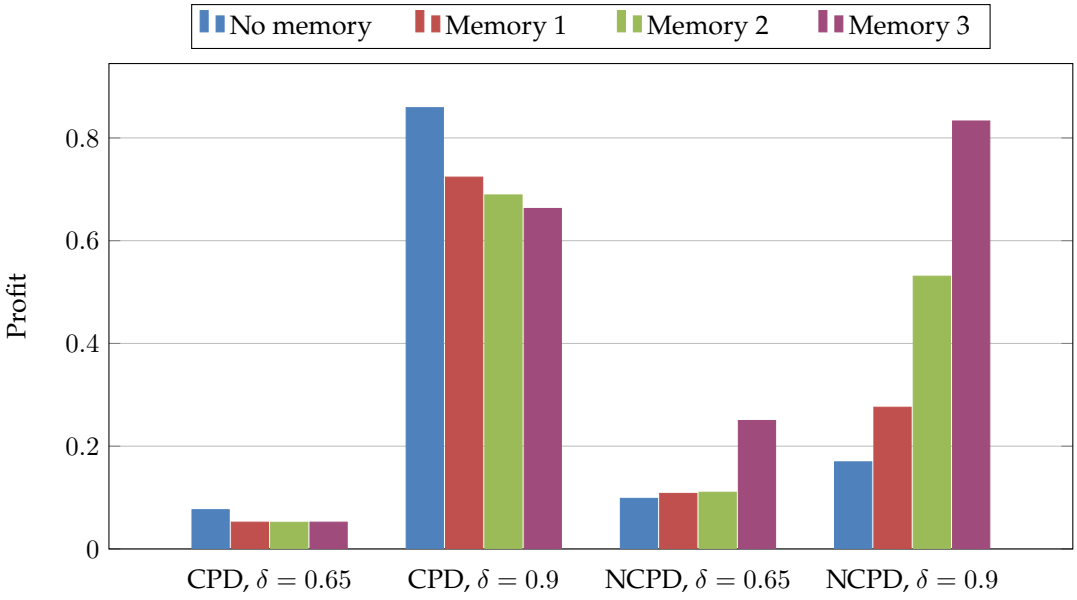
$$\mathbb{P}(a_{n+2}^2 = D | a_n^1 = C \wedge a_{n+1}^1 = C) = 0.67 \geq \mathbb{P}(a_{n+2}^2 = D | a_n^1 = C) = 0.81.$$

5.3 Exploration Scheme in Smoothed Bertrand Competition

A criticism that can be made against Calvano et al. (2020) is the exploration scheme they use. Indeed, they use a β-exponential scheme with the probability of random selection varying with time, following this definition

$$\eta_n = \eta_0 \exp(-\beta n) \tag{5.9}$$

The Borel-Cantelli Lemma implies that exploring with probabilities η_n is done only a finite number of times almost surely. However, this can be problematic for at least two reasons. First, in any realistic setting, players keep experimenting as time goes on because they need to react to a non-stationary



CPD stands for Prisoner Dilemma and NCPD for Non-Convex Prisoner Dilemma.

Figure 5.3: Average profit after $n = 2 \cdot 10^6$ steps

environment or because being stationary (and deterministic) makes them vulnerable to adversarial attacks. Second, experimentation also reflects some uncertainty of the system and simulations on a purely deterministic system may exhibit behavior that would be unstable with little deviations.

Simulations that follow are all executed with the same model as Calvano et al. (2020), that is a Smoothed Bertrand Competition game with a memory of size 1, so it is a stochastic game with state space $S = \boldsymbol{A}$.

β-exponential We start with simulations with the same exploration scheme as Calvano et al. (2020), which is β-exponential. However, we show results on average over the whole experimentation period, which is $4 \cdot 10^6$ steps. On the contrary, Calvano et al. stop simulations when there is no change of best-responses after a fixed number of steps, and compute the profit on these actions. This can not be compared with other exploration schemes, therefore we compute the average profit on a fixed horizon (and check that with twice this horizon we get comparable results in order to be confident on the fact that we have a good approximation of the limit behavior).

The two parameters of importance, as explained by Calvano et al. are α and β, therefore profit is computed for a grid of those parameters, leading to Fig. 5.4. These are averaged on 64 runs for each pair of parameters.

η-greedy The same experiment with an η-greedy exploration scheme gives the results of Fig. 5.5. Although some cooperation is evidenced, this is only with small values of both η and α. This may suggest that what we observe is a lack of convergence and not real cooperation. We did additional simulations with a higher number of steps (up to $20 \cdot 10^6$): the cooperation rate is stable and therefore this suggests that players do cooperate on average, albeit with extreme parameters.

These extreme parameters suggest that this may indicate that the system fails to converge and stays in rest points that are unstable in the continuous time dynamics. Therefore, what we observe could be these unstable rest points and are different from points identified for the β-exponential or smoothed action selection exploration scheme (see below).

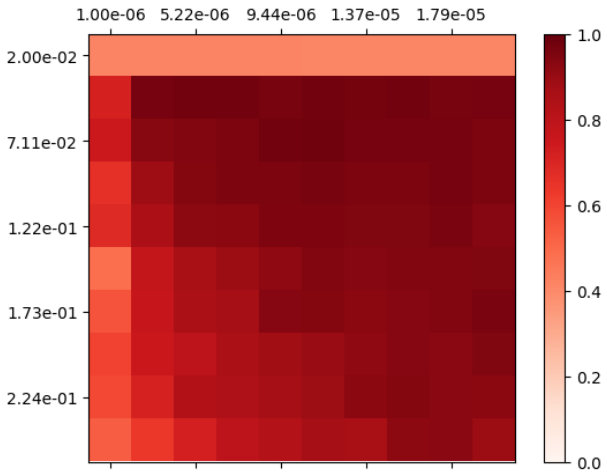


Figure 5.4: Average profit with a β-exponential exploration scheme on smoothed Bertrand competition with $n = 4 \cdot 10^6$
The x-coordinate is β and the y-coordinate is α.

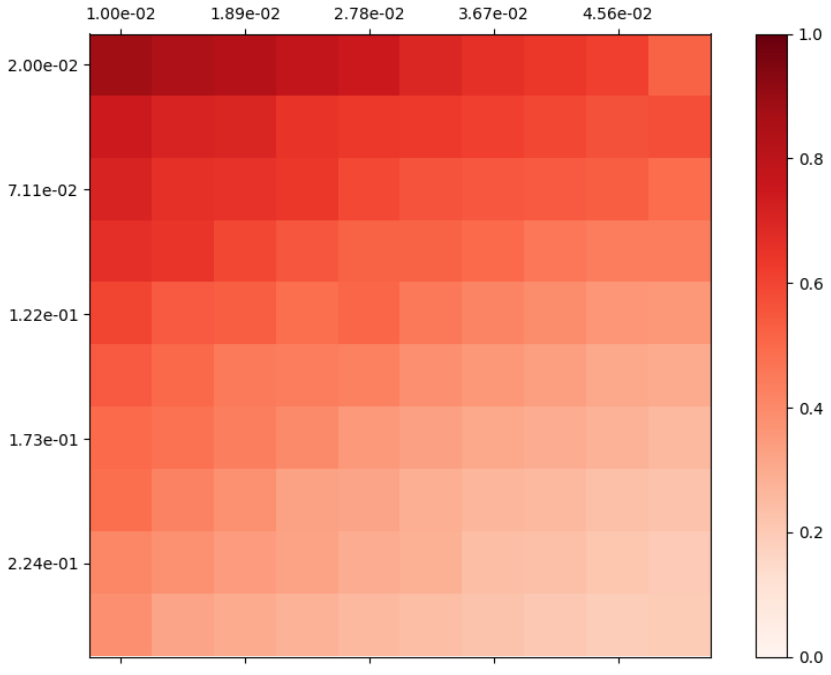


Figure 5.5: Average profit with an η-greedy exploration scheme on smoothed Bertrand competition with $n = 9 \cdot 10^6$
The x-coordinate is η and the y-coordinate is α.

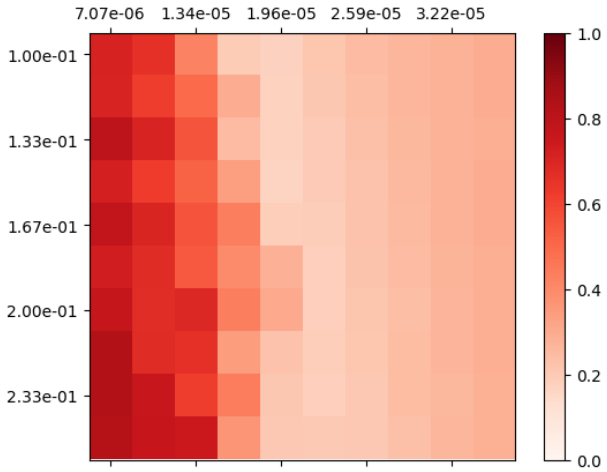


Figure 5.6: Average profit with a smooth exploration scheme with parameter $\eta = \beta\sqrt{n}$ on smoothed Bertrand competition with $n = 4 \cdot 10^6$
The x-coordinate is β and the y-coordinate is α.

Smooth Action Selection The third exploration scheme we try is a smooth one, i.e., actions are selected according to a distribution which is continuously dependent on the Q-values. It is defined as

$$a_n^i \sim_{b^i \in A^i} \frac{\exp\left(\frac{1}{\eta_n} Q(s_n, b^i)\right)}{\sum_{c^i \in A^i} \exp\left(\frac{1}{\eta_n} Q(s_n, c^i)\right)} \tag{5.10}$$

An interesting case is $\eta_n = \beta\sqrt{n}$ as it is similar to Vanishingly Smooth Fictitious Play (Benaïm and Faure, 2013) which is known to be non-regret. This is more sophisticated than constant exploration rate and β-exponential exploration rate as a probability to explore an action depends on its Q-values. Results are drawn in Fig. 5.6.

Contrary to other exploration schemes, an interesting result is that cooperation still holds with higher value of α such as $\alpha = 0.25$. The right part of the figures depicts the average profit when β is high, therefore players mostly play uniformly at random, resulting in an outcome better than the equilibria but without strategic coordination. When β decreases, the profit gradually decreases as players become strategic but they do not cooperate. A lower value of β creates the necessary conditions for cooperation, independently of the value of α.

5.4 Conclusion

Our experiments with Q-learning and different memory size in both Prisoner Dilemma games show that it exhibits cooperative behavior on both games. This cooperative behavior increases when players are more interested in future payoffs, i.e., when the discount factor is closer to 1. However, a less-intuitive result is the decrease in cooperation when the memory increases in the first Prisoner Dilemma game (where pure payoffs form a convex polytope). It is at its highest without memory and decreases with memory of size 1, 2 and 3. On the contrary, it is at its highest with a memory of size 3 in the Non-Convex Prisoner Dilemma game, where players have to alternate between the (C, D) and (D, C) action profiles in order to get the maximum average payoffs.

A limitation of our work is the comparison of systems with different state spaces: there is twice the number of states when memory has size 3 compared to a bounded-memory of size 2. Therefore, an interesting direction, which would increase the confidence in our results, is the study of the limiting behavior independently of a fixed horizon, similarly to what is done in (Calvano et al., 2020). However, our systems do not exhibit last-iterate limiting behavior reduced to a point (in other words, they oscillate between stationary strategies), so this is a challenging task.

Regarding the exploration schemes, we established that Q-learning induces a cooperative behavior with several exploration schemes, albeit for some schemes, extreme parameters are needed. Interestingly, the smooth action selection with vanishing exploration leads to cooperation for a substantial range of parameters and does not need a low value of α.

Establishing formal convergence properties in multiagent systems with such procedures is notably difficult (Sutton and Barto, 2018). In the next chapters, we define procedures inspired from Fictitious Play (FP), which compute empirical actions of other players, and we are going to investigate the formal convergence of such systems.

Chapter 6

Fictitious Play for Stochastic Games

Results of this chapter were published in the following article written with Rida Laraki:

Lucas Baudin and Rida Laraki. "Fictitious Play and Best-Response Dynamics in Identical-Interest and Zero-Sum Stochastic Games". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 17–23, 2022, pp. 1664–1690

In this chapter, we introduce an extension of Fictitious Play (FP) (defined for repeated games in Section 4.2.1) for stochastic games. Most proofs are postponed to Chapter 7 as FP is generalized to SFP for stochastic games.

Contrary to repeated games, a state variable changes during the sequence of play of a stochastic game, and as a consequence the utility function may also change. Furthermore, in addition to getting a high payoff, a player should take into consideration the future state. Therefore, it is not possible to use directly, in stochastic games, the original FP designed for repeated games. An interesting question is whether FP can be used in a repeated stochastic game, i.e., FP would recommand a stationary strategy, then the stochastic game is played on an infinite (or very large) horizon, and so on. Unfortunately, this is also not directly possible since the resulting game would have non-concave utility function and playing FP would bring no guarantee, and there would still be difficulties to handle the large horizon and motivate such a sequence of play.

Therefore, in this chapter, we propose an extension to FP with an additional set of continuation variables, inspired from Q-values of Q-learning. It is shown to converge when all players employ it in identical-interests and zero-sum stochastic games to the set of stationary Nash equilibria (or approximate equilibria in the zero-sum case).

6.1 Fictitious Play (FP)

Synchronicity In this chapter, we consider two different ways to play stochastic games, *asynchronous* and *synchronous* sequences of play. In asynchronous sequences of play, every player provides a single action per step, and there is a state variable which evolves with time. This is the traditional way to play stochastic games. By contrast, in synchronous sequences of play, there is no distinguished state variable and players have to provide an action for every state. Synchronous procedures may not be applied to online control problems but nevertheless, they can be interpreted in various ways: either as a setting where the actual state is not known to the players, as a pre-computation phase of the system or as an algorithm to compute equilibria, as in Vrieze and Tijs (1982). Asynchronous or synchronous procedures refer to procedures that can be used by players to play in each case.

Therefore, an asynchronous FP procedure, introduced in this section, is a behavioral strategy in the usual sense, that is a function that provides a distribution of probability for the action a_n^i given the history of the play prior to n and the current state s_n. Formally, it is a mapping $\bigcup_{n\in\mathbb{N}}[(S \times \boldsymbol{A})^n \times S] \to \Delta(A^i)$. In such an asynchronous extension of FP, there is a unique current state (that every player observes) and actions are chosen by the players only for this state, following Model 5 on page 55.

This contrasts with our other extensions of FP which are *synchronous*. They are not behavioral strategys because there is no specific current state and players provide actions for all states at every

Model 6 Synchronous Sequence of Play

for $n = 1, \dots$ **do**
  for $s \in S$ **do**
    for $i \in I$ **do**
      i chooses an action $a^i_{s,n} \in A^i$
    end for
  end for
  for $s \in S$ **do**
    for $i \in I$ **do**
      i observes $\boldsymbol{a}_n^{-i}$
      i gets $u^i_{s_n}(\boldsymbol{a}_n)$
    end for
  end for
end for

stage, i.e., it is a mapping $\bigcup_{n\in\mathbb{N}}(\boldsymbol{A}^S)^n \to \Delta(A^i)^S$. This is detailed in Model 6.

Technically, the synchronous procedure is also interesting to study because it converges in all identical interest stochastic games even the non-ergodic ones.

Sequence of play (asynchronous) An asynchronous play starts with a state s_0. At every step n, every player i chooses an action a^i_n, receives payoff $u^i_{s_n}(\boldsymbol{a}_n)$ and a new state s_{n+1} is drawn according to $P_{s_n}(\boldsymbol{a}_n)$.

Sequence of play (synchronous) However, in a synchronous play, there is no distinguished state and at every step, every player i chooses an action $a^i_{s,n}$ for every state $s \in S$ and it receives a vector of payoff $\left\{u^i_s(\boldsymbol{a}_{s,n})\right\}_{s\in S}$.

6.1.1 Asynchronous FP (AFP)

Our discrete-time procedures are designed using two estimates per state: one is the empirical action that every player uses, and the other one is the expected continuation payoff that a player estimates starting from this state. First, we define the two estimates, then proceed with a description of the action selection, and finally the updating rules.

Empirical actions We begin by exposing how the (asynchronous) empirical action is computed for every state. Given a state $s \in S$ and a time step n, $\mu_{s,n}$ denotes the number of times that s is the current state between 1 and n, that is

$$\mu_{s,n} = \sharp\{k \mid 1 \le k \le n \wedge s_k = s\}. \tag{6.1}$$

Then the empirical action of player i in state s is defined in $\Delta(A^i)$ as

$$x^i_{s,n} := \frac{1}{\mu_{s,n}} \sum_{k=1}^{n} 1_{s_k=s} a^i_n \tag{6.2}$$

where pure action a^i_n is seen as an element of the Euclidean space $\Delta(A^i)$ and with the convention that if $\mu_{s,n} = 0$, then $x^i_{s,n} = x^i_{s,0}$, which is defined arbitrarily. Thus, $x^i_{s,n+1}$ can be written incrementally as

$$x^i_{s,n+1} = \frac{1_{s_{n+1}=s} a^i_{n+1}}{\mu_{s,n+1}} + \frac{\mu_{s,n} x^i_{s,n}}{\mu_{s,n+1}}. \tag{6.3}$$

Consequently, $x^i_{s,n+1}$ is equal to $x^i_{s,n}$ when s_{n+1} is not equal to s.

Auxiliary Shapley game The second estimate is defined using the payoff of an auxiliary game. Given a state s and a continuation payoff vector $v \in \mathbb{R}^{I \times S}$, we define (following Shapley, 1953) the auxiliary game parameterized by v as the one-shot game where the action set is A^i for every player i and the payoff function is f^i_{s,v^i} where

$$f^i_{s,v^i}(\boldsymbol{a}) := (1-\delta)u^i_s(\boldsymbol{a}) + \delta \sum_{s' \in S} P_{s,s'}(\boldsymbol{a}) v^i_{s'}.$$

Fink (and Shapley) proved that stationary equilibria are the fixed point of an operator based on this auxiliary game.

Update steps α_n is the non-increasing sequence of positive update steps for the payoff estimates whose sum is denoted by $\sigma_n = \sum_{k=0}^{n} \alpha_k$. We make the following assumption on the update steps.

Assumption 6.1 (Discrete Update Steps)**.**

$$\sum_{k=1}^{\infty} \frac{\alpha_k}{\sigma_k} = \infty,$$
$$0 < \alpha_n \leq 1 \textit{ and } \alpha_{n+1} \leq \alpha_n.$$

For instance, α_n may be equal to $\frac{1}{n}$ or a constant $\alpha > 0$.

Payoff estimates Players estimate the continuation payoff in a vector $v^i_n \in \mathbb{R}^S$. Values of this vector are written $v^i_{s,n}$ for state s, at step n for player i. At every step n, the estimator is defined as

$$v^i_{s,n+1} := \frac{1}{\sigma_n} \sum_{k=1}^{n} \alpha_k f^i_{s,v^i_k}(\boldsymbol{x}_{s,k}), \tag{6.4}$$

which can be rewritten

$$v^i_{s,n+1} = \frac{\sigma_{n-1}}{\sigma_n} v^i_{s,n} + \frac{\alpha_n}{\sigma_n} f^i_{s,v^i_n}(\boldsymbol{x}_{s,n}),$$

leading to the incremental form[1]

$$v^i_{s,n+1} - v^i_{s,n} = \frac{\alpha_n}{\sigma_n}(f^i_{s,v^i_n}(\boldsymbol{x}_{s,n}) - v^i_{s,n}). \tag{6.5}$$

Notice that right-hand side of Eq. (6.5) depends on $v^i_{s,n}$, which explains why the sum over k in Eq. (6.4) stops at n and not $n+1$.

In order to control the speed of convergence of $v^i_{s,n}$, we add a factor $\alpha^\star$ to Eq. (6.4)[2],

$$v^i_{s,n+1} - v^i_{s,n} = \alpha^\star \frac{\alpha_n}{\sigma_n} \left(f^i_{s,v^i_n}(\boldsymbol{x}_{s,n}) - v^i_{s,n} \right). \tag{6.6}$$

Remark. *In an identical-interest stochastic game, u^i_s does not depend on i and as a consequence, $v^i_{s,n}$ does not depend on i either. It also holds for zero-sum games because the payoff of player 2 is the negative of that of player 1, therefore it is sufficient to follow player 1's payoff. As such, we omit the superscript i for $v^i_{s,n}$ and f^i_{s,v^i_n} in the rest of the chapter. We show in Section 7.4.4 that if initial values are different (using the incremental updating rule as in Eq.* (6.5)*), convergence results still holds.*

Estimator $v_{s,n}$ can be seen as a mean where recent values of the expected payoffs $f_{s,v_n}(\boldsymbol{x}_{s,n})$ are given less weight than older values. However, if the sequence $f_{s,v_n}(\boldsymbol{x}_{s,n})$ is stationary, v_n will ultimately converge to the same limit as v_n. A similar idea of a fast and a slow update rates is used in (Leslie et al., 2020; Perkins, 2013; Konda and Borkar, 1999; Sayin et al., 2022b).

[1] An alternative definition could be based directly on Eq. (6.5) with a learning rate defined in place of $\frac{\alpha_n}{\sigma_n}$. We prefer writing it with two variables α_n and σ_n in order to express that $v^i_{s,n}$ is a fading average of auxiliary payoffs in the auxiliary game.

[2] Value $\alpha^\star$ is equal to 1 in identical-interest games but may vary in zero-sum games: action profiles converge to A$\alpha^\star$-equilibria of the game.

Remark. *If $\alpha_n = 1$, then $\sigma_n = n$. Thus, the update rates of $x^i_{s,n}$ and v_n are the same and the hypothesis (6.1) holds. It also holds for $\alpha_n = \frac{1}{\log n}$ where $\sigma_n = \log\log n$. In this case, the update rate of $x^i_{s,n}$ is much faster than the one of v_n and, as will be seen, this is the discrete-time analog of the continuous time dynamics in Leslie et al., 2020.*

Action selection We can now define the action selection of our FP procedures. It is an extension of the classical FP procedure. For repeated games, FP is defined as a behavioral strategy where at every stage, every player takes a best response against the empirical action of the opponents up to that stage (see Section 3.3). For stochastic games, we define FP as a best response in the auxiliary Shapley game parameterized by a given continuation payoff $v_{s,n}$, that is for every n,

$$a^i_{n+1} \in \mathrm{BR}^i_{s_{n+1},v_{n+1}}\left(\boldsymbol{x}^{-i}_{s_{n+1},n}\right) := \operatorname*{arg\,max}_{b^i \in A^i} f_{s_{n+1},v_{n+1}}\left(b^i, \boldsymbol{x}^{-i}_{s_{n+1},n}\right).$$

When there are several best responses, our convergence results are independent on the selection rule.

Now we can define precisely our first FP procedure.

Asynchronous FP

$$\begin{cases} \forall s \in S, v_{s,n+1} - v_{s,n} = \alpha^\star \dfrac{\alpha_n}{\sigma_n}\left(f_{s,v_n}(\boldsymbol{x}_{s,n}) - v_{s,n}\right) \\ a^i_{n+1} \in \mathrm{BR}^i_{s_{n+1},v_{n+1}}\left(\boldsymbol{x}^{-i}_{s_{n+1},n}\right) \\ \forall s \in S, x^i_{s,n+1} - x^i_{s,n} = \dfrac{1_{s=s_{n+1}}}{\mu_{s,n+1}}\left(a^i_{n+1} - x^i_{s,n}\right) \end{cases} \tag{AFP}$$

Estimates of continuation payoff $v_{s,n}$ are updated towards the estimated payoff in the auxiliary game $f^i_{s,u^i_n}(x_{n,s})$ for all states s at every step. Empirical actions $x^i_{s,n}$ are updated only for the current state (notice the indicator function) in the direction of the action played a^i_n. These are incremental versions of Eq. (6.3) and Eq. (6.5).

Therefore, this defines a behavioral strategy because the only information needed to update the system variables is the actions played in the current state. Moreover, these equations can be computed in the sense that there is no circular dependencies between variables: $v_{s,n+1}$ is computed using values of step n, a^i_{n+1} is computed using values of step n and $v_{s,n+1}$, $x^i_{s,n+1}$ uses values of step n and a^i_{n+1}.

Remark. *Equations for $v_{s,n}$, a^i_n and $x^i_{s,n}$ are written in this order because of the dependencies between these variables.*

If the stochastic game is not ergodic, then some states may stop being visited and there is no chance of convergence to an equilibrium in those states. This is in another argument that justifies our next synchronous procedure.

6.1.2 Synchronous FP (SyncFP)

To get convergence in non-ergodic stochastic games, we now define a version with synchronous updates on every state. This is an algorithm but not a behavioral strategy and thus is not a learning rule.

Synchronous FP (SyncFP) is defined as follows:

$$\forall s \in S, \begin{cases} v_{s,n+1} - v_{s,n} = \alpha^\star \dfrac{\alpha_n}{\sigma_n}\left(f_{s,v_n}(\boldsymbol{x}_{s,n}) - v_{s,n}\right) \\ a^i_{s,n+1} \in \mathrm{BR}^i_{s,v_{n+1}}\left(\boldsymbol{x}^{-i}_{s,n}\right) \\ x^i_{s,n+1} - x^i_{s,n} = \dfrac{1}{n+1}\left(a^i_{s,n+1} - x^i_{s,n}\right) \end{cases} \tag{SyncFP}$$

Contrary to AFP, an action is provided for every state at every step, as if the state was unknown. This allows to update $x^i_{s,n}$ and $v_{s,n}$ synchronously (but at a potentially different rate).

An alternative to both SyncFP and AFP is fully-asynchronous fictitious play. It can be used in a standard asynchronous play of the stochastic game, it is a behavioral strategy. Both $x^i_{s,n}$ and $v_{s,n}$ are updated only for the current state, leading to system

$$\begin{cases} v_{s,n+1} - v_{s,n} = \dfrac{1_{s_n=s}}{\mu_{s,n+1}} \dfrac{f^i_{s,u_n}(x_{n,s}) - v_{s,n}}{n} \\ a^i_{n+1} \in \mathrm{BR}^i_{s_{n+1},v_{n+1}}\left(\boldsymbol{x}^{-i}_{s_{n+1},n}\right) \\ x^i_{s,n+1} - x^i_{s,n} = \dfrac{1_{s_{n+1}=s}}{\mu_{s,n+1}} \dfrac{a^i_{s,n+1} - x^i_{s,n}}{\mu_{s,n}} \end{cases} \tag{FAFP}$$

6.1.3 Convergence Result

We now state the main convergence results for our FP procedures in identical-interest stochastic games.

Theorem 6.1 (Convergence of FP in identical interest stochastic games)**.** *Under Assumption 6.1, procedures SyncFP and AFP almost surely converge to the set of stationary Nash equilibria in identical-interest ergodic discounted stochastic games and if $\delta < 1/|S|$, then the result also holds for FAFP. The convergence holds also in non-ergodic games for SyncFP.*

As in (Monderer and Shapley, 1996b), our discrete-time proof for identical-interest stochastic games is direct and is not derived from some associated continuous-time system. It is sketched below and detailed in Section 6.4.

When $\delta > 1/|S|$, we conjecture that the result also holds for FAFP, as illustrated by simulations in Section 6.5, but this is still an open problem.

Proof sketch. The central idea is to show that the gap between $f_{s,v_n}(\boldsymbol{x}_{s,n})$ and $v_{s,n}$ is lower-bounded by a sequence whose sum converge. This is possible because $f_{s,v_n}(\boldsymbol{x}_{s,n})$ is mostly non-decreasing (but for the synchronization error of the players that optimize this function which is in $\frac{1}{n^2}$). This is similar to the proof in continuous time (where the payoff function is a Lyapunov function of the system). Another key point is that $v_{s,n}$ moves towards $f_{s,v_n}(\boldsymbol{x}_{s,n})$ at a rate no faster than the updates of $x^i_{s,n}$. This lower bound is used to prove the convergence of $v_{s,n}$, and the convergence to the set of stationary Nash equilibria follows.

□

Theorem 6.2 (Convergence of FP in zero-sum stochastic games)**.** *Under the assumption that $\forall n \in \mathbb{N}, \alpha_n = 1$, there exists $A > 0$, independent of the game, such that procedures SyncFP and AFP almost surely converge to the set of stationary $A\alpha^\star$-Nash equilibria in zero-sum ergodic discounted stochastic games. If $\delta < 1/|S|$, then the result also holds for FAFP. The convergence holds also in non-ergodic games for SyncFP.*

The proof of this latest theorem is sketched in the rest of the chapter. It uses the stochastic approximation framework. We recall the classical theory and the asynchronous extension that we need to modify.

6.2 Best-Response Dynamics

This section extends and studies the best-response dynamics introduced and studied in zero-sum stochastic games by Leslie et al. (2020). We generalize their updating rates and prove that all the extended dynamics converge to stationary equilibria in identical interest and in zero-sum stochastic games. These dynamics are the continuous-time counterpart of AFP and SyncFP as shown in the next section.

As in discrete-time, there are two sets of variables: $\{v^i_s, x^i_s\}_{s\in S, i\in I}$. These variables may have different update rates, and we suppose there is a function $\alpha : \mathbb{R}^+ \to \mathbb{R}^{+*}$ to express the update rates of variables u^i_s. Function α is continuous and non-increasing. We make the following additional assumption on α:

Assumption 6.2.

$$\int_0^t \alpha(y)dy \xrightarrow[t\to\infty]{} +\infty,$$
$$\alpha(t) \geq 0 \text{ and } \alpha \text{ is non-increasing}$$

Synchronicity and asynchronicity Similarly to the FP procedure, we study three types of systems: synchronous, asynchronous and fully-asynchronous ones. In the synchronous kind, variables of all states are updated at the same time. There is no distinguished, current state. In semi-asynchronous systems, there is a current state and variables x_s^i are updated only if they are related to the current states but variables $v_{s,n}$ are updated even if the current state is not s. In fully asynchronous systems, variables x_s^i and v_s are updated if and only if the current state is s.

We start with the synchronous dynamics which is mathematically more tractable.

Synchronous Dynamics As in SyncFP, in the next dynamics, variables of all states are updated at the same time. For $t \geq 0$ and every state s and player i, synchronous best-reply dynamics (SyncBRD) is defined as:

$$\begin{cases} \dot{v}_s^i(t) = \alpha(t)\left(f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s^i(t)\right) \\ \dot{x}_s^i(t) \in \mathrm{BR}^i_{s,v_s^i(t)}\left(x_s^i(t)\right) - x_s^i(t) \end{cases} \tag{SyncBRD}$$

where $\mathrm{BR}^i_{i,v_s^i(t)}\left(x_s^i(t)\right)$ is defined as previously.

Remark. *This is a generalization of the definition of Leslie et al. who studied the case $\alpha(t) = \frac{1}{t+1}$. Replacing $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ by the maximum over actions, that is $\max_{a^i \in A^i} f^i_{s,\boldsymbol{v}(t)}\left(a^i, \boldsymbol{x}_s^{-i}(t)\right)$ is an alternative that would be closer to the system outlined by Sayin et al. and Q-learning in general. It could be an interesting system to study but as noted by Sayin et al., this would result in $v_s^i(t)$ to be different for two players even if the game is zero-sum or identical interest, which poses more theoretical challenges.*

Differential inclusion SyncBRD classically admits a (typically non-unique) solution (Aubin and Cellina, 1984; Benaim et al., 2005). Indeed, one can rewrite it as $\frac{dy}{dt} \in F(t, y)$ where y is a vector with every u_s^i, x_s^i and F is Marchaud (see Definition 4.1). Furthermore, as shown in Lemma 7.4 later on, values are bounded, so the solution is defined on $\mathbb{R}^+$ (Aubin and Cellina, 1984, p. 97).

