Of Population-Based Methods for Multiagent Reinforcement Learning

Paul Muller

Of Population-Based Methods for Multiagent Reinforcement Learning

Paul Fernand Michel Müller

under the direction of

Karl Tuyls, Romuald Elie and Viet-Chi Tran

Publicly defended on the **18th of November, 2022**, with the following jury:

Luciano Campi	University of Milan	Jury President and Examiner
Amy Greenwald	Brown University	Rapporteur
Matt Taylor	University of Alberta	Rapporteur
Doina Precup	McGill University, DeepMind	Examiner
Tristan Cazenave	Université Paris-Dauphine	Examiner
Vianney Perchet	ENSAE Paris and Criteo	Examiner
Viet-Chi Tran	Université Gustave Eiffel	PhD Advisor
Romuald Elie	Université Gustave Eiffel, DeepMind	PhD Advisor
Karl Tuyls	DeepMind	PhD Advisor

A thesis presented for the degree of
Doctor of Philosophy

A joint collaboration between
DeepMind Ltd., Paris office
and
LAMA
Université Gustave Eiffel
France

Atop gilded hills
Perched high, black clouds incoming
A great change inbound

Threaded plain of life
Oh, Independent beings
The best stands afar

Flowing, old mystery
Huge crowd; unwise, unguided
Break the traditions

From ashes to dust
We merge all grain, sculpt it right
The mass is the sea

Learning to man sails
Mapping, charting ocean shells
We prepare to build

Set free from old age
Our legs hare, our wisdom proud
Blissful renewal.

Eyes riddled with sweat,
The wind of time all too fast
Sweet'ning damp regrets

Eyes softened by fate,
The head turns back, hailing sea
A sweet smile bearing

Body iron-made
Mind steel-clad, muscles leap-bound,
Jump to bright futures.

Contents

Remerciements

When three years of one's life come to a close, it is important to stop, breathe, and take a look back. Behind us stands a long line of stories, outcomes, utterances; conversations, actions, events.

Hierarchizing people is the last thing I want to do - however, when writing remerciements, such a hierarchy is naturally induced. I shall thus write straight from the heart, and hopefully strike true.

I cannot start this list without thanking my advisors. Karl, I will be forever grateful that you accepted to take me as your PhD student. Over these three years, you have been an academic father, and the best I could have hoped for. Thank you for having been there for my ups and downs, my questions and pings, and laughed at my jokes. Thank you for everything.

Romuald, thank you for having agreed to co-lead this PhD. Thank you as well for having been my second academic father - same, the best I could have hoped for. Thank you for having been there to correct my theorems, help my Math, and tame my worries. Thank you as well for everything.

Chi, thank you for having been there. I am sorry we didn't have the occasion to work more together - Covid certainly proved to be quite the obstacle. Thank you for having checked up on me, and for having always been there for me. Let's work together in the future !!

The next person I feel I should thank is Bilal. Finding a PhD has been quite an ordeal for me, my profile was uncommon enough that many were put off - I received 32 rejections for research roles in the span of a year. Deepmind itself never contacted me after I applied on our own website ! When, after 7 months of search, I finally received a positive response, I was advised by many to just take it. However, it felt too unrelated to what I actually wanted to research, and I had many a doubt about the thesis's organization. I decided I should seek advice, and contacted Bilal. Without this, I never could have joined Deepmind: Bilal confirmed my doubts, and transferred my resume to Remi Munos. This action, however small it may have been, started a chain of events which led to this day, and undoubtedly changed my life for the better. For this, I must thank you wholeheartedly, Bilal. I owe you a lot.

Julien also has a very special place in my PhD. He has acted as a mentor throughout these three years, always providing the wisest insights and comments into my situation, my research projects, and my goals as a researcher. Julien, I must thank you for your kindness, your patience, and your wisdom. Thank you very much for having been there for, and with, me. Tu as raison. ;)

I must also specially thank Karl, Romuald, Julien and Michael for having re-read and provided many thoughtful comments on my thesis. Thank you very much for having spent some of your valuable time on this conclusion of these beautiful three years.

One more special thanks I must address to Rana, who single-handedly decided to organize my PhD defense, and did so with brio. You are truly extraordinary, Rana, and I feel extremely thankful that our paths have crossed. Thank you enormously for having taken of your time for this young researcher.

These three years have been intense, but made merry by wonderful colleagues, some of whom had to leave the company due to personal constraints. Shayegan, Audrunas, I miss both of you, and hope we can find ways to work together again. :(Mark, Marc, Daniel, Zhe, Bart, Florian, Florent, Eugene, Miruna, Shantanu, Yunhao, Mina, Corentin, Michal, Sylvestre, Jean-Baptiste, Jean-Basten, Morgane, Mohammed, Rana, Laurel, Michael, Ramona, Thomas, Max, Jerry, Ngan, Remi, Ian, Yiran, and the whole London and Paris offices, thank you for having been there for

my PhD, however shortly. I have deeply appreciated our collaborations and discussions. I wish I could write something personal for each of you, but this thesis is already long enough, and there is so much to say ! Know that I have deeply enjoyed our conversations, and hope we can keep working together in the future. If by any mistake I have forgotten to include you, please know - I have deeply enjoyed every interaction I had at DeepMind, so I am deeply happy to have met and worked with you as well.

As a great friend of mine once said, "No man is an island by himself". Several great people, whom I have the deep happiness to call friends, have accompanied me throughout these three years. Alexandre, je te remercie d'avoir toujours été là pour moi. Ta gentillesse et ta patience, ton intelligence et ta sagesse sont pour moi des exemples, et le resteront toute ma vie durant. Wissam, your originality, intelligence and wisdom, motivation and kindness are also exemplary for me. Your muscles will never be as swole as your heart. Danny, thank you for being the uberbrain in my life. You are decidedly the most rigorous, exact, and kind-hearted person I know. Ahmed, please never change, always keep your light. Kevin, thank you for all our river conversations; I hope I can keep helping you grow, improve, and live a good life - you certainly have helped me do so. Miruna, thank you for your curiosity, drive and inspiring life ! Shantanu, you as well I must thank for your curiosity and energy. Let's get swole together, my brother ! Andrei, thank you for your cool and steel-sharp mind ! Vidisha, thank you for your constant optimism and happiness. Dragos and Oznur, thank you for your kindness, openness, and our adventures ! Laurent, parrain, thank you for your always instructive and fascinating conversations, your book recommendations, and your unending enthusiasm - though don't forget to let me talk too ! Laurel, thank you for being as annoying as I am - a too rare trait ! Alexis, je te remercie pour ton flegme, ta rigueur et ta "sassiness" ! Attention à toi, je serai bientôt plus musculairement volumineux que tu ne l'es ! Xiao, I hope you can keep maturing as a great researcher, our conversations have definitely helped me realize a lot about research and myself. Daniel, thank you for your patience and wisdom, for hearing my strange worries and quelling them with kindness, and for your endless reservoir of puns. Please keep them coming !! Ryan, my long-ago colleague, I hope you keep impressing me - and I keep impressing you, let's change the world ! Luc, merci pour ton humour, ton sarcasme, et ta culture quasi-infinie. Continuons à râler ensemble ! Corentin, merci pour ton intelligence, ton flegme et ta culture ! Ramona, thank you for your openness and sarcasm - let's get sassier ! Francois, thank you for all the burger/pizza-movie nights and our conversations; I hope you can find peace, you truly deserve it. Thanks for being a true companion. Banerjee, I know whatever I write you'll mock me. Thanks for laughing at me when I fall in the mud ;) Mina, I thank you for your humanizing ideas, discussions and character. I've felt less like a robot since we've gotten to know each other. ;P Louis-Guillaume, merci pour ton érudition et ta sagesse ! Puisses-tu continuer à m'inspirer autant ! Rana, merci pour ton energie si forte, ton enthousiasme constant, et ton sarcasme. You crack me up everytime ! Rapha' v2 est très content de t'avoir rencontrée !! Zhe, thank you for your optimism, kindness and energy. Let's do more things together !! Ganuche, merci pour ton infinie sagesse, ton calme et flegme admirales, et ton ouverture d'esprit. Tu es une des personnes les plus uniques que je connaisse, et je suis très heureux que ce soit le cas. Michael, we haven't known each other for too long, but already, I thank you for your widsom, kindness and openness. Yunhao, my man, let's watch more independent chinese movies together ! Gilles, mon ami, merci pour ton enthousiasme, ta gentillesse et ton énergie toujours au rendez-vous. To the all friends I haven't mentioned, I love you as well, and hope we meet soon. Please pardon me for not having mentioned you. For everyone - I hope we can spend more time together in the future. Our freedom is our only bound.