In identical-interest games, $u_s^i = u_s$ for every player i. Therefore, for every s, $v_s^i(t)$ and $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ do not depend on i (when initial values are equal), hence we omit the superscript i in our statements. It is similar for zero-sum games.

Asynchronous Dynamics We now describe asynchronous and fully-asynchronous dynamics. In the asynchronous system, expected payoffs are updated at a constant rate, but empirical actions are not. For the fully asynchronous system, both payoff estimates and empirical actions are updated at the same state-dependent rate.

The fully asynchronous system is defined as follows:

$$\begin{cases} \dot{v}_s(t) = \beta_s(t)\,\alpha\left(\int_0^t \beta_s(y)dy\right)\left(f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s(t)\right) \\ \dot{x}_s^i(t) \in \beta_s(t)\left(\mathrm{BR}^i_{s,v_s(t)}\left(\boldsymbol{x}_s^{-i}(t)\right) - x_s^i(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{FABRD}$$

where $\beta_- \in (0, 1]$.

The asynchronous system (in the spirit of the system of Leslie et al. (2020)) is:

$$\begin{cases} \dot{v}_s(t) = \alpha(t)\left(f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s(t)\right) \\ \dot{x}_s^i(t) \in \beta_s(t)\left(\mathrm{BR}^i_{s,v_s(t)}\left(\boldsymbol{x}_s^{-i}(t)\right) - x_s^i(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{ABRD}$$

Value $\beta_s(t)$ is the update rate for state s at time t. If only one state was updated at every time point, then we would have $\beta_s(t)$ equal to 0 for all states s but in one state where it would be equal to 1. If the game is ergodic, then on average every state is reached a strictly positive proportion of the time, greater than β_-. Next section will show in the ergodic case, that this system is formally linked to the AFP procedure. This is a mathematically convenient way to use the ergodicity hypothesis.

Remark. *Note that this is different from the model of continuous-time stochastic game outlined for instance by Neyman, 2017. In this paper, we are interested in using the theory of stochastic approximation, therefore discrete-time and continuous-time system are related through an exponential change of variable in time. As a consequence, we can consider that the state occupation is averaged, which is not the case in some other studies in continuous time of stochastic games.*

Theorem 6.3 (Convergence of ABRD, SyncBRD and FABRD in identical interest stochastic games)**.** *Let $\{v_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ be a solution of ABRD or SyncBRD. Under Assumption 6.2, there is $\boldsymbol{v}_\infty \in \mathbb{R}^{|S|}$ such that:*

(i) *for all s, $f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) \xrightarrow[t\to\infty]{} v_{\infty,s}$ and $v_s(t) \xrightarrow[t\to\infty]{} v_{\infty,s}$*

(ii) *$\boldsymbol{v}_\infty$ is a stationary Nash equilibrium payoff*

(iii) *$\{\boldsymbol{x}_s(t)\}_{s\in S}$ converges to the set of stationary Nash equilibria with payoff $\boldsymbol{v}_\infty$*

It also holds for solutions of FABRD when $\delta < \frac{1}{|S|}$.

A sketch of the proof is provided below. A comprehensive proof with technical lemmas is provided in the next chapter, see Section 7.3.4. We conjecture it also holds for FABRD when $\delta \geq \frac{1}{|S|}$, see simulations in Section 7.5.

Sketch of proof for SyncBRD and ABRD. We suppose that $\{v_s, x_s\}_{s\in S}$ is **a solution of ABRD** (which includes the case SyncBRD). We define, for $s \in S$:

$$\begin{aligned}
\Gamma_s(t) &:= f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) \\
\Delta_s^i(t) &:= \max_{y^i\in A^i} f_{s,\boldsymbol{v}(t)}\big(y^i, \boldsymbol{x}_s^{-i}(t)\big) - f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) \\
&= \max_{y^i\in A^i} f_{s,\boldsymbol{v}(t)}\big(y^i, \boldsymbol{x}_s^{-i}(t)\big) - \Gamma_s(t)
\end{aligned}$$

We are going to lower bound $\Gamma_s(t) - v_s(t)$ for every s so as the differential of v_s is lower-bounded by an integrable function. This guarantees that, as v_s is bounded (see Lemma 7.7), it converges. We will then show that for every player i, $\Delta_s^i(t) \to 0$ and finish the proof of the theorem by showing convergence of Γ_s and studying the limit set of x_s^i.

Let $s \in S$. First, note that $\Delta_s^i(t) \geq 0$. Function Γ_s is differentiable:

$$\frac{d\Gamma_s}{dt} = \delta \sum_{s'} P_{ss'}(x_s)\dot{v}_{s'}(t) + \beta_s(t) \sum_i \Delta_s^i(t) \tag{6.7}$$

where $\beta_s(t) = 1$ for SyncBRD and is already defined for ABRD. See Lemma 7.6 for details.

Lower bound of $\Gamma_s(t) - v_s(t)$ for SyncBRD and ABRD Let $s_-(t) \in \arg\min_{s'\in S}(\Gamma_{s'}(t) - v_{s'}(t))$. Then, for any $s \in S$:

$$\begin{aligned}
\frac{d\Gamma_s}{dt} &\geq \delta \sum_{s'} P_{s,s'}(x_s)\alpha(t)(\Gamma_{s'}(t) - v_{s'}(t)) \\
&\geq \delta \sum_{s'} P_{ss'}(x_s)\alpha(t)\Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) \\
&= \delta\alpha(t)\Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big)
\end{aligned} \tag{6.8}$$

Moreover, for $\tau > 0$:

$$\begin{aligned}
&\Gamma_{s_-(t+\tau)}(t+\tau) - v_{s_-(t+\tau)}(t+\tau) - \Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) \\
&\geq \Gamma_{s_-(t+\tau)}(t+\tau) - v_{s_-(t+\tau)}(t+\tau) - \Big(\Gamma_{s_-(t+\tau)}(t) - v_{s_-(t+\tau)}(t)\Big) \\
&\geq \tau \min_{s' \in S} \frac{d\Gamma_{s'}}{dt} + o(\tau) + v_{s_-(t+\tau)}(t) - v_{s_-(t+\tau)}(t+\tau)
\end{aligned} \tag{6.9}$$

Then, it can be shown that if s' is an accumulation point of $s_-(t+\tau)$ when τ goes to 0, then

$$v_{s'}(t) - v_{s'}(t+\tau) = -\tau\alpha(t)\Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) + o(\tau). \tag{6.10}$$

Since this is valid for every such s', combining (6.8), (6.9) and (6.10) leads to

$$\begin{aligned}
&\Gamma_{s_-(t+\tau)}(t+\tau) - v_{s_-(t+\tau)}(t+\tau) - \Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) \\
&\qquad \geq \tau(\delta-1)\alpha(t)\Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) + o(\tau).
\end{aligned}$$

Now we are going to apply a version of Grönwall Lemma (see details in Lemma 7.7) so we get:

$$\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t) \geq \Big(\Gamma_{s_-(0)}(0) - v_{s_-(0)}(0)\Big) \exp\left(\int_0^t (\delta-1)\alpha(\omega)d\omega\right) \tag{6.11}$$

Therefore, we can use this inequality in $\dot{v}_s(t)$:

$$\dot{v}_s(t) \geq -A\alpha(t) \exp\left(\int_0^t (\delta-1)\alpha(\omega)d\omega\right) \tag{6.12}$$

where $A > 0$. The right hand side term is integrable, and as v_s is bounded (see Lemma 7.4), it converges.

Sum $\sum_{i\in I} \Delta_s^i(t)$ goes to 0 We show that $\Delta_s^i(t) \to 0$. First, we notice that in the SyncBRD and ABRD cases:

$$\begin{aligned}
\int_0^t \sum_{i\in I} \Delta_s^i(\omega)d\omega &\leq \int_0^t \frac{\beta_s(\omega)}{\beta_-} \sum_{i\in I} \Delta_s^i(\omega)d\omega && \text{(by construction, } \beta_s(\omega) \geq \beta_-) \\
&= \frac{1}{\beta_-}\left(\int_0^t \frac{d\Gamma_s}{dt}(\omega) - \delta \sum_{s'} P_{ss'}(x_s(\omega))\dot{v}_{s'}(\omega)d\omega\right) && \text{(expansion of Eq. (6.7))} \\
&\leq \frac{1}{\beta_-}\left(\Gamma_s(t) - \Gamma_s(0)\right) \\
&\quad + \frac{A}{\beta_-(1-\delta)}\left(1 - \exp\left(\int_0^t (\delta-1)\alpha(\omega)d\omega\right)\right) && \text{(integration via Eq. (6.12))}
\end{aligned}$$

So, this integral is bounded. However, as $\sum_{i\in I} \Delta_s^i(\cdot)$ is Lipschitz (see Lemma 7.5), we conclude that $\sum_{i\in I} \Delta_s^i(t) \xrightarrow[t\to\infty]{} 0$ (Lemma 7.7).

In the ABRD case, we have similar inequalities except for the argument of α which is $\int_0^t \beta_s(\omega)d\omega$. As it is bounded between $\beta_- t$ and t, it does not change much of the computations and the integral is bounded as well. See Lemma 7.7 for details.

Convergence of Γ_s In the case of SyncBRD and ABRD, using Eq. (6.8), we can lower bound its derivative:

$$\frac{d\Gamma_s}{dt} \geq \delta\alpha(t)\Big(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\Big) \geq -\delta A\alpha(t)\left(\exp\left(\int_0^t (\delta-1)\alpha(\omega)d\omega\right)\right)$$

This latest term being integrable and Γ_s being bounded, we conclude that Γ_s converges to its $\limsup$ when t goes to $+\infty$. The limit is necessarily the same as v_s, otherwise v_s could not be bounded (see the comprehensive proof for details).

Limit set of $x_s^i(t)$ Let $\tilde{x}$ be an accumulation point of the vector-valued function $x = \{x_s\}$. Then, we previously showed:

$$\Delta_s^i(t) = f_{s,\boldsymbol{v}(t)}\Big(br^i_{s,\boldsymbol{v}(t)}(x_s^{-i}(t)) - x_s^i(t), x_s^{-i}(t)\Big) \xrightarrow[t\to\infty]{} 0$$

So by continuity, for all s:

$$f_{s,\lim \boldsymbol{v}}(br^i_{s,\lim \boldsymbol{v}}(\tilde{x}^{-i}) - \tilde{x}^i_s, \tilde{x}_s^{-i}) = 0$$

So $\tilde{x}$ belongs to the set of Nash equilibria.

□

Theorem 6.4 (Convergence of ABRD in zero-sum stochastic games)**.** *Let $\{v_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ be a solution of ABRD. There exists a constant $A > 0$ (which only depends on δ and r_s) such that if $\alpha^\star > \lim_{t\to\infty}\alpha(t)$, then, under Assumption 6.2:*

(i) *for all s, $\limsup_{t\to\infty} |f_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) - v_s(t)| \leq A\alpha^\star$*

(ii) *$\{x_s(t)\}_{s\in S}$ converges to the set of stationary Nash $A\alpha^\star-$equilibria as $t \to \infty$.*

The proof is in Section 7.3.5. Note that if $\alpha(t) \to 0$, then $\alpha^\star$ can be chosen arbitrarily close to 0 which is the case in (Leslie et al., 2020) ($\alpha(t) = t + 1$). Hence this is an extension of (Leslie et al., 2020).

6.3 From Continuous-Time to Discrete-Time

In the following, we use the stochastic approximation framework, reviewed in Section 4.4. We describe an extension to a new setting, that of correlated asynchronicity.

Correlated Asynchronicity To do similar proofs for systems AFP and FAFP, one needs asynchronous stochastic approximations: the standard stochastic approximation framework Benaim et al., 2005 is not sufficient to track every state and make every update rate depends on $\mu_{s,n}$. We therefore use a theorem published by Perkins and Leslie (2012) which makes it possible to take into account correlated asynchronicity. Indeed, in AFP, variables v_s are updated at every time step, independently of the current state and in FAFP, variables v_s and $\boldsymbol{x}_s$ are updated at the same time. We apply this result to our systems ABRD and FABRD, in order to prove Theorem 6.1 under the ergodicity hypothesis.

Proof of Theorem 6.1 (sketch). Then, it remains to show that the ICT sets of this generalized continuous-time system are contained in the set of stationary Nash equilibria and their associated payoff. To do this, we use results of convergence of the previous section. However, there is no direct implication between the convergence to a set and the fact that this set is ICT. Therefore, the chain transitivity is proven in part using the original definition, and in part with a Lyapunov function. More precisely we want to show that any ICT set is included in:

$$B := \Big\{(\boldsymbol{x}, \boldsymbol{v}) \mid \forall s \in S\ f_{s,\boldsymbol{v}}(\boldsymbol{x}_s) \geq v_s\Big\}$$

$$\text{and } A := \left\{(x, \boldsymbol{v}) \mid \begin{array}{r} \forall s \in S\, \forall i \in I,\ f_{s,\boldsymbol{v}}(\boldsymbol{x}_s) = v_s \\ \wedge\ x_s^i \in \arg\max_{y^i \in A^i} f_{s,\boldsymbol{v}}\left(y^i, \boldsymbol{x}_s^{-i}\right) \end{array}\right\}$$

First, we show that any element $(\boldsymbol{x}, \boldsymbol{v})$ of ICT sets are in B. Otherwise, we look at the chain between $(\boldsymbol{x}, \boldsymbol{v})$ and itself, and conclude to a paradox: any solution is arbitrarily close to B after a time T independently of the starting point. Then, relatively to B, we can define a function $V(\boldsymbol{x}, \boldsymbol{v}) := \sum_{s\in S} f_{s,v_n}\left(\boldsymbol{x}_{s,n}\right)$ which is a Lyapunov function in B. This makes it possible to conclude that any ICT set is included in $V^{-1}(0)$ which is equal to A.

The whole proof is detailed in Section 7.4.

□

6.4 Direct Proof of Discrete Time FP in Identical-Interest Stochastic Games

In this section, we prove that systems SyncFP and AFP converge to the set of stationary Nash equilibria in identical-interest stochastic games. The proofs for the two systems are similar, except for the last part about the convergence of the empirical actions. Therefore, we write the first, identical part, only for AFP (the more complex system), and give the two proofs in the last part.

We consider system AFP defined as

$$\begin{cases} \forall s \in S, v_{s,n+1} - v_{s,n} = \dfrac{\alpha_n}{\sigma_n}\left(f_{s,v_n}(\boldsymbol{x}_{s,n}) - v_{s,n}\right) \\ a^i_{n+1} \in \mathrm{BR}^i_{s_{n+1},v_n}(\boldsymbol{x}^{-i}_{s,n}) \\ \forall s \in S, x^i_{s,n+1} - x^i_{s,n} = \dfrac{1_{s=s_{n+1}}}{\mu_{s,n+1}}\left(a^i_{n+1} - x^i_{s,n}\right) \\ \sigma_n = \displaystyle\sum_{k=1}^{n} \alpha_k \end{cases} \tag{AFP}$$

under this hypothesis on α_n:

$$\sum_k \frac{\alpha_k}{\sigma_k} = \infty, \quad \alpha_n \leq 1, \text{ and } \alpha_{n+1} \leq \alpha_n \tag{H1}$$

Note that this is satisfied for $\alpha_n = 1$ (single timescale, autonomous case) or $\alpha_n = \frac{1}{n}$ (v_n is updated slower than $\boldsymbol{x}_n$).

We are going to show that under H1, $v_{s,n}$ and $f_{s,v_n}(\boldsymbol{x}_{s,n})$ converge to the same equilibrium payoff for every s in an (ergodic for AFP), identical interest stochastic game. This implies that $x_{n,s}$ converges to a stationary Nash equilibrium.

Proof of Theorem 6.1. We define a few notations with

$$\begin{aligned} \Gamma_{s,n} &:= f_{s,v_n}(\boldsymbol{x}_{s,n}), \\ w_n &:= \min_{s\in S} \Gamma_{s,n} - v_{s,n}, \\ s^-_n &\in \arg\min_{s\in S} \Gamma_{s,n} - v_{s,n}, \\ \text{and } \mu_n &= \min_{s\in S} \mu_{s,n+1}. \end{aligned}$$

Intuitively, the absolute value $|w_n|$ denotes the energy of the system, i.e., how far away are estimates $v_{s,n}$ from $\Gamma_{s,n}$. These definitions imply that

$$w_n = \Gamma_{s^-_n,n} - v_{s^-_n,n}.$$

Because of the ergodicity assumption, there exists λ such that with probability 1, there exists n_0 such that

$$\lambda n \leq \mu_n \leq n. \tag{6.13}$$

Sequence to w_n approaches $\mathbb{R}^+$ We bound the changes in w_n:

$$\begin{aligned} w_{n+1} - w_n &= \Gamma_{s^-_{n+1},n+1} - v_{s^-_{n+1},n+1} - (\Gamma_{s^-_n,n} - v_{s^-_n,n}) \\ &\geq \Gamma_{s^-_{n+1},n+1} - v_{s^-_{n+1},n+1} - (\Gamma_{s^-_{n+1},n} - v_{s^-_{n+1},n}) \\ &= \Gamma_{s^-_{n+1},n+1} - \Gamma_{s^-_{n+1},n} - (v_{s^-_{n+1},n+1} - v_{s^-_{n+1},n}) \\ &= \Gamma_{s^-_{n+1},n+1} - \Gamma_{s^-_{n+1},n} - \frac{\alpha_k}{\sigma_k}\left(\Gamma_{s^-_{n+1},n} - v_{s^-_{n+1},n}\right) \end{aligned} \tag{6.14}$$

Now, there exists C (independent of n) such that $|\Gamma_{s^-_{n+1},n} - v_{s^-_{n+1},n}| - |\Gamma_{s^-_n,n} - v_{s^-_n,n}| < \frac{C}{\mu_n}$ (because for every s, $\Gamma_{s,n} - v_{s,n}$ changes of at most $\frac{C}{\mu_n}$ between n and $n+1$, so this is true for the minimum as well).

As a consequence, continuing Eq. (6.14):

$$\begin{aligned}
w_{n+1} - w_n &\geq \Gamma_{s^-_{n+1},n+1} - \Gamma_{s^-_{n+1},n} - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{\mu_n \sigma_n} \\
&\geq f_{s^-_{n+1},v_{n+1}}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) - f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) \\
&\quad + f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) - f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{\mu_n \sigma_n} \\
&\geq \delta \sum_{s' \in S} P_{s',s^-_{n+1}}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right)\left(v_{s',n+1} - v_{s',n}\right) \\
&\quad + f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) - f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{\mu_n \sigma_n} \\
&\geq \delta \frac{w_n}{\sigma_n \alpha_n} + f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) - f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{\mu_n \sigma_n} \\
&\geq (\delta - 1)\frac{\alpha_n w_n}{\sigma_n} + f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) - f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n}\right) - \frac{\alpha_n C}{\mu_n \sigma_n}
\end{aligned}$$

The first order expansion of $f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right)$ for AFP:

$$\begin{aligned}
f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n+1}\right) &= f_{s^-_{n+1},v_n}\left(\boldsymbol{x}_{s^-_{n+1},n}\right) \\
&\quad + \sum_{i \in I} \frac{1_{s=s_n}}{\mu_{s,n}} \left(f_{s^-_{n+1},v_{s,n}}(a^i_n, x^{-i}_{s,n}) - f_{s^-_{n+1},v_{s,n}}\left(\boldsymbol{x}_{s,n}\right)\right) + O\left(\frac{1}{\mu_n^2}\right) \qquad (6.15)
\end{aligned}$$

The expansion Eq. (6.15) would be the same for SyncFP except for the indicator $1_{s=s_n}$ which would disappear and μ_n which would be replaced by n.

In both cases, the first order term is positive (because a^i_n is a best-response in the auxiliary game), therefore there exists $D > 0$ such that:

$$w_{n+1} - w_n \geq (\delta - 1)\frac{\alpha_n w_n}{\sigma_n} - \frac{D}{\mu_n^2} - \frac{\alpha_n C}{\mu_n \sigma_n} \qquad (6.16)$$

which using Eq. (6.13) leads to

$$w_{n+1} - w_n \geq (\delta - 1)\frac{\alpha_n w_n}{\sigma_n} - \frac{D}{\lambda n^2} - \frac{\alpha_n C}{\lambda n \sigma_n} \qquad (6.17)$$

Then, using a discrete-time Grönwall lemma (proven in the next section, see Lemma 6.1), for $n > m$,

$$\begin{aligned}
w_n &\geq w_m \prod_{k=m}^{n} (1 + \frac{\delta - 1}{k}) - \sum_{k=m}^{n} \left[\frac{D}{\lambda k^2} + \frac{\alpha_k C}{\lambda k \sigma_k} \right] \\
&\geq E \prod_{k=m}^{n} (1 + \frac{\delta - 1}{k}) - \sum_{k=m}^{\infty} \left[\frac{D}{\lambda k^2} + \frac{C}{\lambda k^2} \right] \qquad (6.18)
\end{aligned}$$

for some $E > 0$ (independent of m because w_n is bounded). Equation (6.18) is obtained using Lemma 6.2 (also proven in the next section). The right term goes to 0 as the rest of a convergent sum. Furthermore, the left term goes to 0 when n goes to ∞, so $\limsup w_n \geq 0$.

The continuation payoffs u_n converge Sequence w_n can be lower bounded more precisely, since

$$\sum_{k=m}^{\infty} \frac{D + C}{k^2} = \Omega\left(\frac{1}{m}\right)$$

$$\text{and } \prod_{k=m}^{n} \left(1 + \frac{\delta - 1}{k}\right) = \Omega\left(\left(\frac{m}{n}\right)^{\delta - 1}\right),$$

so with $m = [\sqrt{n}]$,

$$w_n \geq \Omega\Big(n^{\frac{\delta-1}{2}}\Big) + \Omega\left(\frac{1}{\sqrt{n}}\right) = \Omega\Big(n^{\frac{\delta-1}{2}}\Big).$$

Therefore, for every s,

$$v_{s,n+1} - v_{s,n} \geq \Omega\Big(n^{\frac{\delta-1}{2}-1}\Big),$$

and as $v_{s,n}$ is bounded (again using Lemma 6.2), it converges.

The payoff of the auxiliary game converges to the same limit Similarly, one can show that

$$f_{s,v_{n+1}}\big(\boldsymbol{x}_{s,n+1}\big) - f_{s,v_n}\big(\boldsymbol{x}_{s,n}\big) \geq \Omega\Big(n^{\frac{\delta-1}{2}-1}\Big),$$

so it converges, and it is the same limit as $v_{s,n}$ (otherwise $v_{s,n}$ could not be bounded).

The limit is an equilibrium payoff in AFP Using (6.15), (valid for every s), writing

$$\Delta_{s,n} := \sum_{i\in I} f_{s,v_{s,n}}\big(a_n^i, x_{s,n}^{-i}\big) - f_{s,v_{s,n}}\big(\boldsymbol{x}_{s,n}\big),$$

we can rewrite the change of the payoff in the auxiliary game as

$$f_{s,v_n}\big(\boldsymbol{x}_{s,n+1}\big) - f_{s,v_n}\big(\boldsymbol{x}_{s,n}\big) = \frac{1_{s=s_n}}{\mu_n}\Delta_{s,n} + O\left(\frac{1}{\mu_n^2}\right). \tag{6.19}$$

Then:

$$\begin{aligned} f_{s,v_{n+1}}\big(\boldsymbol{x}_{s,n+1}\big) - f_{s,v_n}\big(\boldsymbol{x}_{s,n+1}\big) &= \delta \sum_{s'\in S} P_{ss'}(\boldsymbol{x}_{s,n+1})(v_{s',n+1} - v_{s',n}) \\ &= \delta \sum_{s'\in S} P_{ss'}(\boldsymbol{x}_{s,n+1})\frac{\alpha_n}{\sigma_n}(f_{s',v_n}\big(\boldsymbol{x}_{s,n}\big) - v_{s',n}) \\ &\geq \delta|S|P_{ss'}(\boldsymbol{x}_{s,n+1})\frac{\alpha_n}{\sigma_n}w_n \end{aligned} \tag{6.20}$$

Summing (6.19) and (6.20) gives:

$$f_{s,v_{n+1}}\big(\boldsymbol{x}_{s,n+1}\big) - f_{s,v_n}\big(\boldsymbol{x}_{s,n}\big) \geq \delta|S|\frac{\alpha_n}{\sigma_n}w_n + \frac{1_{s=s_n}}{\mu_n}\Delta_{s,n} + O\left(\frac{1}{\mu_n^2}\right) \tag{6.21}$$

However, $\Delta_{s,n} \geq 0$, so summing (6.21) over n gives that $\sum_n \frac{1_{s=s_n}}{\mu_n}\Delta_{s,n} < \infty$ with the same reasoning as above (because $\frac{\alpha_n}{\sigma_n}w_n = \Omega\Big(n^{\frac{\delta-1}{2}-1}\Big)$ and the terms in the left hand side cancel out).

Simple calculations yield that

$$\frac{1}{\mu_n}\sum_{k=1}^{n} 1_{s=s_k}\Delta_{s,k} \xrightarrow[n\to\infty]{} 0 \tag{6.22}$$

However, it is clear that changes in $\Delta_{s,n}$ are of the order of magnitude of the update steps, that is $\frac{1}{\mu_n}$. As a consequence, assuming that $\Delta_{s,n}$ does not go to 0, there exists $A > 0$ such that for $\epsilon > 0$, if $\Delta_{s,n} \geq 3\epsilon$, then $\Delta_{s,n+m} \geq 3\epsilon - \sum_{k=1}^{m}\frac{A}{n+k} \geq 2\epsilon - A\log((n+m)/n)$ for n large enough (well known result of the harmonic series). But then, for $m = [n(\exp(\epsilon/A) - 1)]$, $\forall k \in \{n, n+m\}$,

$$\Delta_{s,k} \geq \epsilon$$

Since the game is ergodic, with probability 1 when m goes to ∞ (and it goes to infinity when n goes to infinity), $\frac{1}{m}\sum_{k=n}^{n+m} 1_{s=s_k}$ is greater than λ which depends only on the game (minimal frequency of visit of s using the law of large numbers).

$$\frac{1}{n+m}\sum_{k=n}^{n+m} 1_{s=s_k}\Delta_{s,k} \geq \frac{1}{n+m}m\lambda\epsilon \geq \frac{n(\exp(\epsilon/A)-1)}{n+n(\exp(\epsilon/A)-1)}\lambda\epsilon \geq \frac{\exp(\epsilon/A)-1}{\exp(\epsilon/A)}\lambda\epsilon$$

This latest inequality contradicts Eq. (6.22), so $\Delta_{s,n}$ goes to 0 almost surely when n goes to infinity, proving that the limit of $v_{s,n}$ is an equilibrium payoff. Then, it is clear that $\boldsymbol{x}_{s,n}$ converge towards the set of Nash equilibria almost surely (otherwise $f_{s,v_n}\big(\boldsymbol{x}_{s,n}\big)$ could not have the same limit as $v_{s,n}$).

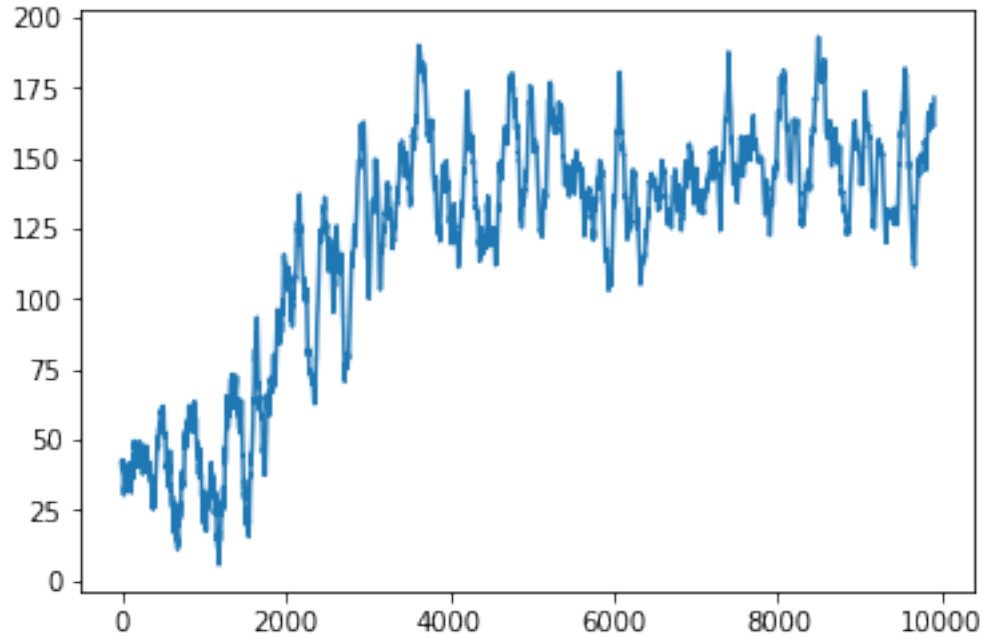


Figure 6.1: (Identical) rewards in the Power Unit Commitment game with our Fully Asynchronous FP procedure with $\delta = 0.5$

For SyncFP The proof is similar but for the indicator function which disappears. As a consequence, it is not needed to use a λ, but we still prove that almost surely, $\Delta_{s,k}$ goes to 0. Therefore, we do not need the ergodicity hypothesis for this case.

□

6.5 Simulation: Power Unit Commitment

In this section, we consider the Power Unit Commitment game defined at page 44, modified to be ergodic: at every step and with probability 0.1, a new state is randomly taken. This is required to make FP explore all states. This is an identical-interest stochastic game.

Agents have to find, for every temperature, a combination of power generation that is greater than demand, under their constraint and without changing too often their production level. The game features 3 players, with parameters described in Appendix A.1.

Fully Asynchronous Fictitious Play is used on this model with a discount factor $\delta = 0.5$, the same learning rate for empirical actions and continuations payoffs (i.e., , $\frac{1}{n}$). It is run 10000 steps.

Fig. 6.1 shows the evolution of the moving average of the laste 100 payoff values. It starts with low payoff as players get the penalty a number of times before they manage to coordinate to action profiles without risk of production lower than demand, and then it is stable around a value of 150. It is noisy as the temperature keeps changing and randomness is introduced in the system to make it ergodic (which may model imperfect information).

Fig. 6.2 shows a random continuation payoff, that progressively stabilizes around level 59. Note that since there are a lot of states (250) and they are only updated when the system is in this particular state, learning takes some time, which explains why a large number of iterations is necessary.

Simulation notebooks are in the supplementary material: `https://www.lamsade.dauphine.fr/~lbaudin/manuscript/`.

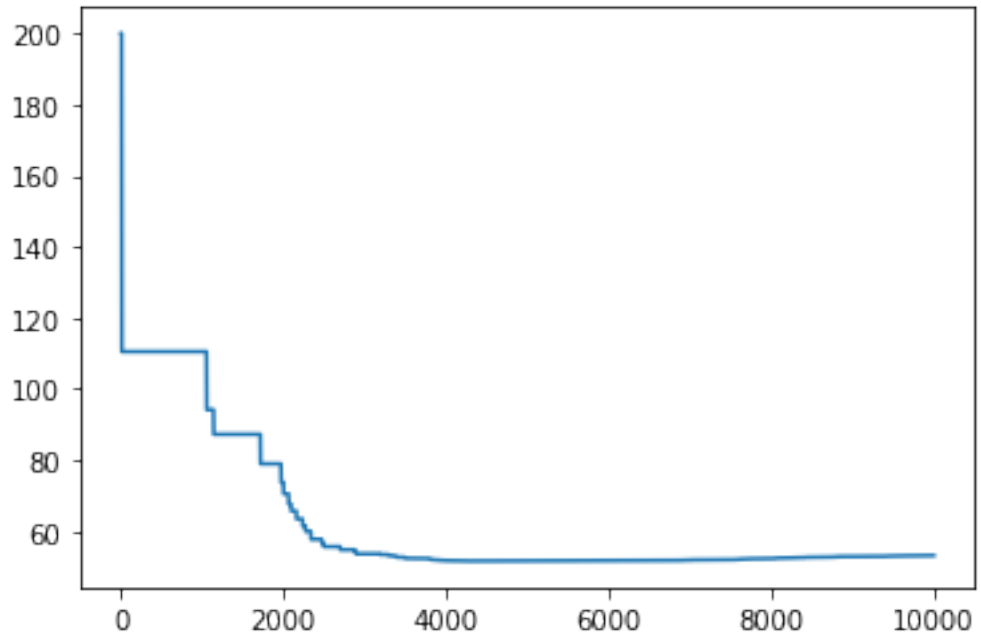


Figure 6.2: Continuation value of a random state in the Power Unit Commitment game with our Fully Asynchronous FP procedure with $\delta = 0.5$

6.6 Conclusion

We defined a number of continuous and discrete time systems to learn stationary equilibria in stochastic games. They combine ideas from FP and Q-learning and are extensions of a continuous-time system of Leslie et al. (2020) who proved its convergence to stationary equilibria in zero-sum stochastic games. We prove their convergence to stationary equilibria in continuous time but also in discrete time; in zero-sum but also in identical-interest discounted stochastic games.

Perspectives An interesting direction is the speed of convergence. As outlined in the proof, there are bounds for zero-sum stochastic games but none for identical-interest ones. To the best of our knowledge, no results are known even in non-stochastic games (Monderer and Shapley, 1996b).

In this chapter, we focused on discounted utilities. Other aggregations are possible and an interesting extension would be limiting average stochastic games. This could be achieved by increasing in AFP the discount factor δ_n from stage-to-stage to 1.

Another interesting direction is the design of a learning procedure that converges to a stationary equilibrium when the players do not observe other players past actions but only their own past actions and the current state. Even when there is only one state (a repeated game) FP and most of its variants are not adapted to this case because there is no way to form a belief about the opponents without observing their actions. Leslie and Collins (2006) shows how, in a repeated game, an actor-critic process can be designed without knowledge of other player actions while retaining some important aspects of a fictitious play process.

A class of procedures that converge to Nash equilibria of the stage game in zero-sum and identical-interest repeated games are no-regret algorithms (Blum and Mansour, 2005; Hofbauer and Sandholm, 2002) with some exploration to be able to estimate what would the payoff be if a player has played differently (we are in a bandit setting).

Another interesting extension would be to have a model-free[3] algorithm in the sense that the game is not known a priori: this is the objective of the next chapter.

[3]This is the definition we use for model-free algorithm, i.e., the algorithm may estimate a model but it is not known a priori. In the litterature, another definition of model-free algorithms is algorithms that do not have complete knowledge of the model *and* that do not attempt to estimate them.

6.7 Postponed Proofs

Discrete-time Grönwall can be found in the literature with various assumptions. For the sake of completeness, we include here a version that matches the assumptions we have in this chapter, with the associated proof. It is a differential version with error terms.

Lemma 6.1 (Discrete-Time Grönwall)**.** *Let $\{y_n\}$, $\{g_n\}$, $\{b_n\}$ sequences of real numbers such that $1 > 1+g_n > 0$ for all n and:*

$$y_{n+1} - y_n \leq g_{n+1} y_n + b_{n+1}$$

Then $y_n \leq y_0 \Pi_{k=0}^{n}(1+g_k) + \sum_{k=0}^{n} b_k$.

Proof. We define $v_n := \frac{y_n - \sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)}$. We show that v_n is decreasing:

$$\begin{aligned}
v_{n+1} - v_n &= \frac{y_{n+1} - y_n}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n+1}(1+g_k)} - \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{g_{n+1} y_n + b_{n+1}}{\Pi_{k=0}^{n}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} \left(\frac{1}{1+g_{n+1}} - 1 \right) - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{g_{n+1} y_n + b_{n+1}}{\Pi_{k=0}^{n}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} \frac{-g_{n+1}}{1+g_{n+1}} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{b_{n+1}}{\Pi_{k=0}^{n+1}(1+g_k)} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \left(1 - \frac{1}{1+g_{n+1}} \right) \\
&\leq 0
\end{aligned}$$

And $v_0 = y_0$, hence the result.

□

Lemma 6.2 (Bound on α_n)**.** *Under hypothesis 6.1, for all n, $\frac{\alpha_n}{\sigma_n} \leq \frac{1}{n+1}$.*

Proof. By induction: for $n = 0$, $\frac{\alpha_n}{\sigma_n} = 1$. Now for $n+1$:

$$\frac{\alpha_{n+1}}{\sigma_{n+1}} \leq \frac{\alpha_n}{\sigma_n} \frac{\sigma_n}{\sigma_{n+1}} \leq \frac{1}{n+1} \left(1 - \frac{\alpha_{n+1}}{\sigma_{n+1}} \right)$$

As a consequence:

$$\frac{\alpha_{n+1}}{\sigma_{n+1}} \frac{n+2}{n+1} \leq \frac{1}{n+1}$$

And the result follows. □

Chapter 7

Smooth Fictitious Play for Stochastic Games

Results of this chapter were published in the following article written with Rida Laraki:

Lucas Baudin and Rida Laraki. "Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions". In: *Advances in Neural Information Processing Systems*. 2022

In this chapter, we adapt procedures of Chapter 6 to smooth best-responses. This is first motivated by no-regret properties of SFP in repeated games, implying that procedures with smooth action selection are more likely to be robust in an adversarial context.

Moreover, our smooth procedures also have the property to explore the stochastic game, and as such can be used in a context where the game is not initially known.

7.1 Smooth Fictitious Play (SFP)

Smooth or Regularized Learning Learning using regularizers is a widespread technique in machine learning, for instance with "follow the perturbed leader" (Cesa-Bianchi and Lugosi, 2006) or so-called stochastic fictitious play (Fudenberg and Levine, 1995). It allows to have no-regret properties (Perchet, 2014; Shalev-Shwartz, 2011). From another point of view, it is a substitute to the simpler ϵ-greedy exploration scheme in reinforcement learning (Sutton and Barto, 2018).

Regularizers are classically some steep concave functions. They are added to the payoff functions u^i and can be given several interpretations (see Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002 for details): it models the uncertainty of the payoff caused by the "trembling hand" of players or is a way to generate strict incentives to explore all the actions. Formally player i maximizes a perturbation of its payoff function $u_s^i + \eta h^i$ under the following hypothesis:

Assumption 7.1.

$$h^i : \Delta(A^i) \to \mathbb{R}^+, \textit{ strictly concave in } x^i,\ C^1 \textit{ on the interior,}$$
$$\lim_{x^i \to \partial\Delta(A^i)} \|\nabla_{x^i} h^i\| = +\infty \textit{ and } \eta > 0$$

A famous variation of FP is smooth fictitious play (Fudenberg and Levine, 1995) where players choose their action according to a regularized payoff function. Formally, a player i draws an action according to a distribution that maximizes $u^i\left(\cdot, \boldsymbol{x}_n^{-i}\right) + \eta h^i\left(\cdot, \boldsymbol{x}_n^{-i}\right)$ (with h^i and η defined above). With a suitable definition of η, smooth FP has no *regret* (up to η), meaning that if played unilaterally by a player i, other players can not trick i into using a suboptimal action. Intuitively, this is in part due to the randomness of the action choice: the distribution assigns positive probability to every action, so a player's behavior remains unpredictable.

In this chapter, we extend the smooth FP procedure to stochastic games. This builds upon the recent extension of FP (Leslie et al., 2020; Sayin et al., 2022b; Baudin and Laraki, 2022a), see Section 4.3 for

more details. Indeed, it is not easy to derive the convergence to a regularized equilibrium by applying directly the definition of smooth fictitious play to the discounted stochastic game in which the players are restricted to play in stationary strategies, because the payoff function in this game is non-linear (nor is it concave, or quasi-concave) with respect to a player stationary strategy. To overcome this difficulty, and following the idea in (Leslie et al., 2020), we update two sets of variables for every state: one concerns the uncorrelated empirical actions and the other is an estimate of the continuation payoffs (i.e., payoffs that players can anticipate to achieve if that state is reached). These continuation payoffs are used as a parameter in an auxiliary game, often called the Shapley operator.

Equilibria In the following, we study the convergence of some discrete and continuous time systems to regularized stationary Nash equilibria, parameterized by the regularizers $(h^i)_{i\in I}$ and parameter η.

Definition 7.1. Regularized Stationary Nash Equilibria of a discounted stochastic game (DSG) are the stationary Nash equilibria of the DSG with the perturbed payoff functions $u_s^i + \eta h^i$, $i \in I$.

Similarly to stationary Nash equilibria, there exists at least one regularized stationary Nash equilibrium (Takahashi, 1964).

Unique Regularizer In the following, we suppose that all players have the same regularizer function, that is $\forall i \in I, h^i = h$.

Smooth Best-Response While playing the stochastic game, players maintain a set of continuation payoffs $(v_s^i)_{s\in S, i\in I}$ that are used to choose their actions, similarly to the previous chapter. Given that the current state is s, for a player i, its action is drawn from the distribution that is the smooth best-response in the auxiliary game with respect to empirical action profile $\boldsymbol{x}_s$, that is

$$\mathrm{SBR}^i_{s,v^i}\left(\boldsymbol{x}_s^{-i}\right) := \underset{y^i\in\Delta(A^i)}{\arg\max}\, f^i_{s,v^i}\left(y^i, \boldsymbol{x}_s^{-i}\right) + \eta h\left(y^i, \boldsymbol{x}_s^{-i}\right). \tag{7.1}$$

This is well and uniquely-defined because of the strict concavity of h.

Example. *If the regularizer h is taken to be the Shannon entropy, that is*

$$h(y) = -\sum_{j\in I}\sum_{a^j\in A^j} y^j(a^j)\log\left(y^j(a^j)\right).$$

Then the smooth best-response function is the logit function

$$\mathrm{SBR}^i_{s,v^i_s}\left(\boldsymbol{x}_s^{-i}\right)(a^i) := \frac{\exp(\eta^{-1} f^i_{s,v^i}\left(a^i, \boldsymbol{x}_s^{-i}\right))}{\sum_{b^i\in A^i}\exp\left(\eta^{-1} f^i_{s,v^i}\left(b^i, \boldsymbol{x}_s^{-i}\right)\right)}.$$

See Mertikopoulos and Sandholm, 2016 for other examples of regularizers.

Smooth fictitious play with known transitions and deterministic payoff To extend smooth fictitious play to stochastic games, we use two sets of variables. The $x^i_{s,n}$ variable is the distribution of empirical actions for every state s prior to step t. The other variable $v^i_{s,n+1}$ can be interpreted as the continuation payoffs starting from s and is defined as the time average of the regularized payoffs up to step t. This leads to the following system:

$$\begin{cases} \forall s \in S, v^i_{s,n+1} = \dfrac{1}{n+1}\displaystyle\sum_{k=0}^{n}\left(f^i_{s,v^i_k}\left(\boldsymbol{x}_{s,k}\right) + \eta h\left(x^i_{s,k}\right)\right) \\ a^i_{n+1} \sim \mathrm{SBR}^i_{s_{n+1},v_{n+1}}\left(\boldsymbol{x}^{-i}_{s_{n+1},n}\right) \\ \forall s \in S, x^i_{s,n+1} = \dfrac{1}{\mu_{s,n+1}}\displaystyle\sum_{k=0}^{n+1} 1_{s=s_k} a^i_k \end{cases} \tag{7.2}$$

where $\mu_{s,n}$ is the number of times s was reached, that is $\sum_{k=0}^{n} 1_{s=s_k}$ and every action a_k^i is embedded into the Euclidean space containing $\Delta(A^i)$. The interpretation is simple: at each stage t, each player best replies to the belief that the other players will play according to the past uncorrelated empirical distribution of actions, and that the future continuation payoffs are equal to the time average of the past estimated perturbed payoffs, calculated using the past empirical frequencies of actions.

This system can be generalized and rewritten in an incremental fashion, leading to our definition of Smooth Fictitious Play (SFP):

$$\begin{cases} \forall s \in S, v_{s,n+1}^i - v_{s,n}^i = \dfrac{\alpha^\star}{n+1}\left(f_{s,v_n^i}^i(\boldsymbol{x}_{s,n}) + \eta h(x_{s,n}^i) - v_{s,n}^i\right) \\ a_{n+1}^i \sim \mathrm{SBR}_{s_{n+1},v_{n+1}}^i\left(\boldsymbol{x}_{s_{n+1},n}^{-i}\right) \\ \forall s \in S, x_{s,n+1}^i - x_{s,n}^i = \dfrac{1_{s=s_{n+1}}}{\mu_{s,n+1}}\left(a_{n+1}^i - x_{s,n}^i\right) \end{cases} \tag{SFP}$$

where $\alpha^\star = 1$ corresponds to Eq. (7.2) above.

Remark. *If there is only one state (e.g., the classical repeated game setting), this procedure is exactly standard smooth fictitious play since the smooth best-response does not depend on the value of $v_{s,n}$.*

Remark. *The regularizer appears both in the smooth best-response SBR_{s,v^i}^i definition, and in the target of the updating equation of $v_{s,n}^i$. This is required as the* stage *utility function of the regularized games is regularized (and not the total, discounted utility of the stochastic game).*

Smooth fictitious play with unknown transitions and perturbed payoff We suppose now that payoff functions are not known, and that each stage payoff is observed with some zero-mean noise which follows a distribution that may depend on the history, the current state and actions taken by the other players. Therefore, at step n, player i gets a random reward R_n^i that is drawn according to a distribution determined by actions $\boldsymbol{a}_n$ and current state s_n whose expectancy is $u_{s_n}^i(\boldsymbol{a}_n)$ and bounded variance conditionally on the history, formally

$$\mathbb{E}\left[R_n^i | \mathcal{F}_{n-1}\right] = u_{s_n}^i(\boldsymbol{a}_n) \tag{7.3}$$

where $\mathcal{F}_{n-1}$ is the σ-algebra that contains all information up to step $n-1$. Furthermore, we require the variance to be bounded, i.e., $\mathrm{var}(R_n^i|\mathcal{F}_{n-1})$ is bounded.

We also suppose that transitions are not known. Therefore, both transitions and expected payoff may be empirically estimated as follows:

$$\begin{cases} \hat{P}_{ss',n}(\boldsymbol{a}) = \dfrac{\sum_{k=1}^{n} 1_{s_k=s \wedge a_k^i=\boldsymbol{a}} 1_{s_{k+1}=s'}}{\sum_{k=1}^{n} 1_{s_k=s \wedge a_k^i=\boldsymbol{a}}} \\ \hat{u}_{s,n}^i(\boldsymbol{a}) = \dfrac{\sum_{k=1}^{n} 1_{s_k=s \wedge a_k^i=\boldsymbol{a}} R_k^i}{\sum_{k=1}^{n} 1_{s_k=s \wedge a_k^i=\boldsymbol{a}}} \end{cases} \tag{MFP.1}$$

Consequently, we define the estimated auxiliary payoff using these two estimators

$$\hat{f}_{s,v,}^i(\boldsymbol{a}) := (1-\delta)\hat{u}_{s,n}^i(\boldsymbol{a}) + \delta \sum_{s' \in S} \hat{P}_{ss',n}(\boldsymbol{a}) v_{s'}^i. \tag{MFP.2}$$

Now we define a model-free version of smooth fictitious play, similar to SFP but using estimators:

$$\begin{cases} \forall s \in S, v_{s,n+1}^i - v_{s,n}^i = \dfrac{\alpha^\star}{n+1}\left(\hat{f}_{s,v_n,n}^i(\boldsymbol{x}_{i,n}) + \eta h(x_{s,n}^i) - v_{s,n}^i\right) \\ a_{n+1}^i \sim \hat{\mathrm{SBR}}_{s_{n+1},v_n}^i\left(\boldsymbol{x}_{s_{n+1},n}^{-i}\right) \\ \forall s \in S, x_{s,n+1}^i - x_{s,n}^i = \dfrac{1_{s=s_{n+1}}}{\mu_{s,n+1}}\left(a_{n+1}^i - x_{s,n}^i\right) \end{cases} \tag{MFP}$$

where $\hat{\mathrm{SBR}}_{s,v}^i$ is defined similarly to $\mathrm{SBR}_{s,v}^i$ but relatively to $\hat{f}_{s,\cdot,}^i$.

Theorem 7.1 (Convergence in identical-interest ergodic DSG)**.** *In an identical-interest ergodic discounted stochastic game, if all players follow SFP (resp. MFP), their empirical actions for all states s, $\boldsymbol{x}_{s,n}$, converge almost surely to the set of regularized stationary Nash equilibria (Definition 7.1) and their expected vector of continuation payoffs $v^i_{s,n}$ converges to the optimal continuation payoff of the limiting equilibrium set.*

This theorem means that even if the trajectories of SFP (resp. MFP) cycle between several stationary regularized equilibria, they all share the same optimal continuation payoff vector.

Theorem 7.2 (Convergence in zero-sum ergodic DSG)**.** *In a zero-sum ergodic discounted stochastic game, if all the players follow SFP (resp. MFP) with the same initial values, their empirical actions for all states s, $\boldsymbol{x}_{s,n}$, converge almost surely to the set of $D\alpha^\star$-regularized Nash equilibria (where $D > 0$ is a constant that only depends on G) and their expected vector of continuation payoffs $v^i_{s,n}$ converges to the corresponding continuation payoff.*

To prove these results, we are going to define in the next section an associated smooth best-response continuous time counterpart to our independent learning algorithms. The proof of Theorems 7.1 and 7.2 are sketched at the end of Section 7.2 and fully detailed in Section 7.3.

Remark. *Theorem 7.2 suggests that it is possible to use a doubling-trick mechanism to converge to the set of 0-regularized stationary Nash equilibria. Indeed, players can compute the duality gap (see Section 7.3 for a definition) and decide to reduce the update rate factor $\alpha^\star$ every time it is below a certain threshold. This is a standard trick also used to achieve no-regret in reinforcement learning.*

7.2 Smooth Best-Response Dynamics

Continuous counterpart of discrete time systems Similarly to Section 6.2, we define a continuous-time system analogous to SFP, that we call Smooth Best-Response Dynamics (SBRD) with

$$\forall s \in S, \begin{cases} \dot{v}^i_s = \alpha(t)\left(f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) + \eta h\big(x^i_s(t)\big)] - v^i_s(t)\right) \\ \dot{x}^i_s = \beta_s(t)\left(\mathrm{SBR}^i_{s,v_{n+1}}\big(\boldsymbol{x}^{-i}_{s,n}\big) - x^i_s(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{SBRD}$$

where α has the same properties as in Section 6.2, in particular Assumption 6.2 (page 74).

Remark. *In zero-sum or identical-interest games, similarly to previous sections, if players have the same initial conditions, then we can omit the superscript i in v^i_s as they are equal.*

Update rates Profile $x^i_s(t)$ evolve towards the smooth best-response at a rate of $\beta_s(t)$. Variable $\beta_s(t)$ corresponds to the frequency at which a state s is visited. The fact that it is bounded below by $\beta_- > 0$ is a mathematically convenient way to exploit the ergodicity of the stochastic game.

Model-Free System In order to also study the convergence of the model-free version of our procedure, we also define the smooth best-response dynamics when the model is progressively learned. The estimators are defined as follows for every action and couple of states:

$$\forall s, s' \in S, \forall \boldsymbol{b}_s \in \boldsymbol{A}, \begin{cases} \dot{\hat{P}}_{ss'}(\boldsymbol{b}_s) = \beta_s(t) a^i_s(t)(\boldsymbol{b}_s)\left(P_{s,s'}(\boldsymbol{b}_s) - \hat{P}_{ss'}(\boldsymbol{b}_s)(t)\right) \\ \dot{\hat{u}}^i_s(\boldsymbol{b}_s) = \beta_s(t) a^i_s(t)(\boldsymbol{b}_s)\left(u^i_s(\boldsymbol{b}_s) - \hat{u}^i_s(\boldsymbol{b}_s)(t)\right) \\ a^i_s(t) = \hat{\mathrm{SBR}}^i_{s,\boldsymbol{v}(n+1)}\big(\boldsymbol{x}^{-i}_s(n)\big) \end{cases} \tag{7.4}$$

where $\boldsymbol{a}_s(t) := \Pi_{i\in I} a^i_s(t)$ is the profile of selected actions and $a^i_s(t)(\boldsymbol{b}_s)$ is the joint probability to select pure profile $\boldsymbol{b}_s$ at time t.

Then, $\hat{f}^i_{s,\cdot,t}$ is defined as in (MFP.2) and we obtain a system MFBRD by using estimators in SBRD.

Solutions Systems SBRD and MFBRD are differential inclusions. A general theory can be found in Aubin and Cellina, 1984. Here, the right-hand side of each system forms a closed, set-valued map (because there are several possible values for $\beta_s(t)$) whose images are convex and uniformly bounded. Under these assumptions, it is known that such systems admit (typically non-unique) solutions.

Theorem 7.3 (Convergence of Smooth BRD in identical-interest and zero-sum games)**.** *Under Assumption 6.2, MFBRD converges to the set of regularized Nash equilibria in identical-interest stochastic games. Furthermore, if $\alpha^\star \geq \limsup_{t\to\infty} \alpha(t)$, then MFBRD converges to the set of $D\alpha^\star$-regularized stationary Nash equilibria in zero-sum games, where D is a positive constant that depends only on the game G. In particular if $\alpha(t)$ goes to 0 then MFBRD converges to the set of regularized stationary Nash equilibria.*

The convergence of MFBRD is helpful to characterize the limit set of discrete-time smooth fictitious play systems. Indeed, they are contained in the internally chain transitive sets (see Definition 7.2 below) of MFBRD using the theory of stochastic approximations. Below, we sketch the proof of continuous-time Theorem 7.3. Complete proofs of these results are in Sections 7.3.4 and 7.3.5.

Sketch of the proof of Theorem 7.3. The proofs for both class of games proceed quite differently: this is not surprising since there are no (to the best of our knowledge) unified convergence proof of even simple FP in potential and zero-sum games.

For identical-interest stochastic games, the key point is showing that the gap between (estimated) auxiliary payoffs $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ and the auxiliary values $v^i_s(t)$ is narrowing. Technically, the difference is bounded below by an (absolutely) integrable function. Since it is the differential of $v^i_s(t)$ and that $v^i_s(t)$ it, which implies that $v^i_s(t)$ converges and then that $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ converges and their limits are necessarily equal. Then, a study of the behavior of actions shows that they belong to the set of regularized equilibria, otherwise $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ could not converge (based on Lipschitz properties of all these quantities).

Regarding zero-sum stochastic games, the first part of the proof studies a rather standard quantity called the duality gap. It goes to 0, which implies first that the value of the auxiliary game is mostly reached by the auxiliary payoffs $f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t))$ and second, that the minmax strategies of the auxiliary game is learned by the players. Then, comparing relative speed of auxiliary values in all states leads to the convergence of values u^i_s and of other variables.

□

7.3 Proofs of Convergence of Best-Response Dynamics and Smooth Best-Response Dynamics

In this section, we provide detailed proofs of results of both non-smooth and smooth best-response dynamics respectively of Section 6.2 and Section 7.2. Since solutions of differential equation SBRD are special solutions of differential equation MFBRD, it is sufficient to study MFBRD. Moreover, we can generalize MFBRD to the following system

$$\forall s \in S, \begin{cases} \dot{v}_s(t) = \gamma_s(t)\alpha\left(\int_0^t \gamma_s(u)du\right)\left(f^i_{s,\boldsymbol{v}(t)}(\boldsymbol{x}_s(t)) + \eta h\big(x^i_s(t)\big) - v_s(t)\right) \\ \forall i \in I, \dot{x}^i_s(t) \in \beta_s(t)\left(\hat{\mathrm{R}}^i_{s,\boldsymbol{v}(t)}\big(\boldsymbol{x}^{-i}_s(t)\big) - x^i_s(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \\ \gamma_s(t) \in [\beta_-, 1] \end{cases} \tag{GBRD}$$

where $\hat{\mathrm{R}}$ is either the smooth best response or standard best-response and $\eta \leq 0$.

System GBRD has the property to be satisfied by solutions of MFBRD and solutions of non-smooth fictitious play, i.e., SFP, AFP and FAFP. In what follows, unless otherwise specified, $t \mapsto (x^i_s, v_s, \beta_s, \gamma_s, \hat{P}_{s,}, \hat{u}^i_{s,})_{s,i}$ is **one of those solutions** (and not any solution of GBRD).

We start with general properties of (smooth) best-response, convergence of estimates and regularity of solutions and then proceed with the study of two special cases: identical-interest stochastic games (subsection 7.3.4) and zero-sum stochastic games (subsection 7.3.5).

7.3.1 Properties of (smooth) best-responses

Lemma 7.1. *Function* $(\boldsymbol{v}, \boldsymbol{x}_s^{-i}) \mapsto \mathrm{SBR}^i_{s,\boldsymbol{v}}\left(\boldsymbol{x}_s^{-i}\right)$ *is continuous in* $\boldsymbol{v}$ *and* $\boldsymbol{x}_s^{-i}$.

Proof. It follows from a simple application of the maximum theorem that the smooth best-response is upper hemicontinuous as a set-valued map. The strict concavity of h implies that there is a single profile that maximizes the smooth auxiliary payoff. Therefore, h is continuous as a single-valued application. □

Lemma 7.2. *Under Assumption 7.1, for* $B > 0$, *there exists* $\zeta > 0$ *such that for all* $\boldsymbol{x}_s^{-i} \in \Pi_{j\in I}\Delta(A^j)$, $\boldsymbol{v} \in \mathbb{R}^S$ *such that* $\|u\|_\infty \leq B$ *and* $a_s^i \in A^i$*:*

$$\mathrm{SBR}^i_{s,\boldsymbol{v}}\left(\boldsymbol{x}_s^{-i}\right)(a_s^i) \geq \zeta$$

Proof. This is a classical property of the smooth best-response under assumptions (7.1).
Point $\mathrm{SBR}^i_{s,v_{n+1}}\left(\boldsymbol{x}_{s,n}^{-i}\right)$ is a maximum of the function $b^i \mapsto f^i_{i,\boldsymbol{v}}\left(b^i, \boldsymbol{a}^{-i}\right) + \eta h\left(b^i, \boldsymbol{a}^{-i}\right)$. However, let $\boldsymbol{a}$ be an interior point of the simplex and $(b^i, \boldsymbol{a}^{-i})$ be on the boundary for player i. Then, the composition of the linear interpolation between $(b^i, \boldsymbol{a}^{-i})$ and $\boldsymbol{a}$ and $\boldsymbol{a} \mapsto f^i_{i,\boldsymbol{v}}(\boldsymbol{a}) + \eta h(\boldsymbol{a})$ is concave because both functions are. However, the slope of this composed function is infinite (because the norm of the gradient goes to ∞ and $\boldsymbol{a}$ is interior), therefore it goes to $-\infty$ (otherwise it cannot be concave), which implies that $(b^i, \boldsymbol{a}^{-i})$ cannot be a maximum. Then, by compacity, smooth best-response are away from the boundary of the simplex. □

7.3.2 Convergence of estimates $\hat{P}_{,}$ and $\hat{u}_{,}$

Lemma 7.3. *There exists* $C > 0$ *such that for all states* s *action profiles* $\boldsymbol{b} \in A^i$ *and* $t \geq 0$,

$$|\hat{u}^i_{s,n}(\boldsymbol{b}) - u^i_s(\boldsymbol{b})| \leq C\exp(-\zeta\beta_- t),$$
$$\text{and } |\hat{P}_{ss',n}(\boldsymbol{b})(t) - P_{s,s'}(\boldsymbol{b})| \leq C\exp(-\zeta\beta_- t).$$

Proof. This is obvious when we are in a system where the model is known, since $\hat{u}^i_{,n} = u^i_{s,}$ etc. Therefore, the following is necessary for smooth systems, i.e., when $\hat{\mathrm{R}} = \mathrm{S}\,\hat{B}R$, so Lemma 7.2 holds. Let $s \in S$ and $\boldsymbol{b} \in A^i$.We write $r(t) = \left|\hat{u}^i_{s,n}(\boldsymbol{b}) - u^i_s(\boldsymbol{b})\right|$, so as $\dot{r} = -\beta_s(t)a^i(\boldsymbol{b})r(t)$.

Then we can deduce from Lemma 7.2 and the fact that $\beta_s(t) \geq \beta_-$ that $\dot{r} \leq -\zeta\beta_- r(t)$.

Therefore, Grönwall Lemma implies that $r(t) \leq r(0)\exp(-\zeta\beta_- t)$. The same proof is valid for $\hat{P}_{ss',n}$. □

7.3.3 Regularity of solutions

Lemma 7.4. *Functions* v^i_i, x^i_s, $\hat{f}^i_{s,v,}$ *are bounded.*

Proof. Because of the definition of the derivative of x^i_s, it stays in the simplex and as such it is bounded.

It is clear that $\hat{f}^i_{s,v_n,n}\left(\boldsymbol{x}_{i,n}\right) \leq (1-\delta)\|u^i_s\|_\infty + \delta\|\boldsymbol{v}^i(t)\|_\infty$. Therefore, as long as $\|v_i(t)\|_\infty$ is lower than $\|u^i_s\|_\infty$, then $\hat{f}^i_{s,v_n,n}\left(\boldsymbol{x}_{i,n}\right) \leq \|u^i_s\|_\infty$. By definition of the derivative of v^i_i, this implies that v^i_i is always smaller than $\|u^i_s\|_\infty$ assuming it is true for the initial value. □

Lemma 7.5. *Functions* v^i_s, x^i_s, $\hat{f}^i_{s,v,}$ *are Lipschitz.*

Proof. All functions are differentiable almost everywhere. The derivatives are bounded by composition since $\alpha(t)$ and $\beta_s(t)$ are bounded and because of Lemma 7.4. □

7.3.4 Convergence in identical-interest games

In identical-interest games, all payoff functions are equal. Therefore, assuming the same initial conditions for all players i for v^i_i, we have $v^i_i(t) = v^i_j(t)$ for all players i and j. So, we can omit the superscript, and the same is true for the payoff functions and the auxiliary payoff functions.

The proof proceeds as follows: we show that the differential of v^i_s becomes a good approximation of the target auxiliary payoff Γ_s, and furthermore that their difference is lower bounded by something integrable. Then we can deduce that there is a limit, for this difference and that it is necessarily 0 and the convergence of actions follows.

Convergence of payoffs

We define

$$\Gamma_s(t) := \hat{f}_{s,v(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big)$$
$$\text{and } s_-(t) \in \arg\min_{s\in S} \Gamma_s(t) - v_s(t).$$

Note that there may be several possible choices of $s_-(t)$, results below are valid for any such choice.