I feel I must also thank figures who have been truly influential in my life.

Monsieur Massias, je vous remercie humblement. Alors que j'étais en doute sur l'orientation de ma vie, vous avez pris du temps d'une soirée qui vous était pourtant consacrée, pour me prodiguer des conseils empreints de sagesse. Cela me toucha à l'époque, et me touche encore aujourd'hui - d'autant que vos conseils m'ont permis d'atteindre la position que j'ai aujourd'hui. Et c'est sans

mentionner vos cours d'une qualité extraordinaire. Merci pour tout, monsieur. Je suis heureux d'avoir été votre élève.

A Daniel, notre instructeur de tir a l'arc - je te remercie d'avoir été là pour nous, d'avoir toi aussi pris de ton temps pour nous instruire dans cette discipline que tu aimes tant, et, dans le même temps, de nous avoir inculqué de fondamentales leçons de vie. Merci pour ta sagesse, pour ton grand cœur, et pour ta présence. Nous espérons bientôt tirer avec toi à nouveau. Tu nous manques.

To Nader, I hesitated to write this in French - but we speak in English. Thank you for being a second father; thank you for your support in all things, your advice, and your views. I hope we can see each other soon - you're always welcome where I am.

To my family, thank you for the person I have become, and the trajectory you have set me on life. The smallest grain of sand may influence the biggest events in life, and in this you have been a most wonderful beach. Thank you for everything. The good and the bad, the Winters and Summers. The cold days by the chimney, the warm days by the pool. The chickens, the birds, the cats. The holidays, the trips, the adventures. The evenings together, the mornings asunder. The white mornings, the yellow skies; the red evenings, the blue dances. Thank you for all these, and for many more.

Finally, the most influential figure of all, Kehui. Thank you for having joined me for the second half of this adventure. Change is never easy, but good change is tremendously rewarding - and never have I been so rewarded. Thank you for freeing me from the walls I burrowed myself in, thank you for making me a better researcher, a better man, a better human. Let's keep growing together.

Résumé Etendu de la Thèse

Dans un monde toujours plus connecté, la capacité de prédire et maintenir un contrôle sur des systèmes toujours plus vastes, utilisés et vulnérables devient vitale. Par ce biais se pose à nous la question de la *stabilité*. Plus en détail, afin de pouvoir garantir le futur contrôle d'un système, il faut tout d'abord pouvoir prédire son évolution ; et comment prédire son évolution si ses dynamiques ne sont pas stables ? Si tous les utilisateurs *e.g.* de la grille énergétique changeaient soudainement de comportement d'une manière imprévisible, nous devrions nous attendre à de lourdes conséquences - présumablement soit une énorme surproduction, donc un gâchis de ressources ; soit une sous-production, donc soit un déficit commercial énergétique, soit des coupures généralisées. Pour contrôler un système, il est donc important de pouvoir connaître ses lois d'évolution dans le futur, c'est à dire, que ce système soit stable.

La Théorie des Jeux apporte une réponse à la question de la stabilisation. En effet, les équilibres de théorie des jeux caractérisent des situations où aucun joueur n'a intérêt à changer de comportement, *i.e.* où le système est stable, du moins en ce qui concerne le comportement de ses usagers. Il est donc important dans cette optique d'être capable de trouver des équilibres de théorie des jeux dans toutes les situations qui nous intéressent. Cependant, à part certains équilibres tels que les équilibres de Nash ou les équilibres corrélés, et ce uniquement dans certains cas, il n'existe pas de méthode générale permettant de calculer des équilibres dans des jeux.

Stabiliser un système contenant des millions, des milliards d'individus n'est de plus pas une tâche aisée, même étant donnée la Théorie des Jeux. En effet, la complexité de calcul des équilibres en théorie des jeux grandit en général exponentiellement avec le nombre de joueurs dans le système considéré. Fort heureusement, il est possible d'approximer ces systèmes, lorsqu'ils peuvent être exprimés sous la forme de jeux où l'identité de chaque acteur peut être ignorée, via l'approximation des Jeux à Champ Moyen. Cette approximation permet d'obtenir un équilibre typiquement $\mathcal{O}\left(\frac{1}{\sqrt{\mathrm{N}}}\right)$-optimal, où N est le nombre de joueurs dans le jeu. Quand $\mathrm{N} = 10^9$, l'erreur d'approximation devient en général minime. La littérature des jeux à champ moyen s'est de plus concentrée sur les équilibres de Nash, laissant de côté d'autres équilibres tels que les équilibres corrélés dont la définition même n'était pas, avant cette thèse, claire, alors que ce concept est prometteur dans l'établissement de gouvernance douce. Il n'existe pas non plus dans le domaine des Jeux à Champ Moyen de méthode générique de calcul d'équilibre.

Cette thèse s'articule donc autour de deux angles généraux:

- La conception de méthodes de calcul d'équilibres de théorie des jeux sur tout type de jeu fini, et
- La généralisation de certains équilibres aux jeux à champ moyen, et leur calcul.

A ces développements s'ajoute aussi une application pratique des équilibres de théorie des jeux dans une situation de football.

L'algorithme principal de cette thèse est nommé PSRO [98], Policy Space Response Oracle. Cet algorithme fut utilisé en tant que support d'inspiration pour la plupart des nouveaux algorithmes de cette thèse.

Introduction à PSRO

PSRO est un algorithme dédié au calcul d'équilibres de théorie des jeux en travaillant sur un sous-jeu du jeu en question (qui consiste en le choix d'une politique déterministe parmi un nombre limité - initialement, une seule, choisie aléatoirement - de toutes les politiques déterministes du jeu, et de la jouer jusqu'à la fin du jeu). La récompense totale obtenue par la politique dans le jeu est la récompense obtenue par le joueur du sous-jeu. A chaque itération de PSRO [98], une solution optimale de ce sous-jeu est calculée, cette solution optimale est ensuite transférée dans le jeu initial - dans lequel elle n'est présumablement plus une solution optimale -, puis une meilleure réponse contre cette solution est calculée dans le vrai jeu. Le sous-jeu est ensuite augmenté via l'addition

de la meilleure réponse contre la précédente solution, et une nouvelle solution optimale du sous-jeu est calculée, cela jusqu'à ce que la nouvelle politique calculée fasse déjà partie du sous-jeu, auquel cas une solution optimale a été trouvée.

Cet algorithme n'est théoriquement pas efficace: sa complexité dans le pire des cas est exponentielle en la taille des jeux en question. Cependant, le pire des cas étant assez pathologique, on constate empiriquement que PSRO est très efficace sur des jeux de taille petite ou moyenne, et la réduction de taille dûe à l'utilisation d'un sous-jeu permet de facilement calculer des solutions optimales.

Cet algorithme sert de base à la prochaine section, consacrée à la généralisation de PSRO à de nouveaux types d'équilibres.

Calculs d'Équilibres dans les Jeux à N-joueurs

PSRO était initialement un algorithme utilisé pour calculer l'équilibre de Nash d'un jeu à deux-joueurs et à somme nulle. Une contribution majeure de la thèse consiste en sa généralisation, d'abord à α-Rank [138], puis aux équilibres corrélés et faiblement corrélés [9], et enfin, la thèse généralise ce développement à tout type d'équilibre pouvant être décomposé suivant la décomposition SMD créée par la thèse.

α-Rank

α-Rank [138] est un nouveau concept de théorie des jeux créé afin de produire un nouveau système d'évaluation de stratégies d'un jeu qui serait robuste à certaines contraintes : multiplication de stratégies similaires, prise en compte de dynamiques non-transitives, complexité polynomiale. Cependant, il n'existe pas de méthode permettant le calcul d'α-Rank dans des jeux complexes sans les transformer en un jeu en forme normale de taille exponentielle en la taille du jeu complexe.