Lemma 7.6. *For solutions of SyncBRD, ABRD and MFBRD the differential of $\Gamma_{s_-} - v_{s_-}$ is lower bounded by, for almost every t,*

$$\frac{d\Gamma_{s_-} - v_{s_-}}{dt} \geq -2C\exp(-\zeta\beta_- t) + \alpha(t)(\delta - 1)(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t))$$

where C is defined in Lemma 7.3 and ζ in Lemma 7.2. For solutions of FABRD, we have the following inequality for almost every t,

$$\frac{d\sum_{s\in S}(\Gamma_s - v_s)_-}{dt} \geq (\delta|S| - 1)\sum_{s\in S}(\delta|S| - 1)\gamma_s(t)\alpha\left(\int_0^t \gamma_s(u)du\right)(\Gamma_s(t) - v_s(t))_-.$$

Proof. First, notice that since $\Gamma_{s_-} - v_{s_-}$ is a minimum of continuous, differentiable almost everywhere functions, it is continuous and differentiable almost everywhere. Let $t \in \mathbb{R}^+, \tau > 0$.

$$\begin{aligned}\Gamma_{s_-(t+\tau)}(t+\tau) - v_{s_-(t+\tau)}(t+\tau) - \Gamma_{s_-(t)}(t) + v_{s_-(t)}(t)&\\ = \Gamma_{s_-(t+\tau)}(t) + \tau\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t) - v_{s_-(t+\tau)}(t)&\\ - \tau\frac{dv_{s_-(t+\tau)}}{dt}(t) + o(\tau) - \Gamma_{s_-(t)}(t) + v_{s_-(t)}(t)&\end{aligned} \tag{7.5}$$

Then, for any τ:

$$\begin{aligned}\frac{dv_{s_-(t+\tau)}}{dt}(t) &= \gamma_s(t)\alpha\left(\int_0^t \gamma_s(u)du\right)\left(\Gamma_{s_-(t+\tau)}(t) - v_{s_-(t+\tau)}(t)\right)\\ &= \gamma_s(t)\alpha\left(\int_0^t \gamma_s(u)du\right)\left(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\right) + o(1)\end{aligned} \tag{7.6}$$

Moreover, for any τ:

$$\Gamma_{s_-(t+\tau)}(t) - v_{s_-(t+\tau)}(t) \geq \Gamma_{s_-(t)}(t) - v_{s_-(t)}(t) \tag{7.7}$$

Now, we need to lower bound $\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t)$.

For any $s \in S$, we use the chain rule and Lemma 7.3 to get writing $\hat{\mathrm{R}}^i_{s,\boldsymbol{v}(t)}\big(\boldsymbol{x}_s^{-i}(t)\big)$ for an element of this set (which is generically a singleton):

$$\begin{aligned}\frac{d\Gamma_s}{dt}(t) \geq \underbrace{-2C\exp(-\zeta\beta_- t)}_{\text{changes in } \hat{u} \text{ and } \hat{P},} + \underbrace{\sum_{j\in I}\beta_s(t)\langle\hat{\mathrm{R}}^j_{s,\boldsymbol{v}(t)}\big(x_s^{-j}(t)\big) - x_s^{-j}(t), \nabla_j\hat{f}_{s,v(t)}(\boldsymbol{x}_s(t)) + \eta h(\boldsymbol{x}_s(t))\rangle}_{\text{changes in } x_s^i,\ \nabla_j \text{ is the gradient with respect to } x_j^i}&\\ \underbrace{+\,\delta\sum_{s'\in S}\hat{P}_{ss',}(\boldsymbol{x}_s(t))\dot{v}_{s'}(t)}_{\text{changes in } v_s}&\end{aligned} \tag{7.8}$$

The second term is positive because $\hat{f}_{s,v} + \eta h$ is concave: $\hat{\mathrm{R}}^j_{s,\boldsymbol{v}(t)}\big(x_s^{-j}(t)\big)$ is the maximum of the fonction, therefore its gradient is null and the opposite of the gradient of a concave function is monotone. Note that in case $\hat{\mathrm{R}}^j_{s,\boldsymbol{v}(t)}\big(x_s^{-j}(t)\big)$ is not a singleton, then this is also valid.

For solutions of SyncBRD, ABRD and MFBRD Then, $\gamma_s(t) = 1$, so the third term of Eq. (7.8) is greater than

$$\delta\alpha(t)\left(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\right),$$

leading to:

$$\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t) \geq -2C\exp(-\zeta\beta_- t) + \delta\alpha(t)\left(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\right) \tag{7.9}$$

Using (7.6), (7.7) and (7.9) in (7.5) leads to:

$$\begin{aligned}&\Gamma_{s_-(t+\tau)}(t+\tau) - v_{s_-(t+\tau)}(t+\tau) - \Gamma_{s_-(t)}(t) + v_{s_-(t)}(t)\\ &\geq \tau\left(-2C\exp(-\zeta\beta_- t) + \alpha(t)(\delta-1)\left(\Gamma_{s_-(t)}(t) - v_{s_-(t)}(t)\right)\right) + o(\tau)\end{aligned}$$

And the result follows.

For solutions of FABRD The third term of Eq. (7.8) is

$$\delta\sum_{s'\in S}\hat{P}_{ss'},(\boldsymbol{x}_s(t))\dot{v}_{s'}(t) \geq \delta\sum_{s'\in S}\hat{P}_{ss'},(\boldsymbol{x}_s(t))\zeta_{s'}(t)(\Gamma_{s'}(t) - v_{s'}(t))_- \tag{7.10}$$

where $\zeta_{s'}(t) = \gamma_{s'}(t)\alpha\Big(\int_0^t \gamma_{s'}(u)du\Big)$. Summing over s, this leads to

$$\begin{aligned}\sum_{s\in S}\frac{d(\Gamma_s - v_s)_-}{dt} &\geq \sum_{s\in S}1_{\Gamma_s(t)-v_s(t)<0}\delta\sum_{s'\in S}\hat{P}_{ss'},(\boldsymbol{x}_s(t))\zeta_{s'}(t)(\Gamma_{s'}(t) - v_{s'}(t))_-\\ &\quad - \sum_{s\in S}1_{\Gamma_s(t)-v_s(t)<0}\zeta_s(t)(\Gamma_s(t) - v_s(t))\\ &= \sum_{s\in S}\left(\sum_{s'\in S}1_{\Gamma_{s'}(t)-v_{s'}(t)<0}\delta\sum_{s\in S}\hat{P}_{s's},(\boldsymbol{x}_s(t)) - 1_{\Gamma_s(t)-v_s(t)<0}\right)\zeta_{s'}(t)(\Gamma_s(t) - v_s(t))_-\\ &= \sum_{s\in S}\left(\sum_{s'\in S}1_{\Gamma_{s'}(t)-v_{s'}(t)<0}\delta\sum_{s\in S}\hat{P}_{s's},(\boldsymbol{x}_s(t)) - 1\right)\zeta_{s'}(t)(\Gamma_s(t) - v_s(t))_-\\ &= \sum_{s\in S}\left(\sum_{s'\in S}\delta|S| - 1\right)\zeta_{s'}(t)(\Gamma_s(t) - v_s(t))_-\end{aligned}$$

□

Lemma 7.7. *Under Assumptions 6.2 and 7.1, for solutions of SyncBRD, ABRD and MFBRD, there exists $\boldsymbol{v}_\infty \in \mathbb{R}^S$ such that for all $s \in S$, $v_s(t) \to v_{\infty,s}$ and $\Gamma_s(t) \to v_{\infty,s}$.*

Proof. We consider the following differential equation:

$$y' = -2C\exp(-\zeta\beta_- t) + \alpha(t)(\delta-1)y \tag{7.11}$$

Lemma 7.6 implies that solutions of (7.11) with initial condition $y(0) = \Gamma_{s_-(0)}(0) - v_{s_-(0)}(0)$ lower bound $\Gamma_{s_-} - v_{s_-}$.

Moreover, solutions of (7.11) are of the form:

$$\left(y(0) - \int_0^t 2C\exp(-\zeta\beta_- v)\exp\left(\int_0^v \alpha(u)(1-\delta)du\right)dv\right)\exp\left(-\int_0^t \alpha(u)(1-\delta)du\right)$$

which is equal to:

$$y(0)\exp\left(-\int_0^t \alpha(u)(1-\delta)du\right) - \int_0^t 2C\exp(-\zeta\beta_- v)\exp\left(-\int_v^t \alpha(u)(1-\delta)du\right)dv$$

Which is greater than:

$$-|y(0)|\exp\left(-\int_0^t \alpha(u)(1-\delta)du\right) - \int_0^t 2C\exp(-\zeta\beta_- v)\exp\left(-\int_v^t \alpha(u)(1-\delta)du\right)dv \tag{7.12}$$

So for every s, $\Gamma_s(t) - v_s(t)$ is greater than (7.12). We study the differential of $v_s(t)$, that is $\alpha(t)(\Gamma_s(t) - v_s(t))$ and show that it is lower bounded by an integrable quantity:

$$\dot{v}_s \geq -\alpha(t)|y(0)| \exp\left(-\int_0^t \alpha(u)(1-\delta)du\right) - \alpha(t)\int_0^t 2C\exp(-\zeta\beta_- v)\exp\left(-\int_v^t \alpha(u)(1-\delta)du\right)dv \quad (7.13)$$

The integral of the first term is:

$$\int_0^t -\alpha(v)|y(0)|\exp\left(-\int_0^v \alpha(u)(1-\delta)du\right) = -|y(0)|\frac{1}{1-\delta}\left(1-\exp\left(-\int_0^t \alpha(u)(1-\delta)du\right)\right) > -\infty$$

And the integral of the second term is:

$$\begin{aligned}
&\int_0^t -\alpha(v)\int_0^v 2C\exp(-\zeta\beta_- w)\exp\left(-\int_w^v \alpha(u)(1-\delta)du\right)dwdv \\
&= \int_0^t\int_0^t -1_{w\leq v}\alpha(v)2C\exp(-\zeta\beta_- w)\exp\left(-\int_w^v \alpha(u)(1-\delta)du\right)dwdv \\
&= \int_0^t\int_w^t -\alpha(v)2C\exp(-\zeta\beta_- w)\exp\left(-\int_w^v \alpha(u)(1-\delta)du\right)dvdw \\
&= \int_0^t 2C\exp(-\zeta\beta_- w)\int_w^t -\alpha(v)\exp\left(-\int_w^v \alpha(u)(1-\delta)du\right)dvdw \\
&= \int_0^t 2C\exp(-\zeta\beta_- w)\left[\frac{1}{1-\delta}\exp\left(-\int_w^v \alpha(u)(1-\delta)du\right)\right]_w^t dw \\
&= \int_0^t 2C\exp(-\zeta\beta_- w)\frac{1}{1-\delta}\left(\exp\left(-\int_w^t \alpha(u)(1-\delta)du\right)-1\right)dw \\
&\geq -\frac{1}{1-\delta}\frac{2C}{\zeta\beta_-}(1-\exp(-\zeta\beta_- t)) > -\infty
\end{aligned}$$

Therefore, both term of the integral of (7.13) are greater than $-\infty$. Since $v_s(t)$ is bounded (Lemma 7.4) and its derivative is $\alpha(t)(\Gamma_s(t) - v_s(t))$, then $v_s(t)$ converges to its $\limsup$. Moreover, the same reasoning can be made with Γ_s (see its differentiation in (7.8)), so it has a limit which is necessarily the same as v_s: otherwise, the derivative of $v_s(t)$ would converge towards $\alpha(t)$ times the difference of the limits and $v_s(t)$ would be unbounded because of Assumption 6.2. □

Convergence of actions

Lemma 7.8. *Under Assumptions 6.2 and 7.1, for solutions of SyncBRD, ABRD and MFBRD, action profile $x^i(s)$ converges to a fixpoint of $\hat{\mathrm{R}}_{s,v_\infty}$. Therefore, $\boldsymbol{x}(t)$ converges to a regularized Nash equilibria of the stochastic game.*

Proof. We use equation (7.8) to bound above the scalar product:

$$\langle \hat{\mathrm{R}}^i_{s,\boldsymbol{v}(t)}\left(\boldsymbol{x}_s^{-i}(t)\right) - x_i^i(t), \nabla_j \hat{f}_{s,v(t)}\left(x_s^i(t)\right) + \eta h\left(x_s^i(t)\right)\rangle \quad (7.14)$$

Since Γ_s and v_s have limits and are Lipschitz (Lemma 7.5), then their derivative goes to 0, and the first term of (7.8) also goes to 0. As a consequence, the limsup of (7.14) is bounded above by 0, and this is positive, so it goes to 0. Therefore, the action profile converges to a fixpoint of $\hat{\mathrm{R}}_{s,v_\infty}$. □

Proofs of non-smooth best-response dynamics

Proof of Theorem 6.3.

- Item (i) is Lemma 7.7.
- Items (ii) and (iii) are Lemma 7.8.
- Regarding FABRD, if $\delta < \frac{1}{|S|}$, Lemma 7.6 can be used to show that $(\sum_{s\in S} \Gamma_s(t) - v_s(t))_-$ goes to 0 like in Lemma 7.7. Subsequent proofs are the same.

□

Proofs of smooth best-response dynamics

Proof of Theorem 7.3 for identical-interest stochastic games. The proof is the combination of Lemmas 7.7 and 7.8. □

7.3.5 Convergence in zero-sum games

In this subsection we suppose that the game is zero-sum, that is there are only two players and for all states, $u_s^1 = -u_s^2$. Therefore, we can omit the superscript with the convention that $u_s = u_s^1$.

Moreover, we suppose that both players use the same regularizer:

$$h\big(x_s^1, x_s^2\big) = h^1\big(x_s^1, x_s^2\big) = -h^2\big(x_s^1, x_s^2\big) \tag{7.15}$$

where both h^1 and h^2 are concave in respectively x_s^1 and x_s^2.

Therefore, with the same initial conditions, continuation payoffs v_s are opposite for both players, and we also omit the superscript.

The proof is inspired from Leslie et al., 2020.

Convergence of payoffs

We define the energy of the system, also known as the duality gap (Benaim et al., 2005):

$$\begin{aligned} w_s(t) = &\max_{b^1\in\Delta(A^1)} \hat{f}_{s,\boldsymbol{v}(t)}\big(b^1, x_2^i(t)\big) + \eta h\big(b^1, x_2^i(t)\big) \\ &- \min_{b^2\in\Delta(A^2)} \hat{f}_{s,\boldsymbol{v}(t)}\big(x_1^i(t), b^2\big) + \eta h\big(x_1^i(t), b^2\big) \end{aligned} \tag{7.16}$$

This is a positive quantity because the first (second) term is greater (lower) than $\hat{f}_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big)$. We are going to show that it is mostly decreasing.

In the rest of the section, we use a $\alpha^\star$ such that:

$$\alpha^\star > \limsup \alpha(t) \tag{H3}$$

A special case is when $\alpha(t)$ goes to 0, in this case $\alpha^\star$ can be taken arbitrarily small.

Lemma 7.9. *The differential of w_s is bounded:*

$$\frac{dw_s}{dt} \leq -\beta_- w_s(t) + D\alpha^\star + D\exp\left(-\zeta\beta_- t\right)$$

Proof. We write $y_s^{i\star}(t) = \mathrm{SBR}^s_{v_s(t),x_s^i(t)}(i)$ and:

$$g(y_s^{1\star}(t)) := \hat{f}_{s,\boldsymbol{v}(t)}\big(y_s^{1\star}(t), x_2^i(t)\big) + \eta h\big(y_s^{1\star}(t), x_2^i(t)\big)$$

where the dependency on $v_s, \hat{P}_{s,}, \hat{r}_s, x_s^i$ is left implicit. This implies:

$$g(y_s^{1\star}(t)) = \max_{b^1\in\Delta(A^1)} \hat{f}_{s,\boldsymbol{v}(t)}\big(b^1, x_2^i(t)\big) + \eta h\big(b^1, x_2^i(t)\big)$$

Therefore, using the envelope theorem, the derivative of g is written using only derivatives on all variables but $y_s^{1\star}(t)$:

$$\frac{dg \circ y_s^{1\star}}{dt} = D_v g \cdot \dot{v} + D_{x_s^2} g \cdot \dot{x}_2^i + D_{\hat{r}_s} g \cdot \dot{\hat{u}}_s + D_{\hat{P}_{s,}} g \cdot \dot{\hat{P}}_s$$

With Lemma 7.3, Lemma 7.4 and Hypothesis H3,

$$\frac{dg \circ y_s^{1\star}}{dt} \leq \alpha^\star \|D_v g\| \|r_s\| + D_{x_s^2} g \cdot \dot{x}_2^i + (\|D_{\hat{r}_s} g\| + \|D_{\hat{P}_{s,}} g\|)\beta_s(t)\zeta C \exp(-\zeta\beta_- t) \tag{7.17}$$

$\hat{f}_{s,\boldsymbol{v}(t)}$ is linear, so:

$$D_{x_s^2} g \cdot \dot{x}_2^i = \hat{f}_{s,\boldsymbol{v}(t)}\big(y_s^{1\star}(t), \dot{x}_2^i\big) + \eta \nabla_{x_2^i} h\big(y_s^{1\star}(t), x_2^i(t)\big) \cdot \dot{x}_2^i \tag{7.18}$$

And since $h = h^1 = -h^2$ (with (7.15)), $\nabla_{x_2^i} h = -\nabla_{x_2^i} h^2$ by definition, and h^2 is concave in x_s^2, so:

$$\nabla_{x_2^i} h^2\big(y_s^{1\star}(t), x_2^i(t)\big) \cdot (y_s^{2\star}(t) - x_2^i(t)) \geq h^2\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) - h^2\big(y_s^{1\star}(t), x_2^i(t)\big) \tag{7.19}$$

It follows from (7.18), (7.19), (7.15) and $\dot{x}_2^i = \beta_s(t)(y_s^{2\star}(t) - x_2^i(t))$ that

$$\begin{aligned} D_{x_s^2} g \cdot \dot{x}_2^i &\leq \beta_s(t)\hat{f}_{s,\boldsymbol{v}(t)}\big(y_s^{1\star}(t), y_s^{2\star}(t) - x_2^i(t)\big) + \eta\beta_s(t) h\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) - \beta_s(t)\eta h\big(y_s^{1\star}(t), x_2^i(t)\big) \\ &\leq -\beta_s(t) g(y_s^{1\star}(t)) + \beta_s(t)\hat{f}_{s,\boldsymbol{v}(t)}\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) + \eta\beta_s(t) h\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) \end{aligned} \tag{7.20}$$

Since g as a function of $v_s, \hat{P}_{s,}, \hat{r}_s, x_s^i$ is Lipschitz, and using (7.17) and (7.20), there exists D such that:

$$\begin{aligned} \frac{dg \circ y_s^{1\star}}{dt} \leq \frac{D}{2}\alpha^\star + \frac{D}{2}\beta_s(t)\exp(-\zeta\beta_- t) - \beta_s(t) g(y_s^{1\star}(t)) \\ + \beta_s(t)\hat{f}_{s,\boldsymbol{v}(t)}\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) + \eta\beta_s(t) h\big(y_s^{1\star}(t), y_s^{2\star}(t)\big) \end{aligned} \tag{7.21}$$

The same reasoning apply for the second term of w_s with the opposite payoff function (notice that in this case, the second line of (7.21) is exactly the opposite, therefore when summed it cancels out), leading to, when summed:

$$\begin{aligned} \frac{dw_s}{dt} &\leq D\alpha^\star + D\beta_s(t)\exp(-\zeta\beta_- t) - \beta_s(t) w_s(t) \\ &\leq D\alpha^\star + D\exp(-\zeta\beta_- t) - \beta_- w_s(t) \end{aligned}$$

because $\beta_- \leq \beta_s(t) \leq 1$ and $w_s \geq 0$ (its first term is greater than $\hat{f}_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big)$ and its second one is lower). □

The following lemma implies that the auxiliary payoff is close to the value of the auxiliary game because the duality gap of the auxiliary game is small enough.

Lemma 7.10. *For all states $s \in S$,*

$$\limsup w_s(t) \leq 2D\alpha^\star\beta_-^{-1}$$

Furthermore, $\max\{w_s - 2D\alpha^\star\beta_-^{-1}, 0\}$ is a Lyapunov function of SBRD (i.e., when payoff and transitions estimate are exact) in the autonomous case.

Proof. Since w_s is positive (the first term is greater than $\hat{f}_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big)$ and the second one is lower), Lemma 7.9 makes it possible to use Grönwall's Lemma on $w_s - 2D\alpha^\star\beta_-^{-1}$ as soon as $\exp(-\zeta\beta_- t) \leq \alpha^\star$.

□

Define ξ such as $\frac{(1-\delta)\xi}{16} = 4D\alpha^\star\beta_-^{-1}$.

Estimates of transitions and payoffs are close to real values for t large enough, so Lemma 7.10 implies that there exists $t_1(\xi)$ such that for $t \geq t_1(\xi)$:

$$\begin{aligned} |\hat{f}_{s,\boldsymbol{v}} - f_{s,\boldsymbol{v}}| &\leq 4D\alpha^\star\beta_-^{-1} = \frac{(1-\delta)\xi}{16} \\ |\hat{f}_{s,\boldsymbol{v}} + \eta h - v_{s,u(t)}| &\leq 2D\alpha^\star\beta_-^{-1} = \frac{(1-\delta)\xi}{32} \\ |f_{s,\boldsymbol{v}} + \eta h - v_{s,u(t)}| &\leq 2D\alpha^\star\beta_-^{-1} = \frac{(1-\delta)\xi}{32} \end{aligned} \tag{A1}$$

where $v_{s,u(t)}$ is the value of the auxiliary game parameterized by $u(t)$ (and functions h, $f_{s,\boldsymbol{v}}\big(x_s^i\big)$ are $\hat{f}_{s,\boldsymbol{v}}\big(x_s^i\big)$ are valued at $x_s^i(t)$, omitted for readability).

We define two distinguished states (notice that we use $f_{s,\boldsymbol{v}}$ and not $\hat{f}_{s,\boldsymbol{v}}$):

- $s_f(t) \in \arg\max_{s\in S} |f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) - v_s(t)|$
- $s_v(t) \in \arg\max_{s\in S} |v_{s,u(t)} - v_s(t)|$

Lemma 7.11. *If (A1) is satisfied (for instance if $t \geq t_1(\xi)$) and*

$$|v_{s_f}(t) - f_{s_f(t),\boldsymbol{v}(t)}\Big(x^{s_f(t)}(t)\Big) - \eta h\Big(x^{s_f(t)}(t)\Big)| \geq \xi$$

and for an $s \in S$,

$$\big||v_{s_f}(t) - v_{s_f,u(t)}| - |v_s(s) - v_{s,u(t)}|\big| \leq \frac{(1-\delta)\xi}{8}$$

then:

$$\frac{d|v_s(s) - v_{s,u(t)}|}{dt} \leq -\frac{(1-\delta)\alpha(t)\xi}{2}$$

Proof. First, using Lemma A.2 of Leslie et al., 2020 on the regularity of the value of a zero-sum (static) game, it follows:

$$\begin{aligned} \left|\frac{dv_{s,u(t)}}{dt}\right| &\leq \delta \max_{s\in S} |\dot{v}_s| \\ &= \delta\alpha(t)|f_{s_f(t),\boldsymbol{v}(t)}\Big(x^{s_f(t)}(t)\Big) + \eta h\Big(x^{s_f(t)}(t)\Big) - v_{s_f}(t)| + \delta\alpha(t)\frac{(1-\delta)\xi}{16} \\ &\leq \delta\alpha(t)\xi(1 + \frac{1-\delta}{16}) \end{aligned} \tag{7.22}$$

We now prove that $v_s(t)$ moves towards $v_{s,u(t)}$ at a constant speed relatively to $\alpha(t)$:

- If $u_s(t) \geq v_{s,u(t)}$, then $|v_{s_f}(t) - v_{s_f,u(t)}| - v_s(t) + v_{s,u(t)} \leq \frac{(1-\delta)\xi}{8}$.

$$\begin{aligned} \dot{v}_s &= \alpha(t)\left(\hat{f}_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) - v_s(t)\right) \\ &\leq \alpha(t)\left(f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) - v_s(t)\right) + \alpha(t)\frac{(1-\delta)\xi}{16} \\ &\leq \alpha(t)\left(f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) + \frac{(1-\delta)\xi}{8} - v_{s,u(t)} - |v_{s_f}(t) - v_{s_f,u(t)}|\right) \\ &\quad + \alpha(t)\frac{(1-\delta)\xi}{16} \\ &\leq \alpha(t)\left(\frac{4(1-\delta)\xi}{16} - |v_{s_f}(t) - v_{s_f,u(t)}|\right) \\ &\leq \alpha(t)\left(\frac{(1-\delta)\xi}{4} - |v_{s_f}(t) - f_{s_f(t),\boldsymbol{v}(t)}\Big(x^{s_f(t)}(t)\Big) - \eta h\Big(x^{s_f(t)}(t)\Big)|\right) \\ &\leq \alpha(t)\left(\frac{(1-\delta)\xi}{4} - \xi\right) \end{aligned}$$

Summing with $v_{s,u(t)}$ and using (7.22):

$$\begin{aligned}\frac{dv_s(t) - v_{s,u(t)}}{dt} &\leq \alpha(t)\left(\frac{(1-\delta)\xi}{4} - \xi + \delta\xi + \delta\xi\frac{1-\delta}{16}\right)\\ &\leq \alpha(t)\xi\left(\frac{1-\delta}{2} - 1 + \delta\right)\\ &\leq -\alpha(t)\left(\frac{2(1-\delta)\xi}{4}\right)\end{aligned}$$

- If $u_s(t) \leq v_{s,u(t)}$, similar calculations yield the same result.

□

Lemma 7.12. *For all $s \in S$, $\limsup_{t\to\infty} |v_s(t) - f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) - \eta h\big(x_s^i(t)\big)| \leq 3\xi$.*

Proof. We define $g(t) = \max\{|v_{s_f}(t) - v_{s_f,u(t)}|, 3\xi\}$.

Now, if $|v_{s_f}(t) - v_{s_f,u(t)}| \leq 2\xi$, then $\frac{dg}{dt} = 0$. If $|v_{s_f}(t) - v_{s_f,u(t)}| \geq 2\xi$ and if t is greater than $t^1(\xi)$, then $|v_{s_f}(t) - f_{s_f(t),\boldsymbol{v}(t)}\big(x^{s_f(t)}(t)\big) - \eta h\big(x^{s_f(t)}(t)\big)| \geq \xi$: indeed, $|f_{s_f(t),\boldsymbol{v}(t)}\big(x^{s_f(t)}(t)\big) + \eta h\big(x^{s_f(t)}(t)\big) - v_{s_f,u(t)}| \leq \xi$ because of Lemma 7.10 and its corollary A1. Similarly, on a neighborhood of t, every s that maximizes $|f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) - v_s(t)|$ satisfies the condition of Lemma 7.11, because $|f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big) + \eta h\big(x_s^i(t)\big) - v_{s,u(t)}| \leq \xi$ according to the same Lemma 7.10. Therefore, Lemma 7.11 can be used and:

$$\frac{dg}{dt} \leq -\frac{3(1-\delta)\alpha(t)\xi}{4}$$

This holds as soon as $g(t) > 2\xi$ and $t > t^1(\xi)$. The integral of α is infinite (Assumption 6.2), so there is a $t^2(\xi)$ such that for $t \geq t^2(\xi)$, $g(t) = 2\xi$.

Then, using A1, we have $|v_{s_f}(t) - f_{s_f(t),\boldsymbol{v}(t)}\big(x^{s_f(t)}(t)\big) - \eta h\big(x^{s_f(t)}(t)\big)| \leq 3\xi$ and by definition of s_f, the inequality of the lemma.

□

Convergence of actions

Lemma 7.13. *For all $s \in S$, $x_s(t)$ converge to the set of 3ξ-Regularized Nash equilibria of the auxiliary game.*

Proof. The previous proof gives that $f_{s,\boldsymbol{v}(t)}\big(x_s^i(t)\big)$ is 3ξ close to $v_{s,u(t)}$, hence the result. □

Proof for non-smooth best-response dynamics

Proof of Theorem 6.4.

- Item (i) is Lemma 7.12.
- Item (ii) is Lemma 7.13.

□

Proof for non-smooth best-response dynamics

Proof of Theorem 7.3 for zero-sum stochastic games. This is a combination of Lemmas 7.12 and 7.13. □

7.4 Proofs of Convergence Fictitious Play in Discrete Time

In this section, we describe how the stochastic approximation framework with differential inclusion (Benaim et al., 2005) can be extended and used to prove result in discrete time in the autonomous case (i.e., α is constant).

7.4.1 Correlated Asynchronous Stochastic Approximation

In this subsection, we expose the stochastic approximation theorem of Perkins and Leslie (2012), which relates the limiting behavior of a discrete-time sequence $(y_n)_{n\in\mathbb{N}} \in (\mathbb{R}^K)^{\mathbb{N}}$ and a function $y : \mathbb{R}^+ \to \mathbb{R}^K$. Note that we are in ambient space $\mathbb{R}^K$: in systems described in this chapter and the previous one, when the model is known, K is equal to $2|S|$ where S is the set of states because we have $|S|$ variables to estimate empirical actions and $|S|$ variables to estimate continuation payoffs. When the model is not known, then there are additional variable to estimate the model and K is even larger.

When there is no ambiguity, we write $s \in K$ for $s \in [1 \dots K]$.

Discrete-Time System An asynchronous system as defined by Perkins and Leslie (2012) is as follows. Assuming $y_n \in \mathbb{R}^K$, one defines a system where updated components of the vector at every step n are $S_n \subseteq K$. We define $\mu_{s,n}$ as the number of times until n that s occurred, that is

$$\mu_{s,n} = \sharp\{u \mid s \in S_u \wedge 0 \leq u \leq n\}. \tag{7.23}$$

We now describe a system where component $y_{s,n}$ is updated at rate $\gamma_{\mu_{s,n}}$ if and only if $s \in S_n$,

$$y_{s,n+1} - y_{s,n} - \gamma_{\mu_{s,n}}(Y_{s,n} + d_{s,n}) \in 1_{s\in S_n}\gamma_{\mu_{s,n}}F_s(y_n) \tag{7.24}$$

where variable $Y_{s,n}$ is a random noise with $\mathbb{E}[Y_{s,n}] = 0$, $d_{s,n}$ goes to 0 when $n \to \infty$ and (γ_n) is a sequence of decreasing update steps typically satisfying Assumption 6.1.

In order to rewrite Eq. (7.24) in a vector form, we define the relative update rate and a diagonal matrix with update rates,

$$\overline{\gamma}_n = \max_{s\in S_n} \gamma_{\mu_{s,n}}, \tag{7.25}$$

$$\text{and } M_{n+1} = \text{diag}\left\{1_{s\in I_n}\frac{\gamma_{\mu_{s,n}}}{\overline{\gamma}_n} \mid s \in K\right\}. \tag{7.26}$$

Then, Eq. (7.24) can be rewritten as

$$y_{n+1} - y_n - \overline{\gamma}_n M_{n+1}(Y_n + d_n) \in \overline{\gamma}_n M_{n+1}F(y_n) \tag{7.27}$$

where y_n is the vector $(y_{s,n})_{s\in K}$.

Continuous-Time System The continuous counterpart is defined as follows. For an $\epsilon > 0$, Ω_K^ϵ is the set of $K \times K$ diagonal matrices with coefficients between ϵ and 1,

$$\Omega_K^\epsilon := \{\text{diag}(\beta_1, \dots, \beta_K); \beta_i \in [\epsilon, 1], \forall i = 1, \dots, K\}.$$

And the continuous system is

$$\frac{dy}{dt} \in \overline{F}(y) := \Omega_K^\epsilon \cdot F(y) \tag{7.28}$$

where the multiplication is between sets (i.e., the resulting set is the multiplication of every pair of the initial sets).

Limit Sets Then, Perkins and Leslie (2012) showed that the limit set of solutions of Eq. (7.27) is internally chain transitive (see Definition 7.2 below) for system Eq. (7.28) under assumptions stated below.

Using this theorem, we are going to link the internally chain transitive sets of differential inclusion $\frac{dy}{dt} \in \overline{F}(y)$ and limit sets of solutions of (7.27). As systems ABRD and SBRD can be written as $\overline{F}$ with a suitable $\mathcal{S}$ and F, making it possible to prove the rest of Theorem 6.1 using the convergence results of the continuous time systems of the previous section, see section 7.4.2.

Definitions We start with the classic definition of Marchaud maps in the stochastic approximation framework. In our systems, as the best-response map BR is piecewise constant (and SBR is continuous, see Lemma 7.1) and the rest of the right-hand side is continuous, right-hand sides of the differential inclusions are Marchaud maps.

Definition 4.1 (Marchaud map)**.** $F : \mathbb{R}^K \rightrightarrows \mathbb{R}^K$ is a Marchaud map if:

(i) F is a closed set-valued map, i.e., $\{(y, z) \in \mathbb{R}^K \times \mathbb{R}^K \mid z \in F(y)\}$ is closed.

(ii) for all $y \in \mathbb{R}^K$, $F(y)$ is a non-empty, compact, convex subset of $\mathbb{R}^K$

(iii) there exists $c > 0$ such that $\sup_{y \in \mathbb{R}^K z \in F(y)} ||z|| \leq c(1 + ||y||)$

We now need the definition of internally chain transitive sets, as stated in Benaim et al., 2005. They will later be used to characterize the limit sets of the discrete-time systems.

Definition 7.2 (Internally chain transitive)**.** A set A is internally chain transitive for a differential inclusion $\frac{dy}{dt} \in F(y)$ if it is compact and if for all $y, y' \in A$, $\epsilon > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $y_1, \dots y_n$ to the differential inclusion and real numbers $t_1, t_2, \dots, t_n$ greater than T such that:

- $y_i(u) \in A$ for $0 \leq u \leq t_i$
- $\|y_i(t_i) - y_{i+1}(0)\| \leq \epsilon$
- $\|y_1(0) - y\| \leq \epsilon$ and $\|y_n(t_n) - y'\| \leq \epsilon$

Definition 7.3 (Asymptotic pseudo-trajectories)**.** A continuous function $z : \mathbb{R}^+ \to \mathbb{R}^m$ is an asymptotic pseudo-trajectory of a differential inclusion if $\lim_{t \to +\infty} \mathbf{D}(\Theta^t(z), \mathcal{E}) = 0$ where $\Theta^t(z)(s) = z(t + s)$ (it is the translation operator), $\mathcal{E}$ is the set of all solutions of the differential inclusion and $\mathbf{D}$ is the distance between continuous functions defined as

$$\mathbf{D}(f, g) := \sum_{k=1}^{\infty} \frac{1}{2^k} \min(\|f - g\|_{[-k,k]}, 1)$$

where $\| \cdot \|_{[-k,k]}$ is the supremum norm on the interval $[-k, k]$.

This two last definitions will be useful with Theorem 4.3 of (Benaim et al., 2005) that establishes that the limit set of asymptotic pseudo-trajectories is internally chain transitive. What is left to prove is that an affine interpolation of the discrete time system is an asymptotic pseudo-trajectories.

Correlated Asynchronous Stochastic Approximations We now state a theorem proven by Perkins and Leslie (2012), which is written here for the sake of completeness, the consistency of notations and to underline the importance of the set $\mathcal{S}$.

Theorem 7.4 (Theorem 3.1 of Perkins and Leslie, 2012)**.** *Suppose that:*

(i) $y_n \in C$ *for all* n *where* C *is compact*

(ii) The set valued application $F : C \rightrightarrows C$ *is Marchaud*

(iii) Sequence γ_n *is such that*

  (a) $\sum_n \gamma_n = \infty$ *and* $\gamma_n \xrightarrow[n \to \infty]{} 0$

  (b) for $z \in (0, 1), \sup_n \gamma_{[zn]}/\gamma_n < A_z < \infty$ *where* $[\cdot]$ *is the floor function and* A_z *is a constant that only depends on* z*.*

  (c) for all n, $\gamma_n \geq \gamma_{n+1}$

(iv) (a) For all $y \in C$, $\mathcal{S}_n, \mathcal{S}_{n+1} \in \mathcal{S}$,

$$\mathbb{P}(S_{n+1} = \mathcal{S}_{n+1} | \mathcal{F}_n) = \mathbb{P}(S_{n+1} = \mathcal{S}_{n+1} | S_n = \mathcal{S}_n, y_n = y)$$

(b) The probability transition between $\mathcal{S}_n$ and $\mathcal{S}_{n+1}$ is Lipschitz continuous in x_n and the Markov chain that S_n form is aperiodic, irreducible and for every $s \in \mathcal{S}$, there exists $S \in \mathcal{S}$ such that $s \in S$.

(v) For all n, Y_{n+1} and S_{n+1} are uncorrelated given $\mathcal{F}_n$

(vi) For some $q \geq 2$, $\begin{cases} \sum_n \gamma_n^{1+q/2} < \infty \\ \sup_n \mathbb{E}(\|Y_n\|^q) < \infty \end{cases}$

(vii) $d_n \to 0$ when $n \to \infty$

Then with probability 1, affine interpolation $\overline{y}$ is an asymptotic pseudo-trajectory to the differential inclusion,

$$\frac{dy}{dt} \in \overline{F}(y)$$

where $\begin{cases} \overline{F}(y) := \Omega^{\epsilon}_{k,\mathcal{S}} \cdot F(y) \\ \Omega^{\epsilon}_{k,\mathcal{S}} := \mathit{diag}(\mathit{conv}(\mathcal{S}) \bigcap [\epsilon, 1]^K) \\ \epsilon > 0 \end{cases}$

7.4.2 Convergence of FP and SFP in identical-interest stochastic games

In this subsection, we first characterize the internally chain transitive sets of $SyncBRD$ and $ABRD$ before using this characterization to prove the convergence of FP in identical interest stochastic games.

Lemma 7.14 (Internally Chain Transitive Sets for ABRD and SyncBRD)**.** *If for all t, $\alpha(t) = 1$ and if L is internally chain transitive either for ABRD or SyncBRD then*

$$L \subseteq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \,\middle|\, \forall s \in S, \forall i \in I,\ f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s) = v_s \wedge x^i_s \in \arg\max_{y^i \in A^i} f^i_{s,\boldsymbol{v}}\left(y^i, \boldsymbol{x}^{-i}_s\right) \right\}.$$

Proof. We show the result for ABRD. We define:

$$A := \left\{ (\boldsymbol{x}, \boldsymbol{v}) \,\middle|\, \forall s \in S, \forall i \in I,\ f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s) = v_s \wedge x^i_s \in \arg\max_{y^i \in A^i} f^i_{s,\boldsymbol{v}}\left(y^i, \boldsymbol{x}^{-i}_s\right) \right\}$$
$$B := \left\{ (\boldsymbol{x}, \boldsymbol{v}) \,\middle|\, \forall s \in S, f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s) \geq v_s \right\}$$

We first show that $L \subseteq B$. In order to do that, we take an element of L and show that any path starting from this element is brought towards B, leading to the fact that the element is necessarily already in B (by definition of internal chain transitivity).

Let $(\boldsymbol{x}, \boldsymbol{v}) \in L$ and suppose that $(\boldsymbol{x}, \boldsymbol{v}) \notin B$, that is:

$$-\xi := \min_{s \in S} f^i_{s,\boldsymbol{v}}(\boldsymbol{x}) - v_s < 0$$

Then for the case of SBRD, for any $T > 0$, there exists $m \in \mathbb{N}$, solutions of SBRD $(\boldsymbol{x}_1, \boldsymbol{v}_1), \ldots (\boldsymbol{x}_m, \boldsymbol{v}_m)$ and $t_1, \ldots, t_m$ greater than T as in Definition 7.2 for $\epsilon = \xi/2$.

Then $\min_{s \in S} f^i_{s,\boldsymbol{v}_1(0)}(\boldsymbol{x}_{1,s}(0)) - v_{1,s}(0) \geq -\xi - \xi/2$.

Now we can use Eq. (7.13) with $\alpha(t) = 1$ and $C = 0$, for all s:

$$f^i_{s,\boldsymbol{v}_1(t_1)}(\boldsymbol{x}_{1,s}(t_1)) - v_{1,s}(t_1) \geq (f^i_{s,\boldsymbol{v}_1(t_1)}(\boldsymbol{x}_{1,s}(t_1)) - v_{1,s}(t_1)) \exp((\delta - 1)t_1) \geq (-\frac{3}{2}\xi) \exp((\delta - 1)T)$$

So for T big enough, then for all s:

$$f^i_{s,\boldsymbol{v}_1(t_1)}(\boldsymbol{x}_{1,s}(t_1)) - v_{1,s}(t_1) \geq -\xi/4$$

Iteratively, we get:

$$f^i_{s,\boldsymbol{v}_n(t_n)}(x_{n,s}(t_n)) - v_{n,s}(t_n) \geq -\xi/4$$

which is contradictory to the fact that $\min_{s\in S} f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s) - v_s = -\xi$.

For SyncBRD, we have the exact same proof.

So $L \subseteq B$.

We can now use a more classic argument to show that $L \subseteq A$ with a Lyapunov function now that the ambient space can be restricted to B. Let us define $V(\boldsymbol{x}, \boldsymbol{v}) := \sum_{s\in S} f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s)$. Then, V is a Lyapunov function for set A with ambient space B. Indeed, on B, $\frac{dv_s}{dt} \geq 0$, so $\frac{df^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s)}{dt} \geq 0$. Therefore, $\frac{df^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s)}{dt} = 0$ for every s if and only if $(\boldsymbol{x}, \boldsymbol{v}) \in A$. Moreover, $V(A)$ has empty interior thanks to Sard's Theorem.

So we can use Proposition 3.27 of Benaim et al., 2005: it applies in case the Lyapunov function is defined on invariant set. So L is contained in A.

□

Lemma 7.15 (Internally Chain Transitive Sets for MFBRD)**.** *If for all t, $\alpha(t) = 1$ and if L is internally chain transitive for MFBRD then*

$$L \subseteq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \,\middle|\, \forall s \in S, \forall i \in I,\ f^i_{s,\boldsymbol{v}}(\boldsymbol{x}_s) + \eta h(\boldsymbol{x}_s) = v_s \wedge x^i_s \in \arg\max_{y^i \in A^i} f^i_{s,\boldsymbol{v}}\left(y^i, \boldsymbol{x}^{-i}_s\right) + \eta h(\boldsymbol{x}_s) \right\}.$$

Proof. The proof is the same as the proof of Lemma 7.14 but an extra step, the convergence of the transition and payoff estimate is done beforehand. □

A second proof of Theorem 6.1 using continuous time Below, we prove a second time Theorem 6.1 in the autonomous case (that is, $\alpha_n = 1$) to show that our extension to the stochastic approximation framework is directly useful.

Proof of Theorem 6.1. For systems (AFP) we now need to apply Theorem 7.4. Variable Y_n is 0 in our case because there is no noise. S_{n+1} is the next state variable, and it has distribution $P_{S_n}(a_n)$. We check the assumptions:

- (i) is guaranteed because every variable of the system is bounded.
- (ii) is guaranteed because the best-response map is Marchaud and the derivative of u is continuous.
- for (iii) and (vi) we use $\alpha_n = 1/n$, so every assumption is trivial to verify.
- (iv) and (v) comes from the definition of a play and the ergodicity hypothesis on the game

Therefore the affine interpolation of a sequence of fictitious play for stochastic games under our assumption is an asymptotic pseudo-trajectory, which implies that its limit set is internally chain transitive by Theorem 4.3 of Benaim et al., 2005.

For system SyncFP Lemma 7.14 states that the limit set is internally chain transitive.

Then Lemma 7.14 concludes the proof: the limit set is internally chain transitive and consequently included in the set of equilibria.

□

Proof of Theorem 7.1 The proof of Theorem 7.1 is similar to that of Theorem 6.1 but with the internally chain transitive set of MFBRD.

7.4.3 Convergence of FP and SFP in zero-sum stochastic games

We consider SBRD and ABRD in zero-sum stochastic games in the autonomous case, that is for $\alpha(t) = \alpha^\star$.

Proof of Theorem 6.2 and Theorem 7.2. The previous proof checks all the hypothesis necessary to apply the stochastic approximation framework we presented in this section. Therefore, a characterization of internally chain transitive sets will be sufficient to conclude.

Lemma 7.10 shows that function $\max\left\{w_s(t) - 2D\alpha^\star\beta_-^{-1}, 0\right\}$ is a Lyapunov function. Therefore, the set where the duality gap is lower or equal than $2D\alpha^\star\beta_-^{-1}$ is an internally chain transitive set. Furthermore, in the proof of Lemma 7.12, we defined a function g which is a Lyapunov function relative to the previous internally chain transitive set.

□

7.4.4 Different Priors and Team Games for asynchronous FP

In this subsection, we suppose that every player has its own v_s^i estimates. We are going to show that, in the asynchronous best-response dynamics, internally chain transitive sets are included into the set where the estimates v_s^i are equal (up to a constant) for every i.

Indeed, suppose that our stochastic game is now a team game. Then every utility function u_s^i can be written $u_s^i = u_s + M_i$ with the convention that $M_1 = 0$ and $u_s^i = u_s^1$. We are going to show that any internally chain transitive set L is included in $\{(\boldsymbol{x}, \boldsymbol{v}) \mid v_s^i = v_s^1 + M_i \; \forall i, s\}$.

We define the following function, which will be shown to converge to 0,

$$V^i(\boldsymbol{x}, \boldsymbol{v}) = \arg\max_{s \in S} |v_s^i - v_s^1 - M_i|. \tag{7.29}$$

Let s that maximizes $|v_s^i - v_s^1 - M_i|$ so that $V^i(\boldsymbol{x}, \boldsymbol{v}) = |v_s^i - v_s^1 - M_i|$. Then if $v_s^i > v_s^1 - M_i$, $V^i(\boldsymbol{x}(t), \boldsymbol{v}(t))$ can be differentiated for almost every t (using the same techniques as in the rest of this section):

$$\begin{aligned}\frac{dV^i}{dt} &= \alpha(t)(f^i_{s,\boldsymbol{v}^i}(\boldsymbol{x}_s(t)) - f^1_{s,\boldsymbol{v}^1}(\boldsymbol{x}_s(t)) - v_s^i(t) + v_s^1(t) \\ &\leq \alpha(t)((1-\delta)M_i + \delta V^i(\boldsymbol{x}, \boldsymbol{v}) + \delta M_i - v_s^i(t) + v_s^1(t) \quad \leq \alpha(t)(\delta - 1)V^i(\boldsymbol{x}, \boldsymbol{v})\end{aligned}$$

And similar calculations for the case $v_s^i \leq v_s^1 - M_i$ give the same results.

Therefore, V^i is a Lyapunov function and $L \subseteq {V^i}^{-1}(\{0\})$, hence the result.

7.5 Simulation: $2 \times 2 \times 2$ games

For illustration purposes, we demonstrate an implementation of our proposed algorithm on a stochastic game with two states, two actions and two players.

There are two states:

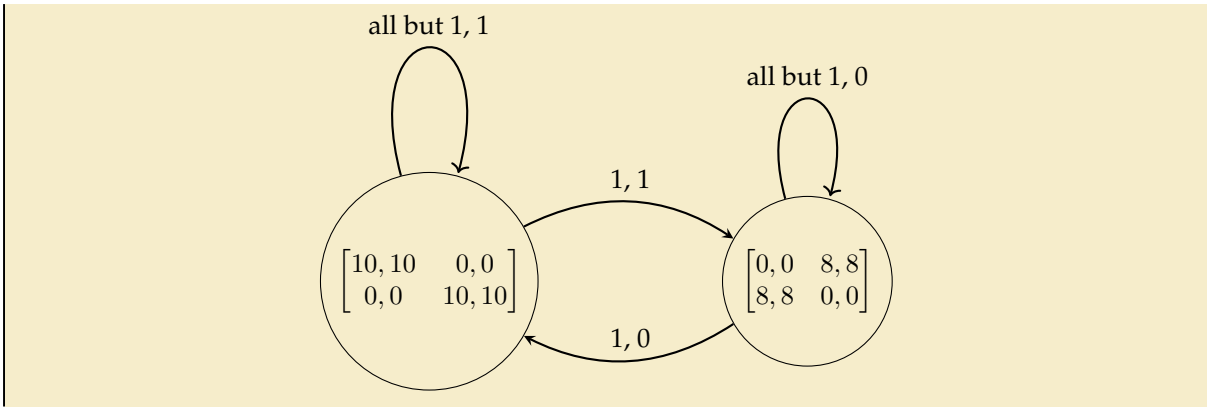
- in the first one, players play a coordination game, with overall higher payoffs
- in the second one, they play an anti coordination game with lesser payoffs.

Transitions are as follows:

- in the first state, the system goes to the first state (with probability $1 - \epsilon$) except if both players play 1
- in the second state, transitions goes to the second state (with probability $1 - \epsilon$) except if players play $(1, 0)$

In this game, players have identical interest, and players prefer the first state to the second one (in the sense that they can get higher payoffs there), so they should play the $(0, 0)$ in the first state and (1, 0) in the second one in order to maximize their overall payoff.

Game 6 (Coordination and Anti-Coordination)**.**



After 10000 iterations of Asynchronous Smooth Fictitious Play with discount factor 0.5, players on average play $(0, 0)$ in the first state and $(1, 0)$ in the second state: they learn that in the first state it is best to select $(0, 0)$ which at the same time yields the higher payoff and which also keep players in the first state, and in the second state it is better to play $(1, 0)$ in order to join the first state.

Simulation notebooks are in the supplementary material: `https://www.lamsade.dauphine.fr/~lbaudin/manuscript/`.

7.6 Conclusion

We defined a number of decentralized, continuous and discrete-time systems with an exploration mechanism so as a player can learn its own payoff function and the transition probability function. When all players use the same procedure, then the system converges to regularized (and so ϵ-approximate) stationary Nash equilibria in discounted stochastic games.

Our procedure has some common characteristics with decentralized Q-learning of Sayin et al. (2021): both do not need the transition probability map or utility functions of the underlying games. However, while decentralized Q-learning estimates also a Q-function (and as such, does not need to observe other player actions), we chose to use the empirical average action of other players. Therefore, our procedure suppose that other players' actions are observable, and it builds a model of other players' behavior.

Completeness of the Convergence Theorems In discrete time, Theorems 7.1 and 7.2 comprise convergence results for identical-interest and zero-sum stochastic games. While in the case of identical-interest games, the convergence is shown to the set of regularized Nash equilibria, it is only shown to an approximate set of these equilibria for zero-sum games. Although a doubling-trick algorithm can progress towards the exact set of regularized equilibria (via an update of the $\alpha^\star$ value), this algorithm is not presented here and could be described in future work.

Moreover, compared to the previous chapter, there is no attempt to extend discrete-time results to other update rates, in particular for continuation payoffs. Similarly, different priors on continuation payoffs and team games are addressed in the asynchronous best-response dynamics, not in the smooth one. We do not expect that overcoming these two differences would be theoretically demanding–even if they would certainly add a layer of complexity.

Perspectives This is an area that has not been widely studied and as such, a number of questions remain to be answered–either regarding our systems or regarding other recent decentralized learning algorithms in stochastic games.

First, the theory of stochastic approximations gives no clue about the speed of convergence of the corresponding discrete-time algorithms. Moreover, even in continuous time, this question remains open even for potential games (Harris, 1998).Therefore, an interesting question in both the stochastic and the repeated game setting is the rate of convergence of both FP and SFP.

Second, and inspired by the convergence of vanishing fictitious play to equilibria in classical repeated games (see (Benaïm and Faure, 2013; Hadikhanloo et al., 2021), it would be of interest to study a vanishing version of our algorithm with a parameter $\epsilon(t)$ that goes to zero in a way that guarantees the convergence to (exact) stationary Nash equilibria of the discounted stochastic game.

Third, the convergence of FP or SFP in repeated games extends to infinite action games and to non-atomic games. It would be of interest to relax our finiteness assumption, at least for the action sets.

A challenging open problem is the design of independent learning algorithms that converge to stationary equilibria in ergodic zero-sum and identical-interest discounted stochastic games without the knowledge of the other player's past actions. The Projected Gradient Methods of Daskalakis and Panageas (2020) and Leonardos et al. (2021) have this minimal information property, but their results are proved only in episodic stochastic games.

Last but not least, obtaining a last-iterate convergence instead of a time average convergence (as in our paper) or best iterate convergence (as in (Daskalakis and Panageas, 2020; Leonardos et al., 2021)) is another interesting direction. However, it is known that there is no necessary last-iterate convergence with smooth fictitious play in the non-stochastic case (Giannou et al., 2021). Therefore, we can imagine that this would require a different algorithm but we emphasize that algorithms without last-iterate convergence such as ours are still of interest because of their behavioral, micro-economics and experimental foundations (Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002).

Chapter 8

Strategic Behavior and No-Regret Learning in Queueing Systems

This chapter is derived from an article not published yet, written with Marco Scarsini and Xavier Venel:

Lucas Baudin et al. *Strategic Behavior and No-Regret Learning in Queueing Systems*. Preprint, 2023. arXiv: 2302.03614 [cs.GT]

This chapter examines a dynamic discrete-time queuing model where at every period players get a new job and must send all their jobs to a queue that has a limited capacity. Players have an incentive to send their jobs as late as possible; however if a job does not exit the queue by a fixed deadline, then the owner of the job incurs a penalty and this job is sent back to the player and joins the queue at the next period. Therefore, stability, i.e., the boundedness of the number of jobs in the system, is not guaranteed. We show that if players are myopically strategic, then the system is stable when the penalty is high enough. Moreover, if players use a learning algorithm derived from a typical no-regret algorithm (exponential weight), then the system is stable when penalties are greater than a bound that depends on the total number of jobs in the system.

8.1 Introduction

In the classical treatment of queues, agents arrive at random times, wait according to a specified regime, and then get served; the service time is also random. In these models there is no room for any strategic behavior of the agents. Even when agents balk or renege, this is modeled as a random event, not as a strategic choice of the agents. Starting with the seminal paper by Naor (1969), strategic elements have been included in queueing models. For instance, in Naor's model, when agents arrive, they rationally choose whether to join the queue or to balk.

The suitable tools to analyze strategic queueing models come from game theory. The literature has considered several strategic models of queueing systems under different stochastic assumptions, different service regimes, and different strategy sets for the players. Various goals have been considered, such as computing Nash equilibria of the games, studying their efficiency, and examining the system stability under various equilibria.

One interesting class of problems, first examined by Glazer and Hassin (1983), deals with situations where players need to be serviced before a fixed deadline, and otherwise pay a steep penalty. In several variations of this model, players play the game repeatedly, for instance, they commute daily to the office and need to get there on time.

In a recent development, Gaitonde and Tardos (2020a) and Gaitonde and Tardos (2021) considered a discrete-time queueing model where agents use learning algorithms to make the decision of which server to choose. The novelty of their analysis was the consideration of spillovers from one period to the other. One of their goals is to establish conditions for the system to be stable.

8.1.1 Our contribution

This chapter draws both on the literature on queues with a fixed deadline and on the theory of learning in games and studies a discrete-time model where agents at every period receive a new job that requires service from a single server and need to decide when their jobs join a queue, taking into account a trade-off between waiting costs and a stiff penalty for being late. The model includes spillovers, since the jobs that cannot be served by the deadline go back to their owners and are to be sent to the server during the following period; these late jobs are then added to the incoming daily new job. From a queuing point of view, the model is deterministic: each player gets exactly one new job at every period. Randomness is due to the actions of the players, which can be mixed, and to the regime of the queue: if several jobs join the queue at the same time, the order in which they get served is uniformly random.

We consider several aspects of the model under the assumption that agents are strategic, but myopic, i.e., at every period they play an equilibrium, but do not take into account the future effect of their actions. This leads us to examine the equilibria of the single-period game. In this framework we show that the structure of equilibria depends on whether the number of jobs in the system is or is not larger than the number of times in each period; moreover it depends on whether the penalty cost for being late is or is not large enough.

When the number of jobs in the system exceeds the number of times in each period, and the penalty cost is large enough, we show that the stage game has a single coarse correlated equilibrium (hence, a single Nash equilibrium), where all players send all their jobs to the queue as early as possible. When the number of jobs in the system does not exceed the number of times in each period, then the stage game has multiple equilibria, whose structure we study. Surprisingly, even if each player has a minmax strategy that guarantees that this player's jobs will meet the deadline, nevertheless, in equilibrium some jobs will be late with positive probability. This implies that the number of jobs in the system in the following period will be larger than in the current period.

In the second part of this chapter we study a model where players use a no-regret learning algorithm to make their choices. As we mentioned before, the number of jobs that each player has may vary from one period to another. To face this, we will adopt a variation of the exponential weight algorithm that takes into account the changing environment.

We describe the model using the language of game theory, but other motivations are possible. For instance, we could consider a revenue-management interpretation where at each period agents buy a priority for their jobs. There are different priority levels and their price is monotone with the priority. Jobs with high priority are served before jobs with lower priority, up to the fixed capacity for each period. When there is a priority conflict, it is resolved at random. The constraint is that a maximum of one agent with the lowest priority is served, a maximum of two agents with the lowest two priorities are served, etc.

8.1.2 Related literature

The analysis of strategic behavior in queueing systems goes back to the seminal paper of Naor, 1969, who studied an M/M/1[1] queue with a first-in first-out (FIFO) policy where the agents' payoff consists of the reward that they get when they get served, minus a waiting cost that is proportional to the time they spend in the queue. Once they arrive, they can decide whether to join the queue or to balk. If they play a Nash equilibrium, their behavior is socially inefficient, in the sense that it does not maximize the social welfare (the sum of all players' payoffs). This is due to the fact that one agent's selfish behavior does not take into account the externalities it creates on other agents. Hassin (1985) showed that optimality can be achieved by a last-in first-out (LIFO) policy. The literature on strategic queueing systems has then exploded. The reader is referred to Hassin and Haviv (2003) and Hassin (2016) for a general treatment of the topic.

Some models have considered strategic agents who can decide when to join a queue. The seminal paper by Glazer and Hassin (1983) considered a model, called ?/M/1, where agents arrive at a facility that every day starts service at time 0, and serves all customers that arrive by some time T according to a FIFO policy. Each day, agents decide whether to visit the facility or not. If they do, they pick their

[1]The two first Ms mean that arrivals and departures from the queue are Markovian, 1 is the number of service nodes (i.e., the number of exits of the queue).

arrival time with the goal of minimizing their expected time in the queue. This queueing literature with a strategic choice of the arrival time has been recently surveyed by Haviv and Ravner (2021).

In the framework of transportation theory, Vickrey (1969) studied a bottleneck model where agents choose the time they leave home to reach the office and used some fluid approximation. Rivera et al. (2018a) studied a discrete version of the Vickrey model and examined Nash and correlated equilibria and their efficiency. Our model is (a variation of) a repeated version of their model with spillover. Kawasaki et al. (2023) extended the study of this bottleneck model to more general preferences and provided conditions for the existence of a pure Nash equilibrium; moreover they examined the link with strong equilibria.

Recently, Gaitonde and Tardos (2020a) proposed a model where several queues receive packets with a fixed, time-independent (but queue dependent) probability and must send them to a number of servers. Each server may process a packet it received with a fixed, time-independent probability. The paper studied the behavior of the system when players use no-regret learning procedures and in particular determined the conditions on the model parameters for the system to be stable. In a subsequent paper, Gaitonde and Tardos (2021) compared the behavior of no-regret, short-term, learners with players who adopt a long-run optimizing behavior. Sentenac et al. (2021) used a similar model and introduced a new cooperative and decentralized algorithm that players can use. Compared to this work, this chapter does assume that players intend to cooperate, in the sense that they use a standard algorithm to maximize their own payoff.

No-regret learning procedures are a family of online reinforcement learning algorithms such that they are at least as good as any constant action selected in hindsight (i.e., given that other players or the environment actions are fixed). The notion, also called Hannan or universal consistency (Fudenberg and Levine, 1998), was originally introduced by Hannan (1957) and is satisfied by multiple algorithms (see, for a recent review, Perchet, 2014). In particular, the exponential weight algorithm (EWA) (Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006) is a simple but efficient reinforcement learning procedure which adjusts weights of actions based on past experiences. However, there is no widely accepted extension of the concept of no-regret learning to systems with (endogenously changing) state variables (for instance Markov Decision Processes) such as ours (the state of the queue), so we propose a simple, multi-level, extension to EWA. Besbes et al. (2015b) and Besbes et al. (2019a) dealt with stochastic optimization and single-agent, multi-armed bandit problems with temporal uncertainty in the rewards. In a recent interesting paper, Crippa et al. (2022) studied—in the context of (non-stochastic) repeated games—strategies that have no-regret compared to dynamic sequences of actions whose number of changes scales sublinearly in time. There is also a large literature on time-varying games, which were studied by Cardoso et al. (2019), Mertikopoulos and Staudigl (2021), Duvocelle et al. (2022b), Zhang et al. (2022), and Anagnostides et al. (2023b), among others. In a game theoretic framework, the reward functions of a player can change across time for two reasons: a change in the game and a change in the strategy of the other players. In these articles, the authors investigate different solution concepts to take this distinction into account. A key difference with our model is that the change of the game is exogenous whereas in our model it is endogenous. We restrict ourselves to the more classical notion of regret.

8.1.3 Outline

Section 8.2 introduces the model and the key concept that will be considered. Section 8.3 analyzes the behavior of myopic strategic players. The key point in this section is the analysis of the one-shot game for any job allocation to the players. Section 8.4 analyzes the repeated game with learning players.

8.2 The Model

We consider a discrete-time game where I is the set of players and time is split into periods of equal length $L \geq 2$. A single, generic period is denoted by t. Throughout this chapter we assume that there are at least $|I| \geq 3$ players.

At the beginning of every period $t \geq 0$ each player i receives one job and independently chooses an action $a_t^i \in \{0, \ldots, L-1\}$, which represents the time during period t at which i's jobs join the queue. The symbol $\boldsymbol{a}_t \coloneqq (a_t^i)_{i \in I}$ denotes the action profile at period t. Since all of player i's jobs join the queue

at the same time, the action set $A \coloneqq \{0, \dots, L-1\}$ does not depend on the number of jobs held by each player.

The choice of a_t^i may depend on the past history of the game. Depending on the actions taken by all players, some jobs are late, i.e., they cannot exit the queue before the end of the period. Late jobs are returned to their respective owners and will join the queue once more during the following period. This produces a *spillover* effect: at every period, the number of jobs that player i handles is the number of i's late jobs in the previous period plus one.

The state of the system at period t is the vector $\boldsymbol{k}_t = (k_t^i)_{i \in I}$, where k_t^i is the number of player i's jobs at the beginning of period t. The state space is $S \coloneqq \mathbb{N}^I$. The total number of jobs at time t is

$$k_t \coloneqq \sum_{i \in I} k_t^i. \tag{8.1}$$

Late jobs at period t incur a penalty cost C_{k_t} that depends on the total number k_t of jobs in the system at period t. At the end of period t, each player i pays a cost $c^i_{\boldsymbol{k}_t}(\boldsymbol{y}_t)$ that is the sum of two components: the waiting cost, i.e., the time spent in the queue by each of i's jobs, and the penalty cost:

$$c^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t) = k_t^i(L - a_t^i) + C_{k_t}\mathbb{E}\left[\text{number of } i\text{'s jobs that are late at } t\right]. \tag{8.2}$$

Given an action profile $\boldsymbol{a}_t$, all of player i's jobs are assumed to have the same probability $p^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t)$ of being late. As a consequence

$$\mathbb{E}\left[\text{number of } i\text{'s jobs that are late at } t\right] = k_t^i p^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t), \tag{8.3}$$

which implies

$$c^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t) = k_t^i(L - a_t^i) + C_{k_t} k_t^i p^i_{\boldsymbol{k}_t}(\boldsymbol{a}_t). \tag{8.4}$$

A dynamic queueing model (DQM) is specified by a period length L, a number of players I (identified with the set of players) and a sequence of penalties $(C_{k_t})_{t \in \mathbb{N}}$ that depend on the total number k_t of jobs in the system.