PSRO est prise comme un candidat prometteur pour permettre de calculer α-Rank dans des jeux complexes. La thèse commence par montrer, suivant le plan de Muller et al. [124], qu'il n'est pas possible d'utiliser PSRO pour calculer α-Rank, pas même en attendant que PSRO ait convergé vers un équilibre de Nash et en réutilisant son sous-jeu pour calculer α-Rank.

La thèse propose donc une nouvelle façon de calculer une meilleure réponse, suivant un autre objectif, qui vise à maximiser non pas la valeur obtenue, mais le nombre d'autres stratégies battues. Étant données quelques conditions, la thèse prouve qu'utiliser cette nouvelle meilleure réponse avec PSRO convergeait vers l'α-Rank du jeu en question.

Equilibres Corrélés

α-Rank est un concept essentiellement descriptif : il permet de calculer l'importance relative de certaines stratégies vis-à-vis d'autres dans un jeu. Cependant, il ne permet pas vraiment de coordonner des agents vers un objectif commun tout en garantissant que tous les agents soient contents de leur situation. C'est le cas des équilibres corrélés : ces équilibres induisent un médiateur, une figure extérieure au jeu qui est chargée de déléguer des recommandations de jeu aux différents joueurs. Ce médiateur choisit une politique globale, et distribue ensuite à chaque joueur le rôle qu'il doit jouer - sans l'informer de ce que les autres vont faire. Quand les joueurs préfèrent écouter le médiateur plutôt que de faire autre chose, le médiateur est un équilibre corrélé.

Similairement à α-Rank, il faut, pour converger vers des équilibres corrélés, changer de type de meilleure réponse, comme prouvé par Marris et al. [110]. D'une manière intéressante, cela n'est pas nécessaire pour les équilibres faiblement corrélés ! Une fois cette opération établie, la thèse prouve la convergence de PSRO utilisant des équilibres (faiblement) corrélés en tant que solution, et les nouvelles meilleures réponses, vers les équilibres des jeux en question.

Equilibres SMD

Prenant de la hauteur, la thèse examine ensuite s'il est possible de généraliser les méthodes ci-dessus afin d'appliquer PSRO à une large classe d'équilibre, créant le concept de décomposition SMD :

Sigma - Métrique - Déviation. Sigma est la fonction d'équilibre, calculant une distribution optimale. Métrique représente une mesure de la non-optimalité d'une stratégie donnée. Enfin, Déviation représente la fonction de Déviation de l'équilibre.

La thèse généralise alors à tout équilibre pouvant être exprimé sous forme de décomposition SMD l'algorithme de PSRO, menant à un algorithme générique : SMDRO. La boucle principale de cet algorithme est la suivante : (i) si la Métrique est nulle, stopper l'algorithme, sinon (ii) calculer la distribution optimale via S; enfin, (iii) calculer une nouvelle déviation contre cette distribution optimale via D. Cet algorithme converge vers l'équilibre SMD du jeu en question.

Équilibres Corrélés dans les Jeux à Champ Moyen

Ayant trouvé une méthode générale de calcul d'équilibres, la thèse cherche ensuite à l'étendre à des systèmes composés d'un grand nombre d'agents. Pour cela, l'approximation des jeux à Champ Moyen [102] semble idéale. Le problème est que le concept d'équilibre corrélé n'est que très jeune dans ce domaine, et sa manipulation n'était pas aisée [34]. De plus, il n'existait pas de borne connue concernant la qualité d'approximation d'un équilibre de Champ Moyen dans un jeu à N joueurs.

La thèse définit donc clairement et précisément ce que sont les équilibres corrélés et faiblement corrélés dans l'approximation des jeux à Champ Moyen, suivant ainsi le plan de Muller et al. [127]. Pour cela, elle s'intéresse d'abord aux jeux symétriques à N joueurs. Son constat principal est de remarquer que la fonction de récompense totale J en ce cas ne dépend que de la politique du joueur courant et de la distribution des politiques des autres joueurs - pas de quel joueur joue quelle politique. Cela veut dire qu'au lieu de considérer une politique jointe, il suffit de considérer la politique d'un seul joueur et la distribution des autres joueurs afin d'avoir des données de valeur. Ce concept passant clairement à la limite, nous en dérivâmes l'expression des équilibres (faiblement) corrélés à Champ Moyen.

La thèse dérive ensuite les propriétés fondamentales des équilibres (faiblement) corrélés, établissant par exemple des liens très forts entre les notions déjà existantes [34] et ses notions nouvellement introduites, et, entre beaucoup d'autres, donnant enfin une borne d'optimalité pour l'utilisation d'un équilibre de jeu à champ moyen dans des jeux à N joueurs : $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, lorsque l'équilibre est une somme finie de diracs. Cette borne est donnée exactement dans le cas où les transitions ne dépendent pas de la dynamique (et ne dépend que linéairement de l'horizon temporel du jeu considéré). Dans le cas où l'équilibre corrélé à Champ Moyen n'est pas une somme finie de diracs, par contre, cette borne tombe à $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$.

Calculs d'Équilibres dans les Jeux à Champ Moyen

Une fois les notions d'équilibres corrélés et faiblement-corrélés établis dans les jeux à champ moyen, la thèse aborde la question du calcul de cesdits équilibres. Deux approches différentes sont suggérées: la première définit et utilise la notion de *regret* à champ moyen, la relie aux équilibres corrélés et faiblement corrélés, et enfin montre que deux algorithmes populaires des jeux à champ moyen minimisent le regret externe, et convergent donc vers des équilibres faiblement corrélés.

La deuxième adapte PSRO au domaine des jeux à champ moyen, ce qui semble à première vue être un exercice facile - mais est en fait bien loin de l'être. Pour le voir, il convient de voir PSRO comme un algorithme qui, à chacune de ses itérations, calcule une table de résultats - une entrée par politique jointe et par joueur. Quand le nombre de joueurs est infini, cette table devient infiniment grande, et donc ne peut pas être manipulée ! Utiliser la même méthode de raisonnement que celle qui fut utilisée pour adapter les équilibres corrélés aux jeux à champ moyen ne marche pas non plus : en effet, si on ne considère plus que la table (Politique - Distribution), on ne considère plus qu'un objet de taille... Infinie. Le principe de l'algorithme décrit par la thèse réside sur une constatation : la table de résultats n'est nécessaire que pour calculer un équilibre ; mais il est possible de calculer des équilibres sans cette table, et c'est exactement ainsi que PSRO à champ moyen procède.

Regret à Champ-Moyen

Le regret à champ moyen est défini comme un analogue du regret dans les jeux à N joueurs : étant donnée une suite de politiques $(\pi_t)_t$, le regret externe est calculé comme étant la récompense maximale obtenue si, au lieu de jouer π_t, on jouait constamment une politique $\pi \in \Pi$, alors que la population jouait cependant toujours π_t.

La notion de regret interne est plus complexe à décrire. Toute politique stochastique est en fait une combinaison de différentes politiques déterministes. Le regret interne est défini comme étant la récompense maximale obtensible lorsque la masse probabilistique mise sur une politique déterministe π_a est mise sur une autre politique π_b pour chaque π_t, l'idée étant de se demander "à chaque fois que l'on m'a recommandé de jouer π_a, aurais-je eu intérêt à jouer π_b à la place ?", le maximum étant calculé sur le choix de π_a et π_b.

La thèse prouve ensuite que deux algorithmes populaires, Online Mirror Descent, et une modification légère de Fictitious Play, minimisent le regret externe à champ moyen, et convergent donc vers des équilibres faiblement corrélés.

PSRO à Champ-Moyen

Comme l'introduction de cette section le mentionne, il n'est pas évident d'adapter PSRO au cadre des jeux à champ moyen. La clef de cette adaptation repose sur l'idée qu'il n'est pas nécessaire d'utiliser une table de résultats pour calculer des équilibres ; certains algorithmes peuvent tout à fait marcher sans nécessiter de table de résultats, comme l'a montré Muller et al. [126] !

La thèse propose donc une adaptation de PSRO pour les jeux a champ moyen qui n'utilise pas, en général, de calcul de table de résultat. Pour être complête, elle exhibe cependant un classe de jeux à champ moyen, très restreinte (fonction de récompense affine en la distribution), pour laquelle il est possible et correct d'utiliser une table de résultats comme pour l'algorithme traditionnel de PSRO. Pour les autres jeux, deux solutions sont proposées, pour deux types d'équilibre.