Given the spillover effect, it is natural to investigate the asymptotic behavior of the number of jobs k_t. In particular, if $k_t > L$ at period t, then there are more jobs than time units; therefore, at least one job is necessarily late. Some jobs can be late even if $k_t \leq L$. In the rest of this chapter, we will be interested in bounding the number of jobs in the system, under various players' behaviors. More formally, we let $\mathcal{H}_t \coloneqq (S \times A)^{t-1} \times S$ denote the set of histories of length t and $\mathcal{H} \coloneqq \cup_{t \geq 1} \mathcal{H}_t$ the set of finite histories.

Definition 8.1 (Strategy)**.** A *strategy* is a function $x^i \colon \mathcal{H} \to \triangle(A)$, where $\triangle(A)$ is the set of distributions over A. When there is no risk of ambiguity, we will let $x_t^i \coloneqq x^i(h_t)$ denote player i's mixed action at period t.

Definition 8.2 (Stability of a strategy profile)**.** A strategy profile $\boldsymbol{x}$ in a dynamic queueing model (DQM) is said to be *stable* if the number of jobs k_t is almost surely bounded, i.e.,

$$\mathbb{P}_{\boldsymbol{x}}\left[\exists M, \forall t \in \mathbb{N}, k_t \leq M\right] = 1.$$

The following assumption will be in place throughout this chapter.

Assumption 8.1. *The following inequality holds: $L \geq I$, i.e., the number of slots in the period is larger than the number of players.*

Without Assumption 8.1, no strategy profile in a DQM could be stable, because at every period at most L jobs leave the system, and I new jobs arrive.

We study the stability of two different families of strategy profiles. In the first case, agents are myopically strategic, i.e., they focus on the situation at the current period and play in a strategic way for that period, i.e., they play at every period a correlated equilibrium. We prove that the DQM is stable under this strategy profile, provided the penalty costs C_{k_t} are large enough (possibly constant in k_t). In the second case, agents use no-regret strategies that depend on their number of jobs. We provide sufficient conditions for the stability of these strategies in the DQM.

8.3 Myopic Strategic Players

In this section, players are assumed to be strategic and myopic. The myopic assumption implies that at period t players only consider the costs that they may pay at the current period without taking into consideration the effect that their actions have on the future of the game. In other words, players repeatedly play one-shot games with payoff functions $(c^i_{\boldsymbol{k}_t})_{i\in I}$.

We now introduce the definitions of myopic Nash equilibrium (NE) and myopic coarse correlated equilibrium (CCE).

Definition 8.3 (Myopic Nash equilibrium (NE))**.** A strategy profile $\boldsymbol{x}\colon \mathcal{H} \to \triangle(A^I)$ is called a *myopic NE* if, for any history h_t with current state $\boldsymbol{k}_t$, the mixed action profile $\boldsymbol{x}(h_t)$ is a Nash equilibrium of the one-shot game with jobs $\boldsymbol{k}_t = (k^i_t)_{i\in I}$.

Definition 8.4 (Myopic coarse correlated equilibrium (CCE))**.** A strategy distribution $\boldsymbol{x}\colon \mathcal{H} \to \triangle(A^I)$ is called a *myopic CCE* if, for any history h_t with current state $\boldsymbol{k}_t$, the distribution $\boldsymbol{x}(h_t)$ is a coarse correlated equilibrium of the one-shot static game with jobs $\boldsymbol{k}_t = (k^i_t)_{i\in I}$.

When the number of jobs is lower than L, any player who plays 0 is sure not to pay the penalty cost. Even in this case, there exist equilibria where some jobs are late with positive probability. Therefore, the number of jobs may not be constant throughout the play, as detailed in Section 8.3.1.

We now analyze the family of one-shot games parameterized by the number of jobs owned by each player.

The rest of the section is organized as follows. We start by focusing on the one-shot game, first when $L \geq k_t$, and then when $L < k_t$. We conclude by analyzing the global queue.

8.3.1 Equilibria in the one-shot game when $L \geq k_t$

We will show that, even when $L \geq k_t$, in equilibrium some jobs are late with positive probability.

We first prove that only the last k_t actions are used in an equilibrium; hence, it is enough to consider the case $L = k_t$.

Lemma 8.1 (Dominated actions for $L > k_t$)**.** *If $L > k_t$, then for each player i, any action $a^i_t \in \{0, \dots, L-k_t-1\}$ is strictly dominated by action $L - k_t$. Furthermore, for any (mixed) action profile $\boldsymbol{y}_t$,*

$$c^i_{\boldsymbol{k}_t}\big(L - k_t, \boldsymbol{y}_t^{-i}\big) = k^i_t k_t. \tag{8.5}$$

Eq. (8.5) shows that the cost $k^i_t k_t$ that player i pays when playing action $L - k_t$ is independent of the other players' actions.

Proof of Lemma 8.1. If player i's jobs join the queue at time a^i_t, then they leave the queue no later than time $a^i_t + l$, where l is the number of jobs that the queue at time a^i_t or before. Since $l \leq k_t$, player i's jobs leave the queue no later than time $a^i_t + k_t$, which is smaller or equal than L if a^i_t is smaller or equal than $L - k_t$. So, the probability $p^i_{\boldsymbol{k}_t}(a^i_t, \boldsymbol{y}_t^{-i})$ that a specific job of player i is late is 0 and player i's cost is $k^i_t(L - a^i_t)$. In particular, the cost of choosing action $L - k_t$ is $k^i_t k_t$, which is strictly smaller than the cost of any $a^i_t \in \{0, \dots, L - k_t - 1\}$. □

Remark. *Lemma 8.1 implies that, in proofs, it is enough to consider the case $L = k_t$: when $L > k_t$, the first actions are never employed by rational players since they are strictly dominated. Therefore, in the rest of this section, all the results will be proved under the hypothesis $L = k_t$ (but still stated for $L > k_t$) and the general case $L \geq k_t$ can be obtained by renaming the actions.*

The next theorem shows that, in any equilibrium, players can be divided into two groups: players in the first group choose the same mixed action and mix only on the two first non-dominated actions $L - k_t$ and $L - k_t + 1$; players in the second group do not mix on $L - k_t$ and put strictly positive weight on $L - k_t + 1$. The way they mix on the remaining non-dominated actions is not specified.

The symbol $\mathsf{SC}(\boldsymbol{y}_t)$ denotes the social cost of a profile $\boldsymbol{y}_t$, that is,

$$\mathsf{SC}(\boldsymbol{y}_t) \coloneqq \sum_{i\in I} c^i_{\boldsymbol{k}_t}(\boldsymbol{y}_t). \tag{8.6}$$

Theorem 8.1 (Structure of Nash equilibria). *If $L \geq k_t$, $C_{k_t} > k_t^2$, and $\boldsymbol{y}_t$ is a Nash equilibrium, then:*

(i) *for each player j and action $a_t^j < L - k_t$, we have $y_t^j(a_t^j) = 0$;*

(ii) *there exists a player i such that*

(a) $y_t^i(L - k_t) > 0$,

(b) *for all $a_t^i > L - k_t + 1$, we have $y_t^i(a_t^i) = 0$;*

(iii) *for every player $j \neq i$, $y_t^j(L - k_t + 1) > 0$;*

(iv) *for every player j, if $y_t^j(L - k_t) > 0$, then $y_t^j = y_t^i$;*

(v) $k_t^2 - k_t + 1 \leq \mathsf{SC}(\boldsymbol{y}_t) \leq k_t^2$.

Proof. (i) *The actions before $L - k_t$ are not chosen.* This is the remark written above. By Lemma 8.1, we know that all actions smaller than $L - k_t$ are strictly dominated. The result follows from the fact that Nash equilibria do not mix on dominated actions. In the rest of the proof, we will assume that $L = k_t$. As mentioned before, if $L < k_t$, then the proof can be easily adapted by translation: action 0 becomes $L - k_t$, 1 becomes $L - k_t + 1$, etc.

(ii)(a) *There exists a player i such that $y_t^i(0) > 0$.* Assume *ad absurdum* that $y_t^j(0) = 0$ for each player j. Then at least one job is late. This implies that

$$\mathbb{E}\left[\text{number of late jobs}\right] = \sum_{j \in I} k_t^j p_{\boldsymbol{k}_t}^j(\boldsymbol{y}_t) \geq 1 \tag{8.7}$$

and there exists i such that $k_t^i p_{\boldsymbol{k}_t}^i(\boldsymbol{y}_t) \geq k_t^i / k_t$. Then,

$$c_{\boldsymbol{k}_t}^i(\boldsymbol{y}_t) \geq k_t^i C_{k_t} p_{\boldsymbol{k}_t}^i(\boldsymbol{y}_t) \geq C_{k_t} k_t^i / k_t.$$

So if $C_{k_t} > k_t^2$, player i's cost is strictly greater than $k_t^i k_t$, which is a contradiction. Therefore, one player i mixes on 0 and $c_{\boldsymbol{k}_t}^i(\boldsymbol{y}_t) = k_t^i k_t$.

(iii) *Every other player mixes on 1.* If this is not the case, then $y_t^j(1) = 0$ for some $j \neq i$. We now prove that i can profitably deviate by playing the pure action 1. Assume that, indeed, i chooses action 1. If j plays the pure action 0, then player i's jobs joining the queue at 1 are not late because the number of jobs joining the queue at 1 is smaller than $k_t = L$, so they are processed before the end of the period. Else, if j does not play 0, then j's action is not smaller than 2. Then the number of jobs that join the queue at 0 or 1 is at most $L - 1$ and these jobs are processed before the end of the period.

So, by deviating to 1, i's jobs are not late and this leads to a cost $c_{\boldsymbol{k}_t}^i\left(1, \boldsymbol{y}_t^{-i}\right)) = k_t^i(k_t - 1)$. Hence, the deviation is profitable and $\boldsymbol{y}_t$ cannot be an equilibrium.

(iv) *(First step) All players who mix on 0 mix on 1, and put the same weight on action 1.* Suppose that at least two players mix on 0. We first prove that they mix on 1 too. Call these two players i and j. Player i satisfies (ii)(a); therefore, by (iii), player j mixes on 1. Symmetrically, player j satisfies (ii)(a); therefore, by (iii), player i mixes on 1. This shows that both players mix on 1.

We now prove that they put the same weight on 1. A job sent by any of these players can be late only if all jobs join the queue at time 1. Indeed, if at least one job joins the queue at 0, then none of the jobs that join the queue at 1 can be late because there are $L - 1$ units of time to process them.

Therefore, a job departing at 1 is late only if all other jobs also depart at 1, i.e., all players play action 1. If this happens, then the probability that a job is late is $1/k_t$. Therefore, the probability that a job owned by player i is late, conditionally on the fact that i plays 1, is equal to

$$p_{\boldsymbol{k}_t}^i(1, \boldsymbol{y}_t^{-i}) = \frac{\mathbb{P}\left[a_j^i = 1 \; \forall j \neq i\right]}{k_t} = \frac{\prod_{j \neq i} y_t^j(1)}{k_t}.$$

Furthermore, since in a Nash equilibrium player i is indifferent between 0 and 1, we have

$$c_{\boldsymbol{k}_t}^i\left(0, \boldsymbol{y}_t^{-i}\right) = k_t^i k_t = c_{\boldsymbol{k}_t}^i\left(1, \boldsymbol{y}_t^{-i}\right) = k_t^i \left(k_t - 1 + \frac{\prod_{j \neq i} y_t^j(1)}{k_t} C_{k_t} \right). \tag{8.8}$$

Thus, we have

$$\prod_{j\neq i} y_t^j(1)C_{k_t} = k_t. \tag{8.9}$$

This applies to any j such that $y_t^j(0) > 0$. Hence $y_t^i(1) = y_t^j(1)$.

(ii) (a) and (iv) *(second step) Players who mix on* 0 *do not mix on any action strictly larger than* 1. If player i mixes on an action $a_t^i > 1$, then one of i's jobs is late whenever all other players play 1. This implies that the probability to be late playing action $a_t^i > 1$ is greater than the probability that all other players play 1, so using (8.9) we get

$$c^i_{\boldsymbol{k}_t}\big(a_t^i, \boldsymbol{y}_t^{-i}\big) \geq k_t^i \left\{ k_t - a^i + \prod_{j\neq i} y_t^j(1)C_{k_t} \right\} = k_t^i k_t + k_t^i(k_t - a_t^i) > k_t^i k_t.$$

Since all players who mix on 0 only mix between 0 and 1 with the same weight on 1, they actually play the same mixed action, which proves (iv).

(v) *Social cost.* First we prove that the social cost is smaller than k_t^2. If $\boldsymbol{y}_t$ is a Nash equilibrium, then for every player j, we have

$$c^j_{\boldsymbol{k}_t}(\boldsymbol{y}_t) \leq c^j_{\boldsymbol{k}_t}\left(0, \boldsymbol{y}_t^{-j}\right) = k_j^i k_t. \tag{8.10}$$

Then, $\mathsf{SC}(\boldsymbol{y}_t) \leq \sum_{j\in I} k_j^i k_t = k_t^2$. We now prove that the social cost is greater than $k_t^2 - k_t + k_t^i$. One player i mixes on 0 and incurs a cost equal to $k_t^i k_t$, and all other players $j \neq i$ mix on 1, so their cost is at least $k_j^i(k_t - 1)$. Thus,

$$\mathsf{SC}(\boldsymbol{y}_t) \geq (k_t - k_t^i)(k_t - 1) + k_t^i k_t = k_t^2 - k_t + 1. \qquad \square$$

Corollary 8.1. *For any Nash equilibrium* $\boldsymbol{y}$*, there is positive probability that a job is late, that is*

$$\mathbb{E}[\text{number of late jobs}] > 0.$$

Proof. By Theorem 8.1 (ii) (a), we know that there exists a player i who mixes on $L - k_t$ with positive probability. We consider two cases, based on this probability. If i plays with probability strictly less than 1, then by Theorem 8.1 (iii) and Theorem 8.1 (iv) we know that all the players are playing simultaneously in $L - k_t + 1$ with positive probability. Therefore, one of them is late.

Let us assume that i plays with probability 1 the action $L - k_t$. If two players mix on $L - k_t$, we can apply Theorem 8.1 to each of them and therefore by (iii), both put positive weight on $L - k_t + 1$. So, i is the only player playing $L - k_t$ and by Theorem 8.1 (iv) every other player plays $L - k_t + 1$ with positive probability. Let $j \neq i$ and suppose that $y_t^j(L - k_t + 1) = 1$. Then, there exists another j' different from both i and j (because $I \geq 3$). Player i plays $L - k_t$, j plays $L - k_t + 1$, so j' is not late when it plays $L - k_t + 2$. Therefore, j' strictly prefers $L - k_t + 2$ to $L - k_t + 1$, so it cannot put positive weight on $L - k_t + 1$, which is a contradiction. $\square$

8.3.2 Equilibria when $L < k_t$

We now study the game when the period length is smaller than the number of jobs, i.e., $L < k_t$.

Theorem 8.2 (CCEs)**.** *If* $k_t > L$ *and* $C_{k_t} > k_t^2 L$*, then there is a unique coarse correlated equilibrium, which is actually a pure equilibrium where all players play* 0*.*

To show that the support of any CCE is the singleton $\mathbf{0}$, we proceed by contradiction. We assume the existence of a CCE whose support is not $\mathbf{0}$; we sum the cost of unilateral deviations to 0 for each player and show that the sum is negative. As a consequence, there is at least one deviation cost which is negative, which shows that this player has a profitable deviation. The proof is not "constructive" in the sense that the dissatisfied player is not designated outright.

Proof of Theorem 8.2. Let $\boldsymbol{\tau}_t$ be a CCE. We have

$$c^i_{\boldsymbol{k}_t}(\boldsymbol{\tau}_t) = \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) = \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{ L - a^i + p^i_{\boldsymbol{k}_t}(\boldsymbol{a}) C_{k_t} \big\}. \tag{8.11}$$

Since $\boldsymbol{\tau}_t$ is a coarse correlated equilibrium, the cost that i obtains by a unilaterally deviating to 0 is not smaller then the cost that i incurs under $\boldsymbol{\tau}_t$, that is,

$$\sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{ L + p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) C_{k_t} \big\} \geq \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{ L - a^i + p^i_{\boldsymbol{k}_t}(\boldsymbol{a}) C_{k_t} \big\}, \tag{8.12}$$

which leads to

$$\sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \{ p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) C_{k_t} - a^i + p^i_{\boldsymbol{k}_t}(\boldsymbol{a}) C_{k_t} \} \geq 0. \tag{8.13}$$

A player who was originally playing 0 is actually not deviating. Therefore, the term of the sum that corresponds to players who play 0 are equal to zero. Therefore, in (8.13) we can sum over all players who do not play 0 and get

$$\sum_{\boldsymbol{a}\in A^I\setminus\{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a}) \left\{ \sum_{i\in I} k_t^i (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) C_{k_t} + k_t^i a^i \right\} \geq 0. \tag{8.14}$$

A job that joins the queue after time 0, has a higher probability of being late:

$$p^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \geq p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) \tag{8.15}$$

The following claim refines the above statement.

Claim 8.3.1. *Given an action profile $\boldsymbol{a} \neq \mathbf{0}$, if*

$$a^j = \max_{i\in I}(a^i), \tag{8.16}$$

then

$$p^j_{\boldsymbol{k}_t}(\boldsymbol{a}) \geq \frac{1}{N_{a^j} k_t} + p^j_{\boldsymbol{k}_t}(0, a^j), \tag{8.17}$$

where

$$N_{a^j} \coloneqq \sum_{i\in I\,:\; a^i = a^j} k_t^i \tag{8.18}$$

is the number of jobs that join the queue at time a^j.

Proof. First of all notice that in (8.16) we have $a^j > 0$. Call M_{a^j} the number of jobs that join the queue at time a^j and leave the system before the deadline L. Since $L \leq k_t$, some jobs are late; more precisely, $N_{a^j} - M_{a^j}$ jobs that join the queue at a^j are late. Then

$$p^j_{\boldsymbol{k}_t}(\boldsymbol{a}) = \frac{N_{a^j} - M_{a^j}}{N_{a^j}}. \tag{8.19}$$

Moreover, $p^j_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-j}) \leq (k_t - L)/k_t$, so

$$p^j_{\boldsymbol{k}_t}(\boldsymbol{a}) - p^j_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-j}) \geq \frac{N_{a^j} - M_{a^j}}{N_{a^j}} - \frac{k_t - L}{k_t} = \frac{-M_{a^j} k_t + L N_{a^j}}{N_{a^j} k_t}. \tag{8.20}$$

However, the maximum number of jobs that can leave the system without being late is L.

Under the action profile $\boldsymbol{a}$, $k_t - N_{a^j}$ jobs join the queue strictly before time a^j. To finish the proof, we consider several cases related to the value of $k_t - N_{a^j}$:

- If $0 < k_t - N_{a^j} \leq L$, no more than $L - (k_t - N_{a^j})$ jobs joining the queue at time a^j can arrive on time, i.e., $M_{a^j} \leq L - (k_t - N_{a^j})$. Using this inequality in (8.20), we obtain

$$\begin{aligned} p^j_{\boldsymbol{k}_t}(\boldsymbol{a}) - p^j_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-j}) &\geq \frac{-(L - k_t + N_{a^j}) k_t + L N_{a^j}}{N_{a^j} k_t} \\ &= \frac{L}{k_t} - \frac{L - k_t}{N_{a^j}} - 1 \\ &= \{L - k_t\} \left\{ \frac{1}{k_t} - \frac{1}{N_{a^j}} \right\} \\ &= \{k_t - L\} \frac{k_t - N_{a^j}}{k_t N_{a^j}} \\ &\geq \frac{1}{k_t N_{a^j}}, \end{aligned}$$

because in this case $k_t - N_{a^j} \geq 1$.

- If $k_t - N_{a^j} = 0$, then all the jobs join the queue at a time greater than 1; therefore $M_{a^j} \leq L - 1$. Since in this case $k_t = N_{a^j}$, after some simplifications (8.20) becomes

$$p^j_{\boldsymbol{k}_t}(\boldsymbol{a}) - p^j_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-j}) \geq \frac{-M_{a^j} + L}{k_t} \geq \frac{-(L-1) + L}{k_t} = \frac{1}{k_t} \geq \frac{1}{k_t N_{a^j}}. \tag{8.21}$$

- if $k_t - N_{a^j} > L$, then no jobs can avoid being late and $M_{a^j} = 0$. Therefore, (8.20) becomes

$$p^i_{\boldsymbol{k}_t}(\boldsymbol{a}) - p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) \geq \frac{L}{k_t} \geq \frac{1}{k_t N_{a^j}}.$$

This concludes the proof of the claim. □

Claim 8.3.2. *For any $\boldsymbol{a} \neq \mathbf{0}$, we have*

$$-\frac{1}{k_t} \geq \sum_{i \in I} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})). \tag{8.22}$$

Proof. We split the players into two groups: players whose jobs join the queue at time a^j as defined in (8.16) and the remaining ones. We have

$$\begin{aligned} \sum_{i \in I} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) = &\sum_{i \in I \colon a^i \neq a^j} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) \\ &+ \sum_{i \in I \colon a^i = a^j} k^i (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})). \end{aligned} \tag{8.23}$$

The first term is nonpositive because of (8.15) and the second term is nonpositive because of (8.17), leading to

$$\sum_{i \in I} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) \leq - \sum_{i \in I \colon a^i = a^j} k^i_t \frac{1}{N_{a^j} k_t}. \tag{8.24}$$

The inequality in (8.24) yields

$$\sum_{i \in I} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) \leq -\frac{N_{a^j}}{N_{a^j} k_t} = -\frac{1}{k_t},$$

which proves the claim. □

The combination of Claim 8.3.2 and Eq. (8.14) yields

$$\begin{aligned} 0 &\leq \sum_{\boldsymbol{a} \in A^I \setminus \{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a}) \left\{ \sum_{i \in I} k^i_t (p^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - p^i_{\boldsymbol{k}_t}(\boldsymbol{a})) C_{k_t} + k^i_t a^i \right\} \\ 0 &\leq \sum_{\boldsymbol{a} \in A^I \setminus \{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a}) \left\{ -\frac{1}{k} C_{k_t} + \sum_{i \in I} k^i_t a^i \right\} \\ 0 &\leq \sum_{\boldsymbol{a} \in A^I \setminus \{0\}} \boldsymbol{\tau}_t(\boldsymbol{a}) \left\{ -\frac{1}{k_t} C_{k_t} + L k_t \right\}. \end{aligned} \tag{8.25}$$

Under the assumptions of the theorem, $C_{k_t} > k^2 L$, so

$$-\frac{1}{k_t} + L k_t < 0. \tag{8.26}$$

Therefore, (8.25) can hold only if $\boldsymbol{\tau}_t(\boldsymbol{a}) = 0$ for all $\boldsymbol{a} \in A^I \setminus \{\mathbf{0}\}$, which implies that the support of $\boldsymbol{\tau}_t$ is $\mathbf{0}$. □

8.3.3 Global Queue

The following theorem is a consequence of the previous results.

Theorem 8.3 (Stability for Strategic Repetition with Coarse Correlated Equilibria)**.** *Consider a DQM with $I \leq L$. If $\inf_k C_k > (L+I)^2 L$, then any myopic coarse correlated equilibrium is stable. Moreover, $\forall t, k_t \leq L+I$.*

Corollary 8.2 (Stability for Strategic Repetition with Nash Equilibria)**.** *Consider a DQM with $I \leq L$. If $\inf_k C_k > (L+I)^2 L$, then any myopic Nash equilibrium is stable. Moreover, $\forall t, k_t \leq L+I$.*

Notice that in particular if the penalty cost is a constant $C > (L+I)^2 L$, then any myopic CCE is stable. The corollary is an immediate consequence of the theorem since any Nash equilibrium is also a coarse correlated equilibrium.

As will be clear in the proof of Theorem 8.3, the queue alternates between two possible regimes. The first regime corresponds to the case $k_t \leq L$. In this regime, stage equilibria have a non-trivial structure and in equilibrium some jobs may be late. As a consequence, the number k_t of jobs in the system may oscillate over time. When this number overcomes the level L, the system enters the other regime, corresponding to $k_t > L$. In this regime the only stage equilibrium is the pure profile $\mathbf{0}$. Therefore, if $I = L$, the number k_t of jobs in the system stays constant; if $I < L$, the number of jobs decreases until the system goes back to the first regime.

Remark. *If the penalty costs are small, there may exist a myopic NE that is not stable. For example, if for all $k \in \mathbb{N}$, $C_k < 1$, the cost of preempting the other players is too large compared to the potential gain: playing at the last stage of the period is strictly dominating for each player. The unique myopic NE is for every player to wait the last stage of the period and there are $k_t - 1$ late jobs. Hence, k_t is almost-surely unbounded.*

Proof of Theorem 8.3. Assume that $I \leq L$ and for all k, $C_k > (L+I)^2 L$. We show by induction that for every t, $k \leq L + I$.

The results is true at period 1. We now show that it is true at every period. From period t to period $t+1$, we need to consider two cases. If $k_t \leq L$, then $k_{t+1} \leq L + I$ because at the next period I new jobs arrive and at most k_t are late. If $L < k_t \leq L + I$, then

$$C_{k_t} > (L+I)^2 L \geq k_t^2 L. \tag{8.27}$$

By Theorem 8.2 there is a unique stage CCE, where all players play 0. This implies that at least L jobs leave the system. Since $I \leq L$ and there are I new jobs in the next period, we have $k_{t+1} \leq k_t \leq L+I$. □

8.4 Learning Players

In this section we study a model where every player independently uses a no-regret strategy. The no-regret property, originally introduced by Hannan (1957), is a property of multiple algorithms used in online reinforcement learning (see, e.g., Perchet, 2014). It specifies that in hindsight, the actions taken by a player are at least (asymptotically) as good as any constant action. Formally, in the one-player case, given a sequence of cost functions l_t indexed by time t and a sequence of actions a_t^i, player i's regret is defined as

$$R_t^i = \max_{b^i \in A} \sum_{u=1}^{t} l_u(b^i) - l_u(a_u^i). \tag{8.28}$$

A strategy satisfies the *no-regret* property if $R_t^i = o(t)$. Well-known no-regret strategies include regret-matching (Hart and Mas-Colell, 2013), stochastic fictitious play (Fudenberg and Levine, 1998), and the exponential weight algorithm (EWA) (Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006), which we study below.

Exponential Weight Algorithm In the following, we use the exponential weight algorithm (EWA), which is known to have no-regret guarantees when the payoff is bounded. Unfortunately, boundedness is not satisfied here, as the number of jobs in the system could grow to infinity, resulting in an unbounded penalty. For this reason, we need to study the efficiency of the algorithm more closely.

In the context of repeated games, weights are classically defined as follows:

$$w_t^i(b^i) = \exp\left(\sum_{u=1}^{t-1} -\eta c^i(b^i, \boldsymbol{a}_u^{-i})\right), \tag{8.29}$$

where η[2] is a positive constant. Eq. (8.29) can be rewritten in a recursive fashion as

$$w_{t+1}^i(b^i) = w_t^i(b^i) \exp\left(-\eta c^i(b^i, \boldsymbol{a}_t^{-i})\right). \tag{8.30}$$

Then, the EWA strategy specifies that action b^i is chosen at time t with probability

$$x_t^i(b^i) = \frac{w_t^i(b^i)}{\sum_{a^i \in A} w_t^i(a^i)}. \tag{8.31}$$

Multi-level EWA (MLEWA) EWA is not designed to handle a changing environment. Here, the number of jobs held by every player changes with time. Therefore, there we need to specify how such information is used. We design a new protocol called MLEWA where each player uses several copies of EWA. This protocol is indexed by a parameter n, which we call a *level*. When the number of jobs that player i owns reaches a new level for the first time, this player starts a new EWA where the weights are initialized as a function of the past. When the number of jobs of player i equals a level that has already been visited, this player follows the recommendation given by EWA for this level and updates the weights following EWA. Notice that at any given period t different players may use an EWA at different levels.

Let τ_n^i be the first time player i has n jobs, with the convention that $\tau_n^i = +\infty$ if player i never has n jobs. Player i's weights $w_{t,n}^i(b^i)$ are now parameterized by two parameters: the period t and the level n. The algorithm at level n is defined for all $t \geq \tau_n^i$ by induction. We start by describing the induction step, which is given by

$$w_{t+1,n}^i(b^i) = \begin{cases} w_{t,n}^i(b^i) \exp\left(-\eta c_{\boldsymbol{k}_t}^i(b^i, \boldsymbol{a}_t^{-i})\right) & \text{if } k_t^i = n, \\ w_{t,n}^i(b^i) & \text{otherwise.} \end{cases} \tag{8.32}$$

Eq. (8.32) implies that $w_{t,n}^i(b^i)$ is updated if and only if $k_t^i = n$.

We now describe the initialization. At τ_n^i, we start a new EWA protocol and define initial weights

$$w_{\tau_n^i,n}^i := w_{\tau_n^i,k_{\tau_n^i-1}^i}^i,$$

that is, weights of a newly encountered state are defined as equal to the weights of the previously visited level.

Algorithm 1 summarizes all the steps of our procedure. Its variables are not indexed by t as there are a fixed number of variables and the algorithm does not access the whole history. Instead, it computes the new values at each step and updates the corresponding variables.

MLEWA is based on a no-regret algorithm adapted to changing states. Eq. (8.28) does not deal with changing states. This is a limitation well identified by Gaitonde and Tardos, where regret is computed assuming the state path is unchanged by a change of action. The situation becomes much more complicated when states change endogenously.

We can now state our main result about the stability of the system when players learn.

Theorem 8.4 (Stability of joint no-regret strategies in the subcritical case)**.** *If $I < L$ and $C_k > 4kL$ for all k, then strategy profiles where all players use MLEWA are stable.*

Several lemmas are needed to prove Theorem 8.4. First, we show that (in a static context) action 0 strictly dominates any other actions for a player who holds a large enough number of jobs. The implication of this dominance in our dynamic model is that that the weight on action 0 converges towards 1 when enough jobs are held by a player. Finally, we expose some results on reinforced random walks.

[2]The regret of EWA depends on η and is typically of the order $\eta t + \frac{A}{\eta}$ where A is a constant. Therefore, η is typically chosen proportionally to $\sqrt{t}$ with a doubling-trick algorithm.

Algorithm 1 multi-level EWA (MLEWA)

$\forall b^i \in A$, initialize $w_1^i(b^i)$
$\forall i \in I, k^i \leftarrow 1$ ▷ level 1
for each step $t \geq 1$ **do**
  for each player i **do**
    select an action $a_t^i \sim x_{k^i}^i$ ▷ proportional to $w_{k^i}^i$
  end for
  for each player i **do**
    $\forall b^i \in A, w_{k^i}^i(b^i) \leftarrow w_{k^i}^i(b^i) \exp(-\eta c_{\boldsymbol{k}_t}^i(b^i, \boldsymbol{a}_t^{-i}))$ ▷ number of late jobs + 1
    $k^i \leftarrow \tilde{k}^i + 1$
    if $w_{k^i}^i$ is not defined **then**
      $\forall b^i \in A$, initialize $w_{k^i}^i(b^i)$ ▷ level k^i
    end if
  end for
end for

8.4.1 Domination by action 0

The following lemma shows that in the static case action 0 is strictly dominant for a player i who has enough jobs.

Lemma 8.2 (Strict Domination by 0 when $k_t^i > 2L^2$ in the static model)**.** *If $k_t^i > 2L^2$, $C_{k_t} > 4k_tL$, and $\boldsymbol{a}$ is an action profile such that $a^i \neq 0$, then $c_{\boldsymbol{k}_t}^i(0, \boldsymbol{a}^{-i}) < c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) - k_t^i$.*

Proof. Let $\boldsymbol{a}$ be a pure action profile such that $a^i \neq 0$. Call

$$k_t^{-i}(0) = \sum_{j \neq i} k_j^i 1_{a^j = 0} \tag{8.33}$$

the number of jobs that join the queue at 0 excluding player's i jobs.

Then:

$$c_{\boldsymbol{k}_t}^i(0, \boldsymbol{a}^{-i}) = k_t^i \left\{ L + \frac{k_t^i + k_t^{-i}(0) - L}{k_t^i + k_t^{-i}(0)} C_{k_t} \right\} = k_t^i \left\{ L + \left\{ 1 - \frac{L}{k_t^i + k_t^{-i}(0)} \right\} C_{k_t} \right\}.$$

For each job, i incurs the waiting cost L and an additional cost due to the probability of being late. At period 0, $k_t^i + k_t^{-i}(0) > L$ jobs join the queue; therefore some jobs will surely be late. Moreover, the probability that a job does not incur the penalty is equal to the probability that this job joins the queue among the first L jobs, which happens with probability $L/(k_t^i + k_t^{-i}(0))$.

- If $k_t^{-i}(0) \geq L$, then under the profile $\boldsymbol{a}$ the queue is full from stage 0 and the jobs sent by i are all late, i.e., $c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) = k_t^i(L - a^i) + k_t^i C_{k_t}$. Then

  $$c_{\boldsymbol{k}_t}^i(0, \boldsymbol{a}^{-i}) - c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) = -\frac{k_t^i L}{k_t^i + k_t^{-i}(0)} C_{k_t} + k_t^i a^i.$$

  The assumption on C_{k_t} implies that $C_{k_t} > k_t(L+1)/L$, so:

  $$c_{\boldsymbol{k}_t}^i(0, \boldsymbol{a}^{-i}) - c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) \leq -\frac{k_t^i}{k_t^i + k_t^{-i}(0)} k_t(L+1) + k_t^i a^i.$$

  Since, by definition, $k_t \geq k_t^i + k_t^{-i}(0)$, it follows that:

  $$c_{\boldsymbol{k}_t}^i(0, \boldsymbol{a}^{-i}) - c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) \leq -k_t^i(L+1) + k_t^i a^i \leq k_t^i(a^i - L) - k_t^i \leq -k_t^i.$$

- Consider now the case $k_0^{-i} < L$. Player i pays the waiting cost $L - a^i$ for each job; at most $L - a^i$ of player i's jobs can leave the system without being late. Consequently the following bound holds:

  $$c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) \geq k_t^i(L - a^i) + (k_t^i - L + a^i) C_{k_t}.$$

Then

$$\begin{aligned} c^i_{\boldsymbol{k}_t}\big(0,\boldsymbol{a}^{-i}\big) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) &= k^i_t L + k^i_t C_{k_t} - \frac{L k^i_t}{k^i_t + k^{-i}_t(0)} C_{k_t} \\ &\quad - k^i_t L + k^i_t a^i - (k^i_t - L + a^i) C_{k_t}, \\ &\leq -\frac{k^i_t L}{k^i_t + k^{-i}_t(0)} C_{k_t} + k^i_t a^i + L C_{k_t} - a^i C_{k_t}, \\ &= \left(\frac{k^{-i}_t(0) L}{k^i_t + k^{-i}_t(0)} - a^i\right) C_{k_t} + k^i_t a^i. \end{aligned} \tag{8.34}$$

Since $k^i_t > 2L^2$, it follows that $k^i_t + k^{-i}_t(0) \geq 2L^2$, so

$$\frac{k^{-i}_t(0) L}{k^i_t + k^{-i}_t(0)} \leq \frac{L^2}{2L^2} < \frac{1}{2}.$$

Hence,

$$c^i_{\boldsymbol{k}_t}\big(0,\boldsymbol{a}^{-i}\big) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \leq \left(\frac{1}{2} - a^i\right) C_{k_t} + k^i a^i < -\frac{C_{k_t}}{2} + k^i_t L,$$

because $1 \leq a^i \leq L$.

The assumption that $C_{k_t} > 4kL$ implies

$$c^i_{\boldsymbol{k}_t}\big(0,\boldsymbol{a}^{-i}\big) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) < -2k_t L + k^i_t L < -k_t L < -k_t < -k^i_t. \qquad \square$$

8.4.2 Action 0 is increasingly preferred when the number of jobs grows

We now use Lemma 8.2 to show that, when the number of jobs in the system is high enough, for strategies $x^i_{t,n}$ defined as in Eq. (8.31), the weight $w^i_{t,n}(0)$ of action 0 increases faster than the other weights. Since $x^i_{t,n}$ is proportional to this weight, players are more and more prone to play action 0 when the state is visited again.

Call $\lambda_{t,n}$ the number of times that the level of player i is n, up to time $t-1$:

$$\lambda_{t,n} = \#\big\{k^i_u = n \mid u \in \{0,\dots,t-1\}\big\}. \tag{8.35}$$

We have

Lemma 8.3 (Preference for 0)**.** *If $n > 2L^2$ and $C_{k_t} > 4kL$, then for all $t \geq \tau^i_n$,*

$$x^i_{t,n}(0) \geq \frac{x^i_{\tau^i_n,n}(0)}{x^i_{\tau^i_n,n}(0) + \Big(1 - x^i_{\tau^i_n,n}(0)\Big)\exp\big(-\eta n \lambda_{t,n}\big)}.$$

Proof. We first prove by induction on t that for every $b^i \neq 0$, we get:

$$\frac{w^i_{t,n}\big(b^i\big)}{w^i_{\tau^i_n,n}(b^i)} \leq \exp\big(-\eta n \lambda_{t,n}\big)\, \frac{w^i_{t,n}(0)}{w^i_{\tau^i_n,n}(0)}, \tag{8.36}$$

If $t = \tau^i_n$ then by definition player i never had n jobs before and $\lambda_{t,n} = 0$.

It follows that both sides are equal to 1 and the result is true.

We now show that, if the result holds for $t-1$, then it holds for t. There are two cases.

If $k^i_{t-1} \neq n$, then all weights are equal at stage t and $t-1$, i.e., $w^i_{t,n}\big(b^i\big) = w^i_{t-1,n}\big(b^i\big)$ for all action b^i. Moreover $\lambda_{t,n} = \lambda_{t-1,n}$, so the inequality is the same at t and $t-1$ and therefore true. If $k^i_{t-1} = n$, then $\lambda_{t,n} = \lambda_{t-1,n} + 1$. By the recurrence hypothesis, we know that

$$\frac{w^i_{t-1,n}\big(b^i\big)}{w^i_{\tau^i_n,n}(b^i)} \leq \exp\big(-\eta n \lambda_{t-1,n}\big)\, \frac{w^i_{t-1,n}(0)}{w^i_{\tau^i_n,n}(0)}. \tag{8.37}$$

Using Lemma 8.2 and Eq. (8.32), for $b^i \neq 0$, we get

$$\frac{w^i_{t,n}(b^i)}{w^i_{t-1,n}(b^i)} = \exp\left(-\eta c^i_{\boldsymbol{k}}\left(b^i, \boldsymbol{a}^{-i}_{t-1}\right)\right) \tag{8.38}$$

$$\leq \exp\left(-\eta n\right) \exp\left(-\eta c^i_{\boldsymbol{k}}\left(0, \boldsymbol{a}^{-i}_{t-1}\right)\right), \tag{8.39}$$

$$= \exp(-\eta n) \frac{w^i_{t,n}(0)}{w^i_{t-1,n}(0)}, \tag{8.40}$$

Multiplying Eqs. (8.37) and (8.40), we obtain the result for t.

Fix now t. Eq. (8.36) implies that

$$w^i_{t,n}(b^i) \leq \exp\left(-\eta n \lambda_{t,n}\right) \frac{w^i_{t,n}(0)}{w^i_{\tau^i_n,n}(0)} w^i_{\tau^i_n,n}(b^i).$$

Therefore,

$$\begin{aligned}
x^i_{t,n}(0) &= \frac{w^i_{t,n}(0)}{\sum_{b^i \in A} w^i_{t,n}(b^i)} \\
&= \frac{w^i_{t,n}(0)}{w^i_{t,n}(0) + \sum_{b^i \neq 0} w^i_{t,n}(b^i)} \\
&\geq \frac{w^i_{t,n}(0)}{w^i_{t,n}(0) + \sum_{b^i \neq 0} \exp\left(-\eta n \lambda_{t,n}\right) \frac{w^i_{t,n}(0)}{w^i_{\tau^i_n,n}(0)} w^i_{\tau^i_n,n}(b^i)} \\
&= \frac{w^i_{\tau^i_n,n}(0)}{w^i_{\tau^i_n,n}(0) + \sum_{b^i \neq 0} \exp\left(-\eta n \lambda_{t,n}\right) w^i_{\tau^i_n,n}(b^i)} \\
&= \frac{x^i_{\tau^i_n,n}(0)}{x^i_{\tau^i_n,n}(0) + \sum_{b^i \neq 0} \exp\left(-\eta n \lambda_{t,n}\right) x^i_{\tau^i_n,n}(b^i)} \\
&= \frac{x^i_{\tau^i_n,n}(0)}{x^i_{\tau^i_n,n}(0) + \left(1 - x^i_{\tau^i_n,n}(0)\right) \exp\left(-\eta n \lambda_{t,n}\right)},
\end{aligned}$$

which proves the lemma. □

Lemma 8.4. *There exists $B^i > 0$ such that*

$$x^i_{\tau^i_n,n}(0) \geq \frac{1}{1 + B^i \exp(-\eta n)}. \tag{8.41}$$

Proof. Let $i \in I$. We can define

$$Z^i \coloneqq \max_{n \leq 2L^2 \text{ s.t. } \tau^i_n < +\infty} \frac{1}{x^i_{\tau^i_n,n}(0)} - 1.$$

For $n = 1$, we initialize the algorithm uniformly so every action has initially a strictly positive weight. By definition of EWA, if an action has a strictly positive weight during the initialization then it is always played with strictly positive probability. When reaching the level $n = 2$, the initialization is done by copying the current distribution of the algorithm of level $n = 1$, hence each action has a strictly positive weight too. Induction proves that at every stage and for every level, the probability to play every action is strictly positive—and strictly lower than 1. It follows that Z^i is strictly positive as the minimum of finitely many strictly positive numbers.

By definition, for every $n \leq 2L^2$ such that $\tau^i_n < +\infty$, one has

$$x^i_{\tau^i_n,n} \geq \frac{1}{1 + Z^i} \geq \frac{1}{1 + Z^i \exp(\eta 2L^2) \exp(-\eta n)},$$

so let $B^i := Z^i \exp(\eta 2L^2)$.

We now prove that this is true also for $n > 2L^2$. The proof is by induction. Assume that it is true for $n \geq 2L^2$ and consider $n+1$ such that $\tau^i_{n+1} < +\infty$. Since the increment in the number of jobs is at most one, this implies that $\tau^i_n < +\infty$.

By Lemma 8.3, for all $t > \tau^i_n$, the weight on 0 satisfies

$$\begin{aligned}
x^i_{t,n}(0) &\geq \frac{x^i_{\tau^i_n,n}(0)}{x^i_{\tau^i_n,n}(0) + \left(1 - x^i_{\tau^i_n,n}(0)\right)\exp\left(-\eta n \lambda_{t,n}\right)} \\
&= \frac{1}{1 + \left(\frac{1}{x^i_{\tau^i_n,n}(0)} - 1\right)\exp\left(-\eta n \lambda_{t,n}\right)} \\
&\geq \frac{1}{1 + (1 + B^i \exp(-\eta n) - 1)\exp\left(-\eta n \lambda_{t,n}\right)} \qquad (8.42) \\
&\geq \frac{1}{1 + B^i \exp\left(-\eta n (1 + \lambda_{t,n})\right)} \\
&\geq \frac{1}{1 + B^i \exp\left(-\eta (n+1)\right)} \qquad (8.43)
\end{aligned}$$

where Eq. (8.42) follows from the recurrence hypothesis and Eq. (8.43) is implied by $t > \tau^i_n$, so $\lambda_{t,n} \geq 1$.

The initial weight when reaching $n+1$ for the first time is equal to the current weight for n packages, it follows that

$$x^i_{\tau^i_{n+1},n+1}(0) = x^i_{\tau^i_{n+1},n}(0) \geq \frac{1}{1 + B^i \exp(-\eta(n+1))}. \qquad (8.44)$$

This proves the result for $n+1$. Hence, it concludes the induction and proves the lemma. □

8.4.3 Reinforced Random Walks

Lemma 8.3 shows that every time the process reaches a given level, there is a reinforcement on the probability to play the action profile where every player plays 0. The next step is to understand how this reinforcement influences the system dynamic. In order to do so, we prove some results on reinforced random walks. We follow the presentation of (Menshikov et al., 2017, p. 47) of nearest neighbor one-dimension random walk. They study random walks that are non-homogeneous *in space* but homogeneous *in time*. The difference is that we suppose there is a reinforcement factor in the drift, resulting in a random walk that is non-homogeneous *in time and space*, but bounded. Furthermore, we suppose that there is more heterogeneity in the weight of our random walk, in the sense that precise probabilities of going up or down are highly dependent of the past but nevertheless bounded.

The proof of Theorem 8.4 requires the following lemma, whose proof can be found in Section 8.6.

Lemma 8.5 (Reinforced Random Walk)**.** *Let $M > 0$ and $(X_t, Z_t, t \geq 0)$ a sequence of random variables in $\mathbb{N}$ such that $X_t/d \leq Z_t \leq X_t$ with $d > 1$, $|X_{t+1} - X_t| \leq M$ and $\mathcal{F}_t = \sigma(X_0, Z_0, \ldots, Z_t, X_t)$. Suppose that there exists a function $r : \mathbb{N}^2 \to \mathbb{R}^+$, reals z_0 and $A > 0$ such that:*

- *for all $t \geq 0$ and $Z_t \geq z_0$, $\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] \leq r(Z_t, \lambda_{t,Z_t})$ almost surely, where $\lambda_{t,z}$ is the number of occurrences of the $Z_t = z$ event for $t' \leq t$,*
- *for $z \geq z_0$, $\sum_m r(z,m) < \frac{A}{z}$*

Then X_t is almost surely bounded.

Proof of Theorem 8.4. Let $X_t := Ik_t$ and $Z_t := \max_{j \in I}\left(Ik^j_t + j\right)$ be random variables that are an encoding of k_t, k^j_t, j where j maximizes k^j_t and is maximal among the maximizers. Indeed, $k_t = X_t/I$, $k^j_t = [Z_t/I]$ and $j = Z_t \mod I$.

By definition,

$$Z_t = \max_{j \in I}\left(Ik^j_t + j\right) \leq \max_{j \in I}\left(I(k_t - 1) + j\right) \leq Ik_t = X_t. \qquad (8.45)$$

Moreover,

$$Z_t \geq \max_{j \in I} I k_t^j \geq I \max_{j \in I} k_t^j \geq I \frac{k_t}{I} = \frac{X_t}{I}. \tag{8.46}$$

Let $z_0 = 2IL^2$. Suppose $Z_t \geq z_0$ and let j such that $Z_t = Ik_t^j + j$. In the following, we write k^j for k_t^j and τ^j for $\tau_{k^j}^j$. Then the probability that j plays 0 is $x_{t,k^j}^j(0)$ which by Lemma 8.3 satisfies

$$x_{t,k^j}^j(0) \geq \frac{x_{\tau^j,k^j}^j(0)}{x_{\tau^j,k^j}^j(0) + \left(1 - x_{\tau^j,k^j}^j(0)\right) \exp\left(-\eta k^j \lambda_{k^j,t}\right)} \tag{8.47}$$

$$= \frac{1}{1 + \left(\frac{1}{x_{\tau^j,k^j}^j(0)} - 1\right) \exp\left(-\eta k^j \lambda_{k^j,t}\right)}. \tag{8.48}$$

Moreover, by Lemma 8.4, there exists $B > 0$ such that $x_{\tau^j,k^j}^j(0) \geq \frac{1}{1+B\exp(-\eta k^j)}$, hence from Eq. (8.48)

$$x_{t,k^j}^j(0) \geq \frac{1}{1 + B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})}. \tag{8.49}$$

At each period, there are I new jobs. By Assumption 8.1, there are less than L new jobs. Since j has more than $[z_0/I] = 2L^2$ jobs, when j plays 0, we know that at least L jobs are not late. Therefore, the number of jobs at the next period has to be smaller or equal compared to the current period. Hence,

$$\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] \leq 1 - x_{t,k^j}^j(0) \leq \frac{B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})}{1 + B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})} \tag{8.50}$$

using Eq. (8.49).

This suggests the following definition,

$$r(Z_t, \lambda_{Z_t,t}) := B\exp(-\eta[Z_t/I])\exp(-\eta[Z_t/I]\lambda_{Z_t,t}) \tag{8.51}$$

which is equal to

$$B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{Z_t,t}),$$

because $j = Z_t \mod I$, $k^j = [Z_t/I]$ and consequently, $\lambda_{Z_t,t}$ (the number of times Z_t was equal to the current value) is lower than $\lambda_{k^j,t}$ (the number of times that j had the current number of jobs). Therefore,

$$\exp(-\eta k^j \lambda_{k^j,t}) \leq \exp(-\eta k^j \lambda_{Z_t,t}). \tag{8.52}$$

Using previous equations,

$$\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] \leq B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t}) \tag{8.53}$$

$$\leq B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{Z_t,t}) \tag{8.54}$$

$$\leq r(Z_t, \lambda_{Z_t,t}) \tag{8.55}$$

where 8.53 comes from Eq. (8.50), 8.54 from Eq. (8.52) and 8.55 from Eq. (8.51).

For all $z \geq z_0$, the sum on m of $r(z,m)$ is

$$B\exp(-\eta[z/I])\frac{1}{1-\exp(-\eta[z/I])},$$

so it is bounded by A/z for some $A > 0$ and Lemma 8.5 applies, so we proved the theorem. □

8.5 Conclusion

We have studied a repeated strategic queueing model with spillover from one period to another. We have focused on the stability of the system when players play learning strategies. Several problems remain open in this model.

Multi-Level regret We have used a multi-level EWA. Although MLEWA is based on a no-regret algorithm, to prove that it is *itself* a no-regret algorithm, we would need an appropriate definition of regret in the context of endogenously changing states. Proving the no-regret property of MLEWA with a suitable definition of regret and using it to prove the system stability would be an interesting generalization. Another promising research direction is the definition of other multi-level algorithms based on different no-regret algorithms. In particular, it would be important to see which stability properties depend on the specific algorithm used and which other properties are general and hold for every no-regret algorithm.

Model Several extensions of the model are conceivable. For instance, a model with more than one server could be studied. In that case the strategy of each player would have two components: the chosen server and the chosen time at which jobs join the chosen server's queue. An apparently simple, but non-trivial generalization would involve the consideration of lower penalty costs.

Importance of the value of C_k Several results of this chapter are based on the value of the penalty C_k. If it is large enough, then the system is stable. We conjecture that a large but constant penalty is not sufficient for the stability in the learning context; i.e., the penalty must depend on the number of jobs in the system, otherwise the number of jobs could be unbounded with a positive probability. This contrasts with the myopic strategic case, where a constant penalty cost guarantees stability, if it is large enough.

8.6 Proof of the Reinforced Random Walk Lemma

Proof of Lemma 8.5. The following proof is inspired by Pemantle (2007).

We first prove that the probability that $\sup_{t\in\mathbb{N}} X_t \in [x-M, x]$ for all x such that $x \geq x_0 + M$, conditionally on the fact that $[x-M, x]$ is reached, is lower bounded by something strictly greater than 0 and that does not depend on x. As we will show, this implies that almost surely there exists x such that $\sup_{t\in\mathbb{N}} X_t = x$.

Denote the event that $[x-M, x]$ is reached by X_t by $A(x)$. We now show that for all $x \geq x_0 + M$:

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x-M, x] \mid A(x)\right] \geq \prod_{z\in\left[\frac{x-M}{d}, x\right]} \prod_{m\in\mathbb{N}} (1 - r(z, m)) \tag{8.56}$$

To prove this result, we fix $x \in \mathbb{N}$ such that $x > dz_0 + M$ and we introduce an extended random process. Define the new state space $\overline{\Omega} = \mathbb{N}^2 \times \mathbb{N} \times \mathbb{N}^x \times \{0, 1\}$. The interpretation of the a state $(x, z, n_0, ..., n_x, i)$ is the following:

- the current state is (x, z),
- the path of Z_t has gone n_r times through the state r,
- i is equal to 1 if and only if X_t went up from a state between x and $x - M$.

Formally, let $(X, Y, N_0, ..., N_x, I)_{t\geq 1}$ be the random process on $\overline{\Omega}$. The first coordinate is equal to X_t whereas all other coordinates are deduced from it. By construction, we know that X_t has a maximal increment of M, hence in order for the supremum to be strictly greater than x, it is necessary for a positive jump from a state between $x - M$ and x, hence

$$\{\sup_{t\in\mathbb{N}} X_t > x\} \subset \{\exists t \geq 1, I_t = 0\}.$$

It follows that

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x-M, x] \mid A(x)\right] \geq \mathbb{P}\left[\forall t \geq 1, I_t = 0 \mid A(x)\right].$$

Moreover,

$$\mathbb{P}\left[\forall t \geq 1, I_t = 0 \mid A(x)\right] \geq \prod_{z\in[\frac{x-M}{d}, x]} \prod_{m\in\mathbb{N}} (1 - r(z, m)).$$

Indeed, by construction of the auxiliary random process, we know that:

- for every $r \in \{0, ..., M\}$, N_r is only increasing,
- conditionally on $Z_t = z \geq \frac{x-M}{d}$, $N_z = n$ and the past, the probability for i to stay equal to 0 is at least $(1 - r(z, n))$,
- conditionally on $z < \frac{x-M}{d}$, the probability for i to stay equal to 0 is 1.

It follows that

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x-M, x] \mid A(x)\right] \geq \mathbb{P}\left[\forall t \geq 1, I_t = 0 \mid A(x)\right] \tag{8.57}$$

$$\geq \prod_{z\in[\frac{x-M}{d},x]} \prod_{m\in\mathbb{N}} (1 - r(z, m)). \tag{8.58}$$

Then, the logarithm of the right hand side is

$$\begin{aligned}
\sum_{z\in[\frac{x-M}{d},x]} \sum_{m\in\mathbb{N}} \log(1 - r(z, m)) &= \sum_{z\in[\frac{x-M}{d},x]} \sum_{m\in\mathbb{N}} -\log(1 + \frac{r(z, m)}{1 - r(z, m)}) \\
&\geq \sum_{z\in[\frac{x-M}{d},x]} \sum_{m\in\mathbb{N}} -\frac{r(z, m)}{1 - r(z, m)} \\
&\geq \sum_{z\in[\frac{x-M}{d},x]} \sum_{m\in\mathbb{N}} -\frac{r(z, m)}{1 - \rho} \\
&= \sum_{z\in[\frac{x-M}{d},x]} -\frac{A}{z(1 - \rho)} \\
&\geq -\frac{A}{1 - \rho} \sum_{z\in[\frac{x-M}{d},x]} \frac{d}{x - M} \\
&= -\frac{A}{1 - \rho} \left(x - \frac{x - M}{d} + 1\right) \frac{d}{x - M} \\
&= -\frac{A}{1 - \rho} \frac{(d - 1)x + M + d}{x - M} \geq -B > 0,
\end{aligned}$$

where $B > 0$ is a positive constant which does not depend on x.

It follows that

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x-M, x] \mid A(x)\right] \geq \exp(-B) > 0. \tag{8.59}$$

The probability that the upper bound of X_t belongs to $[x - M, x]$ is therefore lower bounded by a constant independent of x conditionally on the fact that this interval is reached.

$$\begin{aligned}
\mathbb{P}\left[\sup_t X_t < \infty\right] &= \sum_{k\geq 1} \mathbb{P}\left[\sup_t X_t \in [(k-1)M, kM]\right] \\
&\geq \sum_{k\geq[\frac{x_0}{M}]+1} \mathbb{P}\left[\sup_t X_t \in [(k-1)M, kM] \mid A(kM)\right] \mathbb{P}\left[A(kM)\right].
\end{aligned}$$

However, if X_t is unbounded, then $A(kM)$ happens, so $\mathbb{P}\left[A(kM)\right] \geq \mathbb{P}\left[\sup_t X_t = +\infty\right]$. Therefore, using (8.59):

$$\mathbb{P}\left[\sup_t X_t < \infty\right] \geq \sum_{k\geq[\frac{x_0}{M}]} \exp(-B)\, \mathbb{P}\left[\sup_t X_t = +\infty\right].$$

The right hand side is equal to ∞ if $\mathbb{P}\left[\sup_t X_t = +\infty\right] > 0$, so necessarily $\mathbb{P}\left[\sup_t X_t = +\infty\right] = 0$. □

Chapter 9

Conclusion

The contributions of this thesis are threefold, all related to learning in stochastic or repeated games. First, we carried out an empirical study of the behavior of systems where several players use Q-learning-like algorithms in games. Then, we extended well-known procedures, Fictitious Play (FP) and Smooth Fictitious Play (SFP), to multiple procedures in zero-sum and identical-interests stochastic games. We showed that several of these procedures converged. The last chapter of this thesis is dedicated to the study of a particular stochastic game with a single queue when players use a no-regret procedure.

We first review these contributions and then describe two open problems that arose during my PhD.

Q-learning in Repeated Games In Chapter 5, systems with players using Q-learning in various games were simulated. We showed that memory, formalized as a state variable, could improve cooperation in a non-convex version of the Prisoner Dilemma, whereas this bounded memory degrades cooperation in the (standard) Prisoner Dilemma, somewhat counterintuitively.

Moreover, we studied the influence of exploration schemes on the cooperation in the Smoothed Bertrand Competition game, extending the study of Calvano et al. (2020). Indeed, the original paper of Calvano et al. (2020) supposed that players used an exponentially-decreasing exploration scheme. Such an exploration scheme implies that the nubmer of exploration times is finite, which is not true in other exploration schemes, for instance the ϵ-greedy one. Simulations with different exploration schemes yield different cooperation results, but it appears that cooperation still emerge with all schemes tested, albeit with relatively extreme parameters.

Fictitious Play and Smooth Fictitious Play in Zero-Sum and Identical-Interest Stochastic Games Based on ideas from Leslie et al. (2020) and Sayin et al. (2022a), we proposed, in Chapter 6, an extension of FP to stochastic games. It is shown to converge in both zero-sum and identical-interest stochastic games using the theory of stochastic approximations for most proofs, and therefore continuous time dynamics were also defined and proved to converge.

Then, we defined in Chapter 7 a generalization of SFP for the same classes of stochastic games. This resulted in a number of decentralized, continuous and discrete time systems. These systems feature exploration in order to learn, for a player, both its own payoff function and the transition probability function. We showed that, when used jointly, our algorithms converge to regularized (and so ϵ-approximate) stationary Nash equilibria in discounted stochastic games.

No-Regret Procedures in Queuing Models Introduced by the seminal paper of Naor (1969), strategic queuing systems model situations where a number of agents have packets–also called jobs–that they must submit to a queue. Jobs leave the queue in a first-in first-out order.

In Chapter 8, we studied a queuing game inspired by the class of problems first examined by Glazer and Hassin (1983), where players that do not get their jobs through before a fixed deadline get a deterrent penalty. More precisely, we investigated the behavior of such a system in two cases: first, when players are strategic, meaning that they jointly play a (Nash, correlated or coarse correlated) equilibrium. Second, we supposed that all players use the same procedure that is inspired from the exponential weight algorithm (EWA), known to be no-regret. Contrary to standard EWA, our multi-level

EWA (MLEWA) deals with multiple states, called levels. In this case, a level is the number of jobs held by a player.

We showed that in both cases, if the penalty is deterrent enough, then the system is stable, i.e., the number of jobs is bounded. Therefore, this can be thought of as a mechanism design problem were the system designer has to put sufficiently high penalties in order to make the number of jobs bounded.

Our model could be extended in several directions. First, players could use different algorithms from MLEWA, perhaps multi-level algorithms based on other procedures–for instance FP. The proof of stability could be abstracted from MLEWA to understand what are the features required for a procedure to converge.

Second, an obvious generalization of the game itself would comprise several queues, i.e., multiple servers similarly to Gaitonde and Tardos (2020b). Therefore, the server choice would be an additional strategic ingredient in this game.

Third, bounds on the penalty could be refined. In particular, we do not know whether there are intermediate cases between the stable scenario and the scenario where the limit is infinity (i.e., the diverging one). Is it possible for the system to be infinitely recurrent but not stable?

Open Problems A few other open problems surfaced during this thesis in multiple chapters:

- first, the convergence of simple multiagent systems, such as repeated games with Q-learning, as experimented in Chapter 5. While there were previous attempts to study empirically this system (Babes et al., 2009; Wunder et al., 2010), there is no formal guarantee (Sutton and Barto, 2018), even in simple classes of games.

  In Chapters 6 and 7, we established convergence of several classes of algorithms in identical-interest and zero-sum stochastic games. Most of these algorithms use two-timescales (for the continuations variables and the empirical actions variables). However, the theory of stochastic approximations as outlined in (Benaim et al., 2005) may be used directly in non-autonomous systems. Therefore, the correspondence between arbitrary two-timescales continuous-time systems and discrete-time systems is not straightforward. Extending these results would make the choice of timescales more flexible.

  Another key point is the type of convergence: as explained above, we focused on time-average convergence for Chapters 5 to 7. The last-iterate convergence property characterizes the behavior of the system with certainty, in contrast to the time-average convergence which may feature out-of-equilibria in the long run. Anagnostides et al. (2022) showed that *optimistic* variants of well-known algorithms (in this case mirror descent and gradient descent) have this property in potential and zero-sum games.

- second, the definition of regret in changing environment when the transition is *endogenous*. Indeed, the regret framework fits to the case where the environment is oblivious, meaning that there may be a state variable, but players do not influence it. Instead, if we suppose that the state variable is (at least partially) controlled by players, then the standard no-regret framework is not suitable: a change in actions results in a change in the instantaneous reward but also on the state.

  There is no widely accepted definition of regret in stochastic games with endogenous transitions. Wei et al. (2017) defines regret as the difference between the value of the stochastic game and the achieved payoff. Gaitonde and Tardos (2020b) uses the difference between the achieved payoff and the sum of payoffs that would have been obtained if the player had changed its actions but with the same states, i.e., states are exogenous.

  Indeed, a key difficulty is estimating what would have happened at time t if the current state was different because of different, previous action choices. Another one is the choice of the payoff function.

  At a more abstract level, this would answer the following question: supposing that we have a number of experts that recommend actions in a number of states, which experts should be selected in order to maximize the discounted payoff on the long run?

  Regarding our work, it would be of interest to prove a regret-like properties for SFP. Indeed, the fact that it is randomized should make it more robust to adversarial players, which is usually expressed as being no-regret. In stochastic-games, what is the appropriate notion to use?

Bibliography

[1] S. M. Amadae. *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. Cambridge: Cambridge University Press, 2015.