1. Afin de calculer des équilibres de Nash, la thèse exhibe le problème d'optimisation recherché, et propose un système d'optimisation boîte-noire qui le résout en cherchant, itérativement, la distribution optimale de l'équilibre de Nash. L'algorithme suggéré par la thèse est un algorithme évolutionnaire, Adaptation de Matrice de Covariance - Stratégie d'évolution (Covariance Matrix Adaptation - Evolution Strategy, ou CMA-ES), mais il est certainement possible, étant donné plus d'informations sur la structure interne de nos jeux, de trouver un meilleur algorithme de recherche.

2. Afin de calculer des équilibres (faiblement) corrélés, la thèse utilise les propriétés des algorithmes d'optimisation adversariale : ces algorithmes sont capables de calculer une distribution sur un ensemble d'actions telle que, quelle que soit la suite de vecteurs de récompense choisie par un adversaire, potentiellement la pire, le regret moyen du choix de distribution tend vers 0. Or, la réaction de l'environnement au choix de distribution sur les politiques de PSRO (qui altère les fonctions de transition et de récompense via μ) peut exactement être considérée comme le choix d'une fonction de récompense par un adversaire. La thèse prouve donc qu'il est possible d'utiliser des algorithmes d'optimisation sans regret afin de calculer des équilibres (faiblement) corrélés. La thèse réalise cependant que ces distributions sont souvent à support discret, mais très large, ce qui rend les calculs subséquents très difficiles. Elle propose donc une méthode de compression de ces distributions via la résolution d'un programme d'optimisation linéaire, qu'elle nomme *compression de bandits* (bandit compression).

Équilibres dans les Penaltys au Football

Enfin, la thèse est conclue par une application de la théorie des jeux empirique (Empirical Game Theory, ou EGTA) à un type de situation réelle : les situations de pénalités / tir au but dans les rencontres de football, suivant les méthodes de Tuyls et al. [177]. Plus précisément, elle s'intéresse aux comportements concernant les choix des tireurs et gardiens de but : tirer à gauche, à droite,

au milieu ? Sauter à gauche, à droite, rester sur place ? Reprenant tout d'abord des résultats initialement connus, mais qui n'avaient pas nécessairement été statistiquement vérifiés, la thèse confirme que ces décisions de choix du côté de tir se révèlent être des décisions relatives, étant donnée le constat suivant : tout comme il existe des droitiers et des gauchers, certains joueurs tirent systématiquement avec leur pied droit, ou avec leur pied gauche. Cette préférence change tout à fait leur direction préférée, et leur probabilité de succès. Ainsi, un joueur tirant du pied droit préférera tirer sur la gauche du but, ou au milieu, et inversement pour un joueur tirant du pied gauche. Un ancien article résumait cette dynamique en introduisant la notion de Direction Naturelle, et émettait l'hypothèse qu'il était équivalent pour un gaucher de tirer à droite, et pour un droitier de tirer à gauche. La thèse confirme que cette approximation est statistiquement acceptable, tout en la nuançant : bien qu'elle soit statistiquement vérifiée lorsqu'on considère tous les joueurs à des niveaux d'expérience similaires, elle est statistiquement rejetée pour les joueurs à faible niveau d'expérience ! L'écart d'optimalité, *i.e.* la différence de valeur entre le comportement empirique des joueurs et leurs comportements estimés optimaux via le calcul d'un équilibre de Nash, des joueurs est ensuite estimé.

Enfin, la thèse raffine encore son analyse en introduisant la notion de représentation numérique du style de jeu des joueurs. Après avoir utilisé un algorithme de regroupement non-supervisé, K-means, sur les vecteurs de style de jeu, de nouvelles tables de résultats sont calculées, et d'intéressantes statistiques comportementales sont calculées entre différents groupes de joueurs distincts - différentes destinations de tir et fréquences de réussite, par exemple.

Perspectives

Comme l'introduction de ce résumé l'a souligné, les travaux de cette thèse se concentrèrent sur la question de la stabilité, à des fins portées sur des systèmes larges, y compris systèmes de gouvernance ; mais aussi systèmes multiagents en général - trafic, finance, coordination, ...

Les perspectives d'application de ces travaux peuvent par exemple avoir un impact fort dans le domaine de la gouvernance douce, c'est à dire l'établissement de système de coordination et de gouvernements dans lesquels l'obéissance n'a pas besoin d'être forcée pour être avantageuse : les acteurs obéissent parce que c'est pour eux la meilleure chose à faire. Plus généralement, la question de la stabilité et de la théorie des jeux peut être très importante pour le domaine du Mechanism Design, le domaine d'étude portant sur la recherche du meilleur mécanisme de motivation pour atteindre un optimum social donné.

Idem, des méthodes générales pour atteindre un équilibre donné peuvent s'avérer vitales pour les jeux à somme générale et à N joueurs, où le concept d'équilibre de Nash n'est plus nécessairement souhaitable, pour de nombreuses raisons - problèmes de sélection d'équilibre lorsqu'il en existe plusieurs, pauvre récompense moyenne, non-coordination des agents...

Enfin, le calcul d'équilibre peut aussi être employé afin de réussir à atteindre un haut niveau dans des jeux tels que le Poker [29], Starcraft [179] ou Stratego [148], jeux qui nécessitent un niveau important de capacité de prise de décisions dans des situations d'information partielle. On peut espérer que les méthodes de calcul d'équilibre seront utilisées pour continuer les avancées dans cette direction, ainsi que pour implémenter ce type d'avancées dans des systèmes de décision ou d'aide à la décision, donnant à des utilisateurs humains les meilleurs conseils possibles.

Nous espérons donc voir, dans le futur, de tels domaines d'application de la théorie des jeux, notamment du calcul d'équilibres, se développer ou continuer à se développer.

Contribution Acknowledgements

Most of my work has been done in collaboration with others, whom I once again warmly thank, and I feel it is appropriate to acknowledge their name and contributions to my papers. It has been my pleasure to spend these three years at Deepmind with all of you, and I hope these were only the beginning.

α-PSRO [124]: My work on α-PSRO was done in very close collaboration with Shayegan Omidshafiei, Mark Rowland and Julien Perolat; and I wish to wholeheartedly thank them, and all the other coauthors, for their much-needed help in writing this first paper of mine. Notably, the idea for the algorithm came up during discussions with Shayegan; and while I designed most of the theory of the paper and its experiments, Mark, Shayegan and Julien helped to derive extensive theoretical properties, analyze experiments and provided wonderful technical advice which significantly improved the quality of the paper.

JPSRO [111]: This work was started as the outcome of discussions between Luke Marris, Shayegan Omidshafiei and Karl Tuyls, who suggested to use PSRO as an application for a new correlated equilibrium concept that Luke had developed and for which he sought a use-case. I designed the new PSRO variant, which converged to correlated and coarse-correlated equilibria, and derived the proof of its convergence, while Luke developed maximum-Gini equilibria and wrote code for the experiments with support from Marc Lanctot.

Soccer [177]: This grandiose work was entirely coordinated by Karl Tuyls and Shayegan Omidshafiei. My role in it dealt with applying a game-theoretic analysis to penalty kick set pieces, thereby analyzing player behavior and issuing play recommendations in different situations, while Wang Zhe did the player-vector-derived analysis with advice from Karl and myself.

Correlated Equilibria in Mean-Field Games [127]: This work started at the beginning of 2021 and has been sent to ArXiv in August 2022. I was initially uninvolved in the project, and, as far as I am aware, Julien Perolat, Mathieu Lauriere and Romuald Elie are the ones who created the backbone for the first version of the article. I was initially added to the project to verify that derivations were correct, and, over time, slowly replaced almost the entirety of what had been done: deriving the new definitions of Mean-Field equilibria, providing their N-player intuition, finding their properties - equivalence with existing notions, optimality in N-player games, ... -, and extended existing proofs to a more general settings. Julien and Sarah Perrin worked on, respectively, Online Mirror Descent and Joint Fictitious Play's proof of no-regretness, which I reworked and generalized; Mark Rowland provided extremely insightful examples which deepened our understanding of existence relationships between different equilibria (When Coarse Correlated Equilibria exist, do Correlated Equilibria always exist ? They don't ! When Correlated Equilibria exist, do Nash always exist ? Neither !), and did a lot of work on definitions and properties of Mean-Field regret. Mark, Mathieu Lauriere, Georgios Piliouras and Matthieu Geist provided extremely pertinent comments, which have truly improved the quality of the paper. Finally, Romuald Elie had a tremendously important role in the paper's writing, flow, theorems and ideas, and I cannot thank him enough for having accompanied me there !

Mean-Field PSRO [126]: The idea for the algorithm came after discussions with Georgios Piliouras, and a question of his while we were writing the above paper: *"what good is it to talk about Mean-Field correlated equilibria if we cannot find them ?"*. This extremely pertinent question prompted me to search for a simple algorithm which would find correlated equilibria in Mean-Field games, and, given our then-recent paper with Luke Marris [111], PSRO felt like a great candidate. I however ended up stuck because of Mean-Field games' non-linearity issue, and Mark Rowland

provided the key which unlocked the paper: regret minimization. Without him and his insight, the paper would not have existed, and I wholeheartedly thank him once again for his advice, patience and kindness. Thanks to his idea of using no-regret learners, the non-linearity issue that I had identified was solved, and I could prove how to reach Mean-Field correlated and coarse-correlated equilibria. However, using no-regret learners led to very "wide" equilibria, *i.e.* equilibria which recommend many ν, which made computing best-responses very complex. I therefore introduced a new way to compress these equilibria into new, much sparser and at-least as well-performing (And typically much better performing) equilibria, and proved this behavior; and provided complexity analyses.

Table of Abbreviations

We introduce here in Table 1 common abbreviations used within this thesis, and their intuitive meaning.

Abbreviation	Meaning
RL	Reinforcement Learning, a field interested in maximizing value in a given environment.
BR	Best Response, a value-maximizer.
PSRO	Policy Space Response Oracle, a multiagent reinforcement learning method used to compute equilibria.
CE	Correlated Equilibrium, a type of game-theoretic equilibrium with rather strict constraints.
CCE	Coarse Correlated Equilibrium, a less-constrained variant of a Correlated Equilibrium.
PRD	Projected Replicator Dynamics, a type of perturbed Nash equilibrium solver.
MDP	Markov Decision Process, a type of environment where the dynamics and rewards an agent is subjected to only depend on the agent's current state.
CFR	CounterFactual Regret minimization, a popular algorithm to minimize a wide array of regrets, a game-theoretic measure.
OMD	Online Mirror Descent, a popular optimization algorithm.
FP	Fictitious Play, a popular multiagent reinforcement algorithm.
JFP	Joint Fictitious Play, an alternative definition of Fictitious Play in Mean-Field games.
MF	Mean-Field, an approximation of games with a very large number of players, which considers that these names have an infinity of players and only their spatial distribution matters.
MFG	Mean-Field Game, a game following the Mean-Field hypothesis.
MFCE	Mean-Field Correlated Equilibrium, a variant of CEs in the Mean-Field case.
MFCCE	Mean-Field Coarse Correlated Equilibrium, a variant of CCEs in the Mean-Field case.
MFNE or MFE	Mean-Field Nash equilibrium, a variant of Nash equilibria in the Mean-Field case.
JPSRO	Joint PSRO, a version of PSRO which samples joint actions instead of marginalized, that is, where players may play in a correlated fashion.
MF-PSRO	Mean-Field PSRO, an adaptation of PSRO to the Mean-Field case.
α-PSRO	PSRO adapted to converge to α-Rank's optimal strategic cycles.
PBR	Preference-Based Response, a new type of best-response developed for $\|alpha$-PSRO.
SMD	Sigma - Measure - Deviation, a decomposition framework encompassing several popular game-theoretic equilibria.
SMDRO	SMD-Decomposition PSRO; the general form of PSRO, capable of converging to all equilibria that can be SMD-decomposed.
KL	Kullback-Leibler divergence, a type of divergence often used in information theory and reinforcement learning.

Table 1: Common abbreviations used in this thesis, and their meaning.

Table of Notations

We introduce here in Table 2 common notations used within this thesis, and their intuitive meaning.

Notation	Meaning
π	A policy.
π_i	The policy played by player i.
π_{-i}	The policies played by all players other than i.
Π	The set of policies.
Π_i	The set of policies available to player i.
Π_{-i}	The set of policies available to all players other than i.
σ	A distribution over Π.
$\sigma(\pi)$	The probability of playing π according to σ.
$\mathcal{S}$	The set of states.
$\mathcal{A}$	The set of actions.
r	A reward function.
p	A dynamics function.
J	A payoff function.
V	A value function.
Q	A Q-value function.
$\Delta(\mathcal{X})$	The set of distributions over a finite set $\mathcal{X}$.
$\Delta_N(\mathcal{X})$	The set of possible distributions over a finite set $\mathcal{X}$ that N different players may form.
$P(\mathcal{X})$	The set of distributions over an infinite set $\mathcal{X}$.
$\mathbb{P}(A)$	The probability of event A happening.
ρ	A correlation device, *i.e.* a member of $\mathcal{P}(\Delta(\Pi))$.
μ	The spatio-temporal distribution of an infinite population.
μ^π	The spatio-temporal distribution of an infinite population playing π.
ν	A distribution over policies.
$\mu(\nu)$	The spatio-temporal distribution of an infinite population whose policies are distributed according to ν.
$\langle x, y \rangle$	The dot product between vectors x and y.

Table 2: Common notations used in this thesis, and their meaning.

Chapter 1

Introduction

Our world's connections and transnational links are growing each and every day. Despite political and cultural differences, conflicts - open or hidden - and a rising propensity towards isolationism, our world is increasingly appearing as an interconnected web of billions of individuals, all contributing to a global outcome. The nature of this system, its constantly increasing complexity and chaotic behavior make it almost impossible to analyze using traditional means, let alone to regulate and act on. This is where Artificial Intelligence may play a role.

With rising populations and connections, old questions have been brought to the forefront of political problems under a new light. The question of the Best Society has always been an important one, and ancient Greeks were already looking for an answer to it, 2500 years ago [7, 154] - without, unfortunately, unquestionable success. The Best Society is typically surmised to be a democracy, and defended as such by Leo Strauss [171], among many others [8, 50, 183] - though this view was not historically shared by all [33, 153].

However, other types of old questions, of a much less attractive nature than that of the Best Society, have also come back into play. Most of these questions circle around the capture and retention of Power, and the best representative of those who contemplated them is none other than Niccolo dei Macchiavelli [53]: Machiavel.

Whichever type of question we choose to ask regarding the form of society, every grand author agrees that societal stability is a necessary property of all political systems[1]; the clearest example of emphasis of stability being the doctrines developed by Chinese philosophers such as Kongzi, Mengzi or Mozi during the Spring and Autumn period and the Warring States period [43], a time of great upheaval and constant wars in ancient China.

The questions of *stably* organizing very large systems of agents, *i.e.* beings endowed with the capacity to choose their own actions, is not, however, unique to human-composed systems, but it is also present in human-*composing* systems, among many others. Indeed, our bodies (and all pluricellular organisms') are composed of an immense number of different cells which have *"agreed" to cooperate*, *i.e.* to leave out current rewards in favour of future ones that their cooperation allows them to obtain. Collective self-organization may also be found in the very large living systems that bees or ants create, with results of impressive complexity.

The "new light" mentioned above, under which old questions are being re-examined, has not yet been explicitized, yet it is apparent that it is the sheer interconnectedness of our world, and the consequences of local decisions on the whole. Whereas it mattered little to ancient Gergovians whether ancient Athenians decided to tax more or less their agricultural products (though it did matter!), such decisions now have high impact on global markets, local production and imports, which in turn affect the country's economic well-being, creating a feedback loop with consequences approaching that of the butterfly's tornado - non-negligible impacts on societal stability.

[1] Even Orwell's 1984 [142] can be interpreted as viewing stability as necessary: despite the inherent instabilities of 1984's society (Constant shortages and war), these are *mastered* to allow for total, and *legitimate*, societal control: *"during hard times, we must bond together as one!"*

How, then, can we ensure that our multi-agent systems of interest are *stable*? One avenue of answer rests upon *Game Theoretic Equilibria*.

In this work, we generalize the above settings - political, bio-cellular, ... - to any *game* where different *agents* interact with one another. By agent, we mean any entity which holds control over its actions - an ant, a cell, a human being are all agents, given the right scope. We focus on *games* because many, if not all, common interactions between agents can be modeled as a game - investing, driving, and even working - with varying degrees of fun. This abstraction allows us to generalize our analysis to other multiagent systems, such as financial systems, multiplayer games, sports, epidemiology, traffic routing, etc. Being able to find equilibria in one of those typically means being able to find equilibria in all of them, allowing our work to be widely applicable and, hopefully, useful for society in future works.

1.1 Motivation

This thesis looks at the question of computing game-theoretic equilibria using population-based methods.

Why are we interested in computing game theoretic equilibria? Let us start by providing an intuition as to what they are. Game theoretic equilibria represent, in a game, *stable* situations for the agents who are currently acting. This typically means that the agents do not have an incentive to change their behaviour, *i.e.* there is no way for them to *deviate* such that they find themselves in a more advantageous situation. The concept of *deviation* is central to game-theoretic equilibria. A deviation is a method that agents use to find a behavior that improves their situation. For example, in one type of equilibrium, agents are being recommended a strategy to follow, but can only decide not to follow it, *i.e.* to deviate, *before* having learnt of their recommended strategy. In another type of equilibrium, they may decide to deviate *after* having learnt what strategy they were recommended to follow. We can imagine equilibria where any pair of agents may decide to jointly change their actions to both be in a better situation - while perhaps making things worse for others -, and any other type of deviation-insensitivity defines a type of game-theoretic equilibrium: given a deviation, any situation of play where agents do not have an incentive to deviate, *i.e.* the situation is *insensitive* to the deviation considered, is an equilibrium.

We make the point that game-theoretic equilibria are the appropriate concept to model and optimize *stable* multiagent systems, hence our interest in their computation, since, at equilibrium, agents have *no incentive to deviate*. We may wonder why it is important for us to find stable systems. We must first note that stability is not always a desired property - in some situations, it is interesting to keep a system unstable and take advantage of its instabilities, when those are predictable [107]. Many situations do however require stability, such as traffic routing [23], mechanism design [17, 131], or min-max optimization [148]. Intuitively, this may be understood by instabilities being typically detrimental to (1) the predictability of a system's behavior - which is an important property when trying to improve said system ! -, and (2) the conservation of optimal properties of said system: intuitively, an unstable system will rarely - though that could happen depending on the game of interest - increase the welfare of its constituents; but more often be detrimental to it, at least in the games and situations we are typically interested in. A simple example of this would be a human pyramid: if any member of the pyramid decides to *"deviate"* and leave the figure, then the whole thing collapses. The cooperative games we study will typically have such a flavour.

We decide to use population-based methods to find game-theoretic equilibria. This is because population-based methods model a population of agents, and make them learn by making them interact with one another. By modulating the agents with which a given agent interacts and to which it adapts, population-based methods are capable of finding Nash equilibria in very complex games. Intuitively, they attempt to model the actual behavior of members of the system, hence their appropriateness in treating our general problem. We know that they are able to find Nash equilibria in complex games, and their intrinsic simulation of system-constituent behavior makes

them a prime candidate to generalize to computing any type of equilibrium: if we manage to adapt the method's modulation to fit new deviation types, it makes intuitive sense that the population would manage to learn the deviation's equilibrium.

1.2 Related Work

In this section, we broadly introduce a few game theoretical equilibria, introduce well-known equilibrium-computation methods, show their scaling limitations, and introduce Mean-Field games as a potential solution to the scaling problem, given their identity as an approximation of N-player games which allows one to scale to games with a very large number of players.

The question of computing equilibria in games can be said to be rather young, as its first appearance is only 184 years old, and its extensive study has arguably only started when Nash introduced the famous Nash equilibrium [133]. It has now been studied for more than 70 years, while its precursor, Cournot equilibrium, was formalized in 1838 [45]. The centralized version of Nash equilibria, correlated equilibria, introduced by Aumann [10], has only been studied for about 50 years. Its interest lies in the fact that they coordinate agents' behaviors, allowing for potentially more complex and subtle behavior. In this arguably very short time, game-theorists have developed many algorithms to reach either Nash or correlated equilibria in simple games - games with two players and a net sum of rewards per player equal to zero - via exact solvers. Computing equilibria in more complex games, general-sum N-player games, typically requires abandoning the safety of exact solvers, and relying on *iterative* approaches.

Correlated equilibria in normal-form games - very simple games with only one state and which can be written as a matrix - can be approached using an iterative process called *no-regret learning*. No-regret learning aims to produce a sequence of actions characterized by the fact that one can't make assertions such as "If every time I had done action A in this sequence, I had done action B, I would have been better off !". We can compute such no-regret sequences in rather efficient ways [21] in normal-form games.

Iterative approaches have been developed to find Nash equilibria in 2-player 0-sum games in theory, and in practice, it has helped find equilibria in complex team games where $2N$ players are distributed among two teams. They have been met with impressive success, and a partial taxonomy of this domain can be seen as composed of three clusters:

- Methods that simulate a population of agents which learn to adapt to one another. This type of method was used to learn extremely strong AIs in Capture the Flag [88] and Starcraft [181]. Such methods are typically based on classical methods such as Fictitious Play [163] or Double Oracle [117], which are proven to converge to Nash equilibria in 2-player 0-sum games. They have typically been used to find Nash equilibria.

- Methods that are based on no-regret optimization. Many of these methods are based on a classical algorithm called CFR, **C**ounter**F**actual **R**egret minimization. As its name suggests, this method attempts, at every decision point, to minimize its "regret", a measure of how much an action would have presumably benefited it, compared to the payoff it actually received. By increasing the likelihood of playing "good" actions, and decreasing that of playing "bad" actions, the algorithm is proven to converge to Nash equilibria in 2-player, 0-sum games, and to a complex type of equilibrium[2] in N-player general-sum games when used in self-play. Note that Morrill et al. [122] determined that CFR may be altered so that it converges to correlated equilibria. These methods however involve using a model as large as the game, and attempts at scaling said model have typically been met with mitigated results, with Poker being a notable exception [30, 119].

[2]CounterFactual Coarse Correlated Equilibria (CFCCE); for more information regarding this equilibrium, its relationship with CFR, and how to alter CFR so it converges to other types of equilibria than CFCCEs, we invite the reader to consult Morrill et al. [120].

- Methods that use regularization to modify the game in such a way that their new Nash equilibria are very easy to reach. By then slowly lowering this regularization to zero and starting from the previous equilibrium every time it is lowered, it is possible to reach the true Nash equilibrium of the game. This type of method has been recently used to approximate the Nash of the tremendously complex 2-player 0-sum game of Stratego [148]. Despite being known since 2001 [84], these methods have only now started to become popular due to their innate instability, which was recently solved by Perolat et al. [146] by changing the type of regularization used, which stabilized the process without removing convergence properties. However, their behavior is unclear in general-sum N-player games, and it is also unclear how such approaches may be adapted to lead to other types of equilibria.

Overall and in summary, while methods exist to compute some equilibria in certain settings, these are either restrictive (restricted to 2-player 0-sum games, normal-form games, or unalterable types of equilibria), or computationally unscalable.

One potential, underexplored way to overcome the scalability issue in the number of players is by approximating away the combinatorial complexities that arise in games with a very large number of players. If these games are symmetric, then agents' identities do not matter; only their state distribution does. We can approximate their state distribution by making the approximation that there is an infinity of players, and only consider their distribution. This frees us from the need to consider combinatorial issues due to sampling noise. This is the central idea of Mean-Field games [102]. Of course, this approximation can only be interesting for what it may bring back to the finite-player game currently under investigation - if no insight derived from the Mean-Field case were to transfer to the N-player case, then such approximations would be of little use in our context. Thankfully, this is not the case: under reasonable continuity conditions, a Mean-Field Nash equilibrium is an $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-approximate Nash equilibrium of its corresponding N-player game. This means that in games with a very high number of players, Mean-Field-approximation-derived equilibria are almost optimal. It is however not yet clear whether this is also the case for other equilibria. Furthermore, methods for finding Mean-Field Nash equilibria exist, but none exist for Mean-Field correlated and coarse-correlated equilibria.

1.3 Research Statement and Research Questions

From our motivation and given the current state of the art, a clear problem emerges:

Research Statement: *Finding game-theoretic equilibria is an important part in organizing real-life situations dealing with multiagent systems. Given an equilibrium concept, we are limited by our ability to find the chosen equilibrium, especially in very large games, by (1) a lack of generic methods which work in all games, and (2) a difficulty to scale to very large games.*

Indeed, problem (1) may be justified as follows: many different types of equilibria exist - Nash equilibria [134], correlated equilibria and coarse-correlated equilibria [10], quantal response equilibria [116], α-Rank [138] - a new equilibrium concept recently introduced to overcome some of Nash equilibria's limitations in N-player games -, Pareto-equilibria[3]... -, all of which are best suited for different use-cases. However, there does not exist a generic method able to converge to all of these equilibria, and for some of these, such as α-Rank, it is even unknown whether the equilibrium may even be reached outside of normal-form games.

Problem (2) stems from the fact that we aim to work on real-world systems, which can be characterized by their sheer size, at least in term of number of agents: were we to find a general method which converges to generic equilibria, we would like it to be able to scale to large systems.

We refine our desiderata into several questions.

[3]We define a Pareto equilibrium as any distribution over non-Pareto-dominated joint strategies.

Research Questions:

1. Given a game theoretic equilibrium concept, how does one reach it in *any finite game*[4]? *Chapter 3*

2. What are the Mean-Field equivalents of N-player equilibria? How can we use them to approximate N-player equilibria when N is very large? *Chapter 4*

3. How can we compute equilibria in Mean-Field games? *Chapter 5*

4. How can we apply game-theoretic equilibria to optimize real-world scenarios? *Chapter 6*

Question (1) deals with the question of computing general equilibria in all games. Given the lack of prior work on the question of finding a given equilibrium concept in any finite game, we chose to focus on this question only, leaving the question of finding an equilibrium *which maximizes a given metric* to future work, and on the question of scale.

Question (2) sets up question (3) by addressing the question of what equilibria become when passed to the Mean-Field limit - *i.e.* when considering that there is an infinity of players -, while question (3) asks how to use Mean-Field games to scale-up equilibrium computations. Indeed, the Mean-Field simplification allows one to scale to games with a very high-number of players, potentially greatly simplifying equilibrium computations. Finally, question (4) addresses the usability of these methods on real-world problems. Figure 1.1 illustrates the inclusion relationships between considered equilibrium sets.

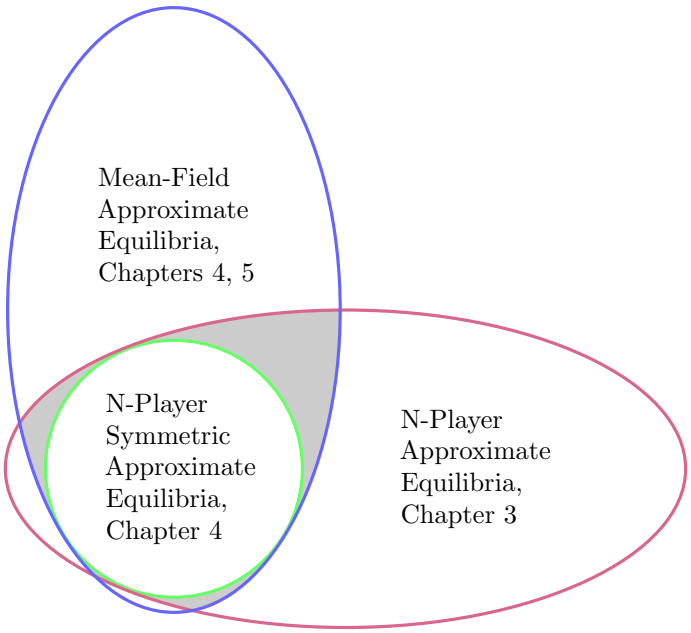


Figure 1.1: Visualization of the typical inclusion relationships between equilibrium sets, and plan of the thesis's contributions. It is yet unclear whether the grey area only represents the equilibria of asymmetric games which are asymptotically symmetric, or if other equilibria are included in this region.

[4]Any game with a finite number of states and actions.

Each research question is answered in a given chapter, indicated in italics at the end of its line. We provide explicit answers to each of these questions in Chapter 7, summarising the limitations and contributions of this thesis.

1.4 Plan of the Dissertation and Contributions

Chapter 2 provides the necessary background concepts that the following chapters build on - provided the reader already has a good understanding of linear algebra, topology, and real analysis.

Most of this dissertation is focused on adapting the Policy Space Response Oracle (PSRO) [98] algorithm to converge to different equilibria. PSRO was initially created to learn Nash equilibria in 2-player 0-sum games, but this work shows that, with few alterations, it can be used to converge to any chosen equilibrium. The issue with the algorithm is that, to converge, it potentially requires as many steps as there are deterministic policies in the game. This is unfortunately non-scalable; however, many games' equilibria do not require mixing over all policies, and, we have empirically noticed that a reasonable number of steps of the algorithm was typically enough to reach acceptable equilibrium approximations in reasonably-sized games. The algorithm can also be combined with deep learning out of the box, potentially making it the most interesting algorithm for our task of generalizing equilibrium computation.

Chapter 3 provides an answer to question (1) via two published papers, and a new development unique to this thesis. The first paper adapts PSRO to converge to α-Rank equilibria; the second, to correlated and coarse-correlated equilibria. The thesis-specific development generalizes these methods to *all* game-theoretic equilibria of a certain form, which includes Nash, α-Rank and correlated equilibria, among others. In more details:

[ICLR 2020] A Generalized Training Approach for Multiagent Learning [124] [Section 3.1]: This paper adapts PSRO to converge to α-Rank-optimal policies. It investigates thoroughly the relationship between α-Rank and Nash equilibria, proves via counterexamples that PSRO *must* be modified to reach α-Rank-optimal policies, and provides a modification of PSRO, α-PSRO, which, under suitable conditions, does converge to α-Rank-optimal policies. This convergence is illustrated in several games, demonstrating the empirical capabilities of the algorithm.

[ICML 2021] Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers [110] [Section 3.2]: My contribution to this paper regards how PSRO can be adapted to converge to correlated and coarse-correlated equilibria (a relaxation of correlated equilibria). In it, we provide modifications to PSRO, which, we prove, allow it to converge to either coarse-correlated equilibria or correlated equilibria. This theoretical result is completed by experiments demonstrating the algorithm's convergence on several games.

[This dissertation] Converging to General Equilibria in N-player, General-Sum Games [Section 3.3]: This development starts with a generic, abstract framework meant to represent as many game theoretic equilibria as possible, and several classical game theoretic concepts are shown to be expressible in our equilibrium framework. From this formalization, it describes how to adapt PSRO to converge to any such formalized equilibrium in a guaranteed way, providing a generalization to the two above papers, and a conclusive answer to research question (1).

Chapter 4 provides a partial answer to question (2) via part of a paper still under review, which defines what correlated and coarse-correlated equilibria exactly *are* in Mean-Field games, proves that they provide accurate approximations for their equivalents in N player games, and details how coarse-correlated equilibria may be learnt in all Mean-Field games.

[Under review, journal] Learning Correlated Equilibria in Mean-Field Games [127] [Chapter 4]: The part of this paper which addresses question (2) provides a new definition for Mean-Field correlated and coarse-correlated equilibria. This is justified in the paper through an exploration of the simplifications one can apply to symmetric games, and the realization that these simplifications are compatible with taking the number of agents to infinity. This concept defined, the paper directs an in-depth exploration of their properties, including the new definition's relationship with existing ones, notably its relationship with Nash equilibria, *e.g.* conditions under which one can extract Nash equilibria from correlated equilibria; but also equivalence with other existing formalisms of this equilibrium [34], existence conditions and relationships between equilibria - when coarse-correlated equilibrium exist, do correlated equilibria always exist? How about when correlated equilibria exist, do Nash equilibria always exist? We provide answers to these questions. We also describe how to use a Mean-Field (coarse-)correlated equilibrium in an N-player game, and derive optimality bounds for using the Mean-Field equilibrium in N-player games.

Chapter 5 addresses question (3) via first, the other part of the above paper, which link notions of Mean-Field regret with notions of Mean-Field (coarse-)correlated equilibria, and proves that two popular algorithms minimize said regret; and second, a published paper which adapts PSRO to converge towards correlated and coarse-correlated equilibria in Mean-Field games. In more details:

[Under review, journal] Learning Correlated Equilibria in Mean-Field Games [127] [Sections 5.1, 5.2 and part of 5.3.4]: The second part of this paper, which addresses question (3), introduces notions of Mean-Field external and swap regret. The minimization of these regret types is linked to respectively reaching Mean-Field coarse-correlated and correlated equilibria, as in the N-player setting. Interestingly, we find that no-regret learning algorithms converge to their respective equilibria *in value*, *i.e.* their deviation-incentive tends to zero; but *also* in distribution, *i.e.* their Wasserstein distance to the set of (coarse-)correlated equilibria also tends to zero. We also prove that Online Mirror Descent and an alteration of Fictitious Play, Joint Fictitious Play, minimize external regret, thereby converging towards coarse correlated equilibria. Their converged-to equilibria are analyzed, and shown to eliminate dominated strategies at a rate of $\mathcal{O}\left(\frac{1}{T}\right)$. Mean-Field PSRO is also qualitatively analyzed, shown to potentially never remove dominated strategies, and modifications that guarantee the absence of dominated strategies in the algorithm are proposed.

[AAMAS 2022] Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO [126] [Section 5.3]: This paper, historically an offshoot of the above paper (which was published before said above paper was even finished !), initially explored solely the question of converging towards correlated equilibria in Mean-Field games via adapting PSRO; but it was quickly generalized to an algorithm able to converge towards Mean-Field Nash, correlated and coarse-correlated equilibria. The difficulty in adapting PSRO to the Mean-Field case resides in the loss of linearity of the evaluation function with respect to strategies played, even in the case of a restricted game - a game where one can only play a subset of policies: whereas in N-player games, the expected value for playing a policy π_i when others play joint policy π_{-i} can always be expressed linearly, this property is lost in the Mean-Field limit. We go over this limit by using, for Nash equilibria, evolutionary strategies which search the combination of policies which minimizes restricted-game exploitability; and for coarse-correlated and correlated equilibria, no-external-regret and no-internal-regret learners to find a coarse-correlated or correlated equilibrium. These procedures yield very "large" equilibria, which require mixing many different recommendations, thus increasing computational load. We thus devised *bandit compression*, an algorithm which provably reduces all equilibria's complexity while providing no-worse (and empirically (much) better) approximations when applied to approximate equilibria. Convergence proofs, equilibrium existence and complexity analyses are provided.

Chapter 6 answers question (4) partially, by showing an example of real-world application of

Game Theoretic principles: Soccer. The published work in question deals with the question of what AI could bring to Soccer (and what Soccer could bring to AI):

[JAIR 2021] Game Plan: What AI can do for Football, and What Football can do for AI [177] [Chapter 6]: My role in this paper consisted in doing the game-theoretic analysis of players' behavior. Using average player behavior data in penalty set pieces, we tested whether players play equivalently when shooting left or right when they are left-footed or right-footed, whether they tended to act optimally (*i.e.* played according to a Nash equilibrium), and how we could potentially cluster their behavior, providing deeper insight into how best to advise and drill them.

Finally, Chapter 7 summarizes the thesis' contributions, notably its answers to each research question, and provides directions for future work and areas where our developed methods may be most useful.

Auxiliary contributions:

On top of the contributions mentioned above, I have had the astounding luck of spending the three years of my PhD at Deepmind, thanks to Karl and Romuald, who were kind enough to provide their wisening mentorship and dedicate part of their busy agendas to me, mentor me and help me become the scientist I am today. Working at DeepMind also provided me the chance to tangentially (and less tangentially) contribute to other works. These include, but are not restricted to:

OpenSpiel [99]: I implemented the current PSRO and Mean-Field PSRO implementations, several subalgorithms, and upgraded the α-Rank computation code. I have also coded new Mean-Field games.

MuJoCo Soccer [106]: I worked on PSRO-derived methods to train agents, reaching state of the art performance for 2v2 Boxheads.

Stratego [148]: I worked on a PSRO-derived approach to solve the game. However, FFoReL's performance outshone PSRO and this line of research was paused in favour of supporting FFoReL, leading to DeepNash.

Navigating the Landscape of Games[139] : I have been involved in the theoretical and technical discussions dealing with the paper, providing feedback and insights.

Scalable Deep Reinforcement Learning Algorithms for Mean Field Games [103]: I have implemented an environment for the paper and run quite a few experiments, including baseline runs, showing Deep-OMD and Deep Fictitious Play's performance compared to other approaches.

Multiagent off-screen behavior prediction in football [140]: I participated in discussions regarding the paper, provided feedback and insights.

Temporal Difference and Return Optimism in Cooperative Multi-Agent Reinforcement Learning [159]: I participated in discussions regarding the paper, provided feedback and insights, and empirical developments.

I have also spent time developing software for DeepMind through my work on Stratego, notably decentralized and distributed multiagent reinforcement learning systems; multiagent reinforcement learning losses and other tools, and new environments. I am tremendously thankful to everyone at DeepMind for the wonderful opportunities they provided me, and the great times we have had working on them !

Chapter 2

Atop Gilded Hills: Background

The scope of this chapter is to provide the reader with the background knowledge upon which the following chapters build. It will detail several concepts: Game Theory, which is a central theme of the thesis, Reinforcement Learning, Multiagent Reinforcement Learning, and Mean-Field games. All of the results here are known, and only their presentation is novel.

2.1 *Circling Adaptation*: Concepts of Game Theory

In this section, we will introduce the reader to the concepts of Game Theory that are most important to understand this dissertation. We will first start with a general introduction to the idea itself of Game Theory, providing an intuition to its contributions and application subjects; then, we will provide mathematical definitions of Game Theory's most important concepts for this dissertation.

2.1.1 What is Game Theory ?

Imagine you are playing a game of Rock-Paper-Scissors with a friend of yours, and that neither of you have ever played such a game before. Perhaps the first strategy your friend will choose will be to consistently play his favourite item - say, Rock. Since neither of you have understood the game at this point, it is even possible that he would win most games using such a tactic! But if, considering the game's dynamics, you consistently pick Rock's enemy, Paper, then he stands to lose consistently against you. Unless he decides to adapt and plays Scissors. In which case you should adapt to playing Rock. And perhaps now you start wondering what your opponent's next move will be given that he knows your move, an idea known as Theory of Mind [67].

However, why would your opponent not consider what you consider that he will consider that you consider when making a move? This sentence, complex to understand, points at a significant issue with adapting approaches: why would they ever end, why would they ever find a stable solution? There is always further to look, more complex models to find, better ways to adapt against a given opponent - but this also means that others will always be able to find better ways to adapt against you, potentially leading to an endless cycle of strategic changes.

Studying such dynamics is a part of Game Theory. Another part, very significant in this manuscript, is one which solves (through avoidance) the above problem : finding stable equilibria, that is, equilibria from which no-one has any incentive to deviate. In Rock-Paper-Scissors, for example, a stable equilibrium is to play Rock, Paper and Scissors in a uniform-random fashion - this is actually the game's Nash equilibrium. We note two things : Whatever the other player plays, we know we will not lose anything, and indeed, this is the game's min-max strategy. However, if your opponent were not to be max-mining and just kept playing Rock, we note that playing the Nash equilibrium would not be the most profitable strategy for you.

From this simple example, we can draw two conclusions which correctly generalize:

- Dynamic, adapting systems may circle endlessly between different strategies and never settle on any
- Static equilibria are great concepts to guarantee a minimal payoff, but they may not be the strategies that will maximize payoffs in all cases : they are typically low-as-possible-risk, high-as-possible-reward strategies (Yet the high-as-possible-rewards may be quite low, as in RPS)

In the following sections, we will first introduce general definitions, then study a few different static equilibria, and then see an intriguing correspondence between adapting systems and static equilibria - it turns out that the cycle produced by adaptive systems can be, when it is averaged and under suitable conditions, a static equilibrium!

2.1.2 General definitions

We provide here a series of definitions which will be useful for this section.

Given a countable set X, we note $\Delta(X)$ the set of distributions