[2] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. *On Last-Iterate Convergence Beyond Zero-Sum Games*. Mar. 22, 2022. arXiv: 2203.12056 [cs]. URL: http://arxiv.org/abs/2203.12056 (visited on 10/24/2022). preprint.

[3] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. *On the Convergence of No-Regret Learning Dynamics in Time-Varying Games*. Jan. 26, 2023. arXiv: 2301.11241 [cs]. URL: http://arxiv.org/abs/2301.11241 (visited on 03/09/2023). preprint.

[4] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. *On the convergence of no-regret learning dynamics in time-varying games*. Tech. rep. arXiv:2301.11241, 2023.

[5] Itai Arieli and Yakov Babichenko. "Average Testing and Pareto Efficiency". In: *Journal of Economic Theory* 147.6 (Nov. 2012), pp. 2376–2398.

[6] John Asker, Chaim Fershtman, and Ariel Pakes. *Artificial Intelligence and Pricing: The Impact of Algorithm Design*. 28535. National Bureau of Economic Research, Inc, Mar. 2021.

[7] Stephanie Assad, Robert Clark, Daniel Ershov, and Lei Xu. "Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market". In: *SSRN Electronic Journal* (2020).

[8] Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Vol. 264. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 1984.

[9] Robert J. Aumann. "Subjectivity and Correlation in Randomized Strategies". In: *Journal of Mathematical Economics* 1.1 (1974), pp. 67–96.

[10] Monica Babes, Michael Wunder, and Michael Littman. "Q-Learning in Two-Player Two-Action Games". In: (2009), p. 4.

[11] Lucas Baudin and Rida Laraki. "Fictitious Play and Best-Response Dynamics in Identical-Interest and Zero-Sum Stochastic Games". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 17–23, 2022, pp. 1664–1690.

[12] Lucas Baudin and Rida Laraki. "Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions". In: *Advances in Neural Information Processing Systems*. 2022.

[13] Lucas Baudin, Marco Scarsini, and Xavier Venel. *Strategic Behavior and No-Regret Learning in Queueing Systems*. Preprint, 2023. arXiv: 2302.03614 [cs.GT].

[14] Richard Bellman. "A Markovian Decision Process". In: *Journal of Mathematics and Mechanics* 6.5 (1957), pp. 679–684. JSTOR: 24900506.

[15] Michel Benaim. "A Dynamical System Approach to Stochastic Approximations". In: *SIAM Journal on Control and Optimization* 34.2 (Mar. 1996), pp. 437–472.

[16] Michel Benaim, Josef Hofbauer, and Sylvain Sorin. "Stochastic Approximations and Differential Inclusions". In: *SIAM Journal on Control and Optimization* 44.1 (Jan. 2005), pp. 328–348.

[17] Michel Benaïm and Mathieu Faure. "Consistency of Vanishingly Smooth Fictitious Play". In: *Mathematics of Operations Research* 38.3 (Aug. 2013), pp. 437–450.

[18] Ulrich Berger. "Brown's Original Fictitious Play". In: *Journal of Economic Theory* 135.1 (July 2007), pp. 572–578.

[19] Ulrich Berger. "Fictitious Play in 2×n Games". In: *Journal of Economic Theory* 120.2 (Feb. 2005), pp. 139–154.

[20] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Non-Stationary Stochastic Optimization". In: *Operations Research* 63.5 (Oct. 2015), pp. 1227–1244.

[21] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Non-stationary stochastic optimization". In: *Oper. Res.* 63.5 (2015), pp. 1227–1244.

[22] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Optimal exploration-exploitation in a multi-armed bandit problem with non-stationary rewards". In: *Stoch. Syst.* 9.4 (2019), pp. 319–337.

[23] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Optimal Exploration–Exploitation in a Multi-armed Bandit Problem with Non-stationary Rewards". In: *Stochastic Systems* 9.4 (Dec. 2019), pp. 319–337.

[24] David Blackwell. "An Analog of the Minimax Theorem for Vector Payoffs". In: *Pacific Journal of Mathematics* 6.1 (Mar. 1, 1956), pp. 1–8.

[25] David Blackwell. "Controlled Random Walks". In: *Proceedings of the International Congress of Mathematicians, 1954* III (1956), pp. 336–338.

[26] Avrim Blum and Yishay Mansour. "From External to Internal Regret". In: *Learning Theory*. Ed. by Peter Auer and Ron Meir. Red. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, and Gerhard Weikum. Vol. 3559. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 621–636.

[27] Vivek S. Borkar. "Asynchronous Stochastic Approximations". In: *SIAM Journal on Control and Optimization* 36.3 (May 1998), pp. 840–851.

[28] Vivek S. Borkar. "Stochastic Approximation with Two Time Scales". In: *Systems & Control Letters* 29.5 (Feb. 1997), pp. 291–294.

[29] Michael Bowling and Manuela Veloso. "Rational and Convergent Learning in Stochastic Games". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 1021–1026.

[30] George W Brown. "Iterative Solution of Games by Fictitious Play". In: *Activity analysis of production and allocation* 13.1 (1951), pp. 374–376.

[31] Zach Y Brown and Alexander MacKay. "Competition in Pricing Algorithms". In: (Oct. 2021), p. 65.

[32] Lucian Busoniu, Robert Babuska, and Bart De Schutter. "A Comprehensive Survey of Multiagent Reinforcement Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2 (Mar. 2008), pp. 156–172.

[33] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. "Artificial Intelligence, Algorithmic Pricing, and Collusion". In: *American Economic Review* 110.10 (Oct. 1, 2020), pp. 3267–3297.

[34] Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. "Competing against Nash equilibria in adversarially changing zero-sum games". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 921–930.

[35] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge; New York: Cambridge University Press, 2006.

[36] Vincent Conitzer and Tuomas Sandholm. "AWESOME: A General Multiagent Learning Algorithm That Converges in Self-Play and Learns a Best Response against Stationary Opponents". In: *Machine Learning* 67.1-2 (May 2007), pp. 23–43.

[37] Ludovico Crippa, Yonatan Gur, and Bar Light. *Regret minimization with dynamic benchmarks in repeated games*. Tech. rep. arXiv:2212.03152, 2022.

[38] Constantinos Daskalakis and Ioannis Panageas. *Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization*. Dec. 2, 2020. arXiv: 1807.04252 [cs, math, stat]. URL: http://arxiv.org/abs/1807.04252 (visited on 10/26/2022). preprint.

[39] Stefano DeMichelis and Fabrizio Germano. "On the Indices of Zeros of Nash Fields". In: *Journal of Economic Theory* 94.2 (2000), pp. 192–217.

[40] Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. "Multiagent Online Learning in Time-Varying Games". In: *Mathematics of Operations Research* (July 1, 2022), moor.2022.1283.

[41] Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. "Multiagent online learning in time-varying games". In: *Math. Oper. Res.* forthcoming (2022).

[42] Paul Erickson. *The World the Game Theorists Made*. Chicago : London: The University of Chicago Press, 2015. 390 pp.

[43] A. M. Fink. "Equilibrium in a Stochastic n-Person Game". In: *Hiroshima Mathematical Journal* 28.1 (Jan. 1, 1964).

[44] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. "Counterfactual Multi-Agent Policy Gradients". In: *AAAI*. 2018.

[45] Dean P. Foster. "A Proof of Calibration via Blackwell's Approachability Theorem". In: *Games and Economic Behavior* 29.1-2 (Oct. 1999), pp. 73–78.

[46] Dean P. Foster and Rakesh V. Vohra. "Calibrated Learning and Correlated Equilibrium". In: *Games and Economic Behavior* 21.1-2 (Oct. 1997), pp. 40–55.

[47] Dean P. Foster and H. Peyton Young. "Learning, Hypothesis Testing, and Nash Equilibrium". In: *Games and Economic Behavior* 45.1 (Oct. 2003), pp. 73–96.

[48] Drew Fudenberg and David K. Levine. "Consistency and Cautious Fictitious Play". In: *Journal of Economic Dynamics and Control* 19.5-7 (July 1995), pp. 1065–1089.

[49] Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press Series on Economic Learning and Social Evolution 2. Cambridge, Mass: MIT Press, 1998. 276 pp.

[50] Drew Fudenberg and Eric Maskin. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information". In: *Econometrica* 54.3 (May 1986), p. 533. JSTOR: 1911307.

[51] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, 1991.

[52] Jason Gaitonde and Éva Tardos. "Stability and learning in strategic queuing systems". In: *Proceedings of the 21st ACM Conference on Economics and Computation*. EC '20. Virtual Event, Hungary: Association for Computing Machinery, 2020, pp. 319–347.

[53] Jason Gaitonde and Éva Tardos. "Virtues of patience in strategic queuing systems". In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21. Budapest, Hungary: Association for Computing Machinery, 2021, pp. 520–540.

[54] Jason Gaitonde and Eva Tardos. "Stability and Learning in Strategic Queuing Systems". Mar. 15, 2020. arXiv: 2003.07009 [cs, math].

[55] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and P. Mertikopoulos. "Survival of the Strictest: Stable and Unstable Equilibria under Regularized Learning with Partial Information". In: *Annual Conference Computational Learning Theory*. 2021.

[56] Amihai Glazer and Refael Hassin. "$?/M/1$: on the equilibrium distribution of customer arrivals". In: *European J. Oper. Res.* 13.2 (1983), pp. 146–150.

[57] Saeed Hadikhanloo, Rida Laraki, Panayotis Mertikopoulos, and Sylvain Sorin. *Learning in Nonatomic Games, Part I: Finite Action Spaces and Population Games*. 2021.

[58] Peter Hammerstein and Susan E. Riechert. "Payoffs and Strategies in Territorial Contests: ESS Analyses of Two Ecotypes of the spiderAgelenopsis Aperta". In: *Evolutionary Ecology* 2.2 (Apr. 1988), pp. 115–138.

[59] James Hannan. "Approximation to Bayes Risk in Repeated Play". In: *Contributions to the Theory of Games* 3 (1957), pp. 97–139.

[60] Christopher Harris. "On the Rate of Convergence of Continuous-Time Fictitious Play". In: *Games and Economic Behavior* 22.2 (Feb. 1998), pp. 238–259.

[61] John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass: MIT Press, 1988. 378 pp.

[62] Sergiu Hart and Andreu Mas-Colell. "A General Class of Adaptive Strategies". In: *Journal of Economic Theory* 98.1 (May 2001), pp. 26–54.

[63] Sergiu Hart and Andreu Mas-Colell. "A Simple Adaptive Procedure Leading to Correlated Equilibrium". In: *Econometrica* 68.5 (2000), pp. 1127–1150. JSTOR: 2999445.

[64] Sergiu Hart and Andreu Mas-Colell. *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*. World Scientific Series in Economic Theory v. 4. New Jersey: World Scientific, 2013. 296 pp.

[65] Sergiu Hart and Andreu Mas-Colell. "Uncoupled Dynamics Do Not Lead to Nash Equilibrium". In: *The American Economic Review* 93.5 (2003), pp. 1830–1836. JSTOR: 3132156.

[66] Sergiu Hart and Andreu Mas-Colell. "Uncoupled Dynamics Do Not Lead to Nash Equilibrium". In: *The American Economic Review* 93.5 (2003), pp. 1830–1836. JSTOR: 3132156.

[67] Hado Hasselt. "Double Q-Learning". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010.

[68] Hado van Hasselt, Arthur Guez, and David Silver. "Deep Reinforcement Learning with Double Q-Learning". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, 2016, pp. 2094–2100.

[69] Refael Hassin. "On the optimality of first come last served queues". In: *Econometrica* 53.1 (1985), pp. 201–202.

[70] Refael Hassin. *Rational Queueing*. Boca Raton, FL: CRC Press, 2016, pp. xiii+378.

[71] Refael Hassin and Moshe Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Boston, MA: Kluwer Academic Publishers, 2003, pp. xii+191.

[72] Moshe Haviv and Liron Ravner. "A survey of queueing systems with strategic timing of arrivals". In: *Queueing Syst.* 99.1-2 (2021), pp. 163–198.

[73] Josef Hofbauer. "From Nash and Brown to Maynard Smith: Equilibria, Dynamics and ESS". In: *Selection* 1.1-3 (Jan. 2001), pp. 81–88.

[74] Josef Hofbauer and Ed Hopkins. "Learning in Perturbed Asymmetric Games". In: *Games and Economic Behavior* 52.1 (July 2005), pp. 133–152.

[75] Josef Hofbauer and William H. Sandholm. "On the Global Convergence of Stochastic Fictitious Play". In: *Econometrica* 70.6 (2002), pp. 2265–2294. JSTOR: 3081987.

[76] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press, 1998.

[77] Junling Hu and Michael P. Wellman. "Nash Q-Learning for General-Sum Stochastic Games". In: *The Journal of Machine Learning Research* 4 (null Dec. 1, 2003), pp. 1039–1069.

[78] L. P. Kaelbling, M. L. Littman, and A. W. Moore. "Reinforcement Learning: A Survey". In: *Journal of Artificial Intelligence Research* 4 (May 1, 1996), pp. 237–285.

[79] Sham M. Kakade and Dean P. Foster. "Deterministic Calibration and Nash Equilibrium". In: *Journal of Computer and System Sciences* 74.1 (Feb. 2008), pp. 115–130.

[80] Ryo Kawasaki, Hideo Konishi, and Junki Yukawa. "Equilibria in bottleneck games". In: *Internat. J. Game Theory* forthcoming (2023).

[81] Ardeshir Kianercy and Aram Galstyan. "Dynamics of Boltzmann Q Learning in Two-Player Two-Action Games". In: *Physical Review E* 85.4 (Apr. 26, 2012), p. 041145.

[82] J. Kiefer and J. Wolfowitz. "Stochastic Estimation of the Maximum of a Regression Function". In: *The Annals of Mathematical Statistics* 23.3 (Sept. 1952), pp. 462–466.

[83] Vijaymohan R. Konda and Vivek S. Borkar. "Actor-Critic–Type Learning Algorithms for Markov Decision Processes". In: *SIAM Journal on Control and Optimization* 38.1 (Jan. 1999), pp. 94–123.

[84] Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Red. by Fritz John, Lawrence Sirovich, Joseph P. LaSalle, and Gerald B. Whitham. Vol. 26. Applied Mathematical Sciences. New York, NY: Springer New York, 1978.

[85] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. *Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games*. Sept. 28, 2021. arXiv: 2106.01969 [cs]. URL: http://arxiv.org/abs/2106.01969 (visited on 05/17/2022). preprint.

[86] David S. Leslie and E J Collins. "Individual Q-Learning in Normal Form Games". In: *SIAM Journal on Control and Optimization* 44.2 (2005), p. 20.

[87] David S. Leslie and E. J. Collins. "Generalised Weakened Fictitious Play". In: *Games and Economic Behavior* 56.2 (Aug. 1, 2006), pp. 285–298.

[88] David S. Leslie, Steven Perkins, and Zibo Xu. "Best-Response Dynamics in Zero-Sum Stochastic Games". In: *Journal of Economic Theory* 189 (Sept. 2020), p. 105095.

[89] Nick Littlestone and Manfred K. Warmuth. "The weighted majority algorithm". In: *Inform. and Comput.* 108.2 (1994), pp. 212–261.

[90] Michael L. Littman. "Markov Games as a Framework for Multi-Agent Reinforcement Learning". In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 157–163.

[91] Jason R. Marden, Gürdal Arslan, and Jeff S. Shamma. "Joint Strategy Fictitious Play With Inertia for Potential Games". In: *IEEE Transactions on Automatic Control* 54.2 (Feb. 2009), pp. 208–220.

[92] Mikhail Menshikov, Serguei Popov, and Andrew Wade. *Non-Homogeneous Random Walks: Lyapunov Function Methods for Near-Critical Stochastic Systems*. Cambridge: Cambridge University Press, 2017.

[93] Panayotis Mertikopoulos. "Online Optimization and Learning in Games: Theory and Applications". Habilitation à diriger des recherches. Université Grenoble Alpes, 2019.

[94] Panayotis Mertikopoulos and William H. Sandholm. "Learning in Games via Reinforcement and Regularization". In: *Mathematics of Operations Research* 41.4 (Nov. 2016), pp. 1297–1324.

[95] Panayotis Mertikopoulos and Mathias Staudigl. "Equilibrium tracking and convergence in dynamic games". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. 2021, pp. 930–935.

[96] Dov Monderer and Lloyd S. Shapley. "Fictitious Play Property for Games with Identical Interests". In: *Journal of Economic Theory* 68.1 (Jan. 1996), pp. 258–265.

[97] Dov Monderer and Lloyd S. Shapley. "Potential Games". In: *Games and Economic Behavior* 14.1 (May 1, 1996), pp. 124–143.

[98] James D. Morrow. *Game Theory for Political Scientists*. Princeton, N.J: Princeton University Press, 1994. 376 pp.

[99] P. Naor. "The regulation of queue size by levying tolls". In: *Econometrica* 37.1 (1969), pp. 15–24.

[100] John F. Nash. "Equilibrium Points in *n* -Person Games". In: *Proceedings of the National Academy of Sciences* 36.1 (Jan. 1950), pp. 48–49.

[101] J. von Neumann. "Zur Theorie Der Gesellschaftsspiele". In: *Mathematische Annalen* 100 (1928), pp. 295–320.

[102] Abraham Neyman. "Continuous-Time Stochastic Games". In: *Games and Economic Behavior* 104 (July 2017), pp. 92–130.

[103] OECD. *Algorithms and Collusion : Competition Policy in the Digital Age*. 2017. URL: http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

[104] Wojciech Olszewski. "Chapter 18 - Calibration and Expert Testing". In: *Handbook of Game Theory with Economic Applications*. Ed. by H. Peyton Young and Shmuel Zamir. Vol. 4. Elsevier, Jan. 1, 2015, pp. 949–984.

[105] Robin Pemantle. "A Survey of Random Processes with Reinforcement". In: *Probability Surveys* 4 (none Jan. 1, 2007).

[106] Vianney Perchet. "Approachability, Regret and Calibration: Implications and Equivalences". In: *Journal of Dynamics & Games* 1.2 (2014), pp. 181–254.

[107] Steven Perkins. "Advanced Stochastic Approximation Frameworks and Their Applications". University of Bristol, Sept. 2013.

[108] Steven Perkins and David S. Leslie. "Asynchronous Stochastic Approximation with Differential Inclusions". In: *Stochastic Systems* 2.2 (Dec. 2012), pp. 409–446.

[109] Thomas Rivera, Marco Scarsini, and Tristan Tomala. *Efficiency of correlation in a bottleneck game*. Tech. rep. SSRN3219767, 2018.

[110] Thomas J. Rivera, Marco Scarsini, and Tristan Tomala. *Efficiency of Correlation in a Bottleneck Game*. SSRN Scholarly Paper ID 3219767. Rochester, NY: Social Science Research Network, July 25, 2018.

[111] Herbert Robbins. "Some Aspects of the Sequential Design of Experiments". In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535.

[112] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407.

[113] Julia Robinson. "An Iterative Method of Solving a Game". In: *The Annals of Mathematics* 54.2 (Sept. 1951), p. 296. JSTOR: 1969530.

[114] Muhammed O. Sayin, Francesca Parise, and Asuman Ozdaglar. *Fictitious Play in Zero-Sum Stochastic Games*. June 2, 2022. arXiv: 2010.04223 [cs, math]. URL: http://arxiv.org/abs/2010.04223 (visited on 07/27/2022). preprint.

[115] Muhammed O. Sayin, Francesca Parise, and Asuman Ozdaglar. "Fictitious Play in Zero-Sum Stochastic Games". In: *SIAM Journal on Control and Optimization* 60.4 (2022), pp. 2095–2114. eprint: https://doi.org/10.1137/21M1426675.

[116] Muhammed O. Sayin, Kaiqing Zhang, David S. Leslie, Tamer Basar, and Asuman Ozdaglar. "Decentralized Q-Learning in Zero-sum Markov Games". June 4, 2021. arXiv: 2106.02748 [cs, math].

[117] Flore Sentenac, Etienne Boursier, and Vianney Perchet. "Decentralized learning in online queuing systems". In: *Advances in Neural Information Processing Systems* 34 (2021). Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, pp. 18501–18512.

[118] Shai Shalev-Shwartz. "Online Learning and Online Convex Optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2011), pp. 107–194.

[119] L. S. Shapley. "Some Topics in Two-Person Games". In: *Advances in Game Theory. (AM-52), Volume 52*. Ed. by Melvin Dresher, Lloyd S. Shapley, and Albert William Tucker. Princeton University Press, 1964, pp. 1–28.

[120] Lloyd S. Shapley. "Stochastic Games". In: *Proceedings of the National Academy of Sciences* 39.10 (Oct. 1, 1953), pp. 1095–1100. pmid: 16589380.

[121] Yoav Shoham, Rob Powers, and Trond Grenager. "If Multi-Agent Learning Is the Answer, What Is the Question?" In: *Artificial Intelligence* 171.7 (May 2007), pp. 365–377.

[122] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go without Human Knowledge". In: *Nature* 550.7676 (Oct. 2017), pp. 354–359.

[123] Richard S. Sutton. "Learning to Predict by the Methods of Temporal Differences". In: *Machine Learning* 3.1 (Aug. 1988), pp. 9–44.

[124] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2018. 526 pp.

[125] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. "Fast Convergence of Regularized Learning in Games". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 2989–2997.

[126] Masayuki Takahashi. “Equilibrium Points of Stochastic Non-Cooperative n-Person Games”. In: *Journal of Science of the Hiroshima University, Series AI (Mathematics)* 28.1 (1964), pp. 95–99.

[127] Gerald Tesauro. “Practical Issues in Temporal Difference Learning”. In: *Machine Learning* 8.3-4 (May 1992), pp. 257–277.

[128] Fernando Vega-Redondo. *Economics and the Theory of Games*. Cambridge, UK ; New York: Cambridge University Press, 2003. 512 pp.

[129] William S. Vickrey. “Congestion theory and transport investment”. In: *Amer. Econ. Rev.* 59.2 (1969), pp. 251–260.

[130] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1944. 625 pp.

[131] O. J. Vrieze and S. H. Tijs. “Fictitious Play Applied to Sequences of Games and Discounted Stochastic Games”. In: *International Journal of Game Theory* 11.2 (June 1982), pp. 71–85.

[132] C. J. C. H. Watkins. “Learning from Delayed Rewards”. King’s College, Oxford, 1989.

[133] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. “Online Reinforcement Learning in Stochastic Games”. Dec. 2, 2017. arXiv: 1712.00579 [cs].

[134] Jörgen W. Weibull. *Evolutionary Game Theory*. 1. MIT Press paperback ed., [Nachd.] Cambridge, Mass.: MIT Press, 2004. 265 pp.

[135] Michael Wunder, Michael Littman, and Monica Babes. “Classes of Multiagent Q-Learning Dynamics with Epsilon-Greedy Exploration”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Madison, WI, USA: Omnipress, June 21, 2010, pp. 1167–1174.

[136] H. Peyton Young. “The Evolution of Conventions”. In: *Econometrica* 61.1 (Jan. 1993), p. 57. JSTOR: 2951778.

[137] Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. “No-regret learning in time-varying zero-sum games”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 26772–26808.

Glossary

Q-learning . 3, 13, 14, 19–22, 30, 35, 54–57, 59–61, 63, 64, 67–69, 74, 82, 123, 124

action a stationary element of the action set, different from a strategy which may depend on the history of play in repeated or stochastic games. 40

asynchronous In the context of FP, procedures that do not need an action for all states but only the current one. 69, 70

auxiliary Shapley game . 6

behavioral strategy Strategy that describes the behavior of a player in a repeated game or a stochastic game.. 69

CCE coarse correlated equilibrium. 106, 109, 111, 114

DQM dynamic queueing model. 108, 114

DSG discounted stochastic game. 41, 86, 88

ergodic for a stochastic game. 42

EWA exponential weight algorithm. 7, 30, 106, 107, 114, 115, 118, 123

FIFO first-in first-out. 106

FP Fictitious Play. 3, 5, 6, 14, 19, 24–27, 30, 35, 36, 48–52, 69, 72–74, 78, 81, 82, 85, 102–104, 123, 124, 133

FTL Follow the Leader. 48, 49

FTRL Follow the Regularized Leader. 48, 50

ICT internally chain transitive. 57, 58

internal regret . 52

LIFO last-in first-out. 106

MARL Multiagent Reinforcement Learning. 34, 35, 55–57

MDP Markov Decision Process. 41, 54, 59

MLEWA multi-level EWA. 7, 30, 115, 116, 121, 124

NE Nash equilibrium. 106, 109

online . 47

potential game . 40

procedure (Computable) function from histories to actions that decides which action is chosen by a given player. 43, 72

regret . 50, 52

repeated game. 40

RL reinforcement learning. 54

SBRD Smooth Best-Response Dynamics. 88

SFP Smooth Fictitious Play. 14, 24, 27, 31, 35, 49, 50, 69, 85, 103, 104, 123, 124

stationary . 41

stochastic game . 35, 40, 41, 56, 69

strategy formally an action of the repeated game, so it provides an action at every step, possibly depending on the history of play. 40

SyncFP Synchronous FP. 72

synchronous Procedures that need an action for all states at every step. 69, 70

Appendix A

Python Code of Example Games

A.1 Power Unit Commitment

```
S = tuple[int, tuple[int]]

class PowerGen(GameWithTransitions[S]):

    def __init__(self, maxPlayer: int):
        super().__init__(maxPlayer)
        self.max_temp = 5
        self.cost_change = [10, 1, 0, 0, 0][0:maxPlayer]
        self.cost_unit = [2, 6, 10, 4, 2][0:maxPlayer]
        self.min_prod = [5, 0, 0, 0, 3][0:maxPlayer]
        self.max_prod = [10, 5, 2, 1, 6][0:maxPlayer]
        self.reward = 10
        self.d:Callable[[int], int] = lambda temp: temp+5
        self.C = 100
        self.max_action_total =
            max([self.max_action((0, (0,)), i) for i in range(maxPlayer)])
        self.encode_actions =
            np.array([self.max_action_total**i for i in range(0, maxPlayer+1)])

    def rewardstuple(self, state: S, actions: tuple[int]) -> tuple[S, list[NumExpr]]:
        temp, old_actions = state
        new_temp = max(0, min(self.max_temp, temp + random.randint(-1, 1)))
        p = np.array(old_actions)
        a = np.array(actions)

        rew = min(self.d(temp), np.sum(a + self.min_prod))*self.reward
        rew -= np.dot(self.cost_change,  np.abs(p-a))
        rew -= np.dot(self.cost_unit, a + self.min_prod)
        if np.sum(a+self.min_prod) < self.d(temp) :
            rew -= self.C

        rew_players: list[NumExpr] = [rew] * self.maxPlayer

        if random.random() >= 0.9:
            actions = tuple([
                random.randint(0, self.max_prod[i] - self.min_prod[i]+1)
                for i in range(self.maxPlayer)
            ])
```

```
        return ((new_temp, actions), rew_players)

    def rewards(self, state: S, actions: list[int]) -> tuple[S, list[NumExpr]]:
        return self.rewardstuple(state, tuple(actions))

    def max_action(self, state: S, player: int) -> int:
        return self.max_prod[player] - self.min_prod[player] + 1

    def state_count(self) -> int:
        return (self.max_temp + 1) * self.max_action_total ** self.maxPlayer
    ...
```

A.2 Prisoner Dilemma

```
class PrisonerDilemma(StaticGame):
    def __init__(self):
        super().__init__(2)

    def max_action(self, state: None, player: int) -> int:
        return 2

    def rewards(self, state: None, actions: List[int]) -> Tuple[None, List[NumExpr]]:
        r: List[NumExpr] = []
        if actions == [0, 0]:
            r = [3, 3]
        elif actions == [0, 1]:
            r = [0, 4]
        elif actions == [1, 0]:
            r = [4, 0]
        elif actions == [1, 1]:
            r = [1, 1]
        else:
            raise NotImplementedError
        return None, r
```

A.3 Smoothed Bertrand Competition

```
class Cournot(game.StaticGame):
    def __init__(self, n:int, C, A, mu:float, possible_prices, k:int=1):
        """
        n : nb of players
        C : array of costs
        A : array of product quality (int), size of n+1 (a0), param for demand
        mu : param for demand
        m : number of possible actions (discretisation of prices)
        k : size of memory
        possible_prices : int (action) -> price
        """
        super().__init__(n)

        self.nb_players = n
        self.C = C
        self.A = A
```

```
        self.mu = mu
        self.m = possible_prices.shape[0]
        self.k = k
        self.possible_prices = possible_prices
        if self.nb_players == 2:
            self.compute_rewards()

    def compute_rewards(self):
        nactions = self.possible_prices.shape[0]
        self.pre_computed_rewards =
            np.zeros((nactions, nactions, 2))
        for a in range(self.possible_prices.shape[0]):
            for b in range(a, self.possible_prices.shape[0]):
                demand = self.get_demand(np.array([a,b]))
                reward = (self.possible_prices[[a, b]] - self.C) * demand
                self.pre_computed_rewards[a,b] = reward
                self.pre_computed_rewards[b,a] = reward[::-1]

    def max_action(self, state: None, player: int) -> int:
        return self.m

    def get_demand(self, actions):
        '''
        action : array of size nb_players
        '''
        ExpForDemand = np.exp((self.A[1:]-self.possible_prices[actions])/self.mu)
        SommeExp = np.sum(ExpForDemand) + np.exp(self.A[0]/self.mu)
        return ExpForDemand / SommeExp

    def rewards(self, state: None, actions: List[int]) -> Tuple[None, List[NumExpr]]:
        '''
        actions : of size nb_players
        returns reward an array
        '''
        if self.nb_players == 2:
            r = list(self.pre_computed_rewards[tuple(actions)])
            return None, r
        r: List[game.NumExpr] = []
        actions = np.array(actions)
        Demand = self.GetDemand(actions)
        reward = list((self.possible_prices[actions] - self.C) * Demand)
        return None, reward
```

RÉSUMÉ

Cette thèse est consacrée à l'étude de la dynamique de systèmes multi-agents dans lesquels les agents apprennent via des algorithmes. Formellement, il s'agit d'apprentissage en ligne dans les jeux stochastiques.

L'apprentissage en ligne est un champ des mathématiques et de l'informatique dans lequel on cherche à optimiser une fonction d'utilité ou de perte tout en interagissant avec l'environnement. À chaque étape, un agent choisit une action et observe ensuite ce qu'elle lui rapporte. Dans un jeu stochastique, les fonctions d'utilité des joueurs sont paramétrées par une variable d'état dont l'évolution peut être influencée par les joueurs.

Cette thèse étudie des procédures d'apprentissage, certaines originales et d'autres déjà connues, qui peuvent être utilisées par des agents qui interagissent dans un environnement modélisé par un jeu stochastique. Nous analysons les dynamiques résultant de ces systèmes, par exemple en prouvant que le comportement moyen des agents converge vers un équilibre.

MOTS CLÉS

théorie des jeux, apprentissage en ligne, jeux répétés, jeux stochastiques

ABSTRACT

This thesis is dedicated to the study of the dynamics of multiagent systems with learning agents. This is formalized as online learning in stochastic games.

Online learning is a field in mathematics and computer science that examines how to optimize a utility or loss function while interacting in an environment. It is typically supposed that interactions are a sequence of actions taken by an agent whose behavior is specified in algorithmic terms. In a stochastic game, utility functions are parameterized by a state variable evolution may be influenced by agents.

This thesis studies new and existing learning procedures which can be used by agents to interact in a changing environment modeled as a stochastic game, and analyses the resulting dynamics, for instance via the time average behavior of the players.

KEYWORDS

game theory, online learning, repeated games, stochastic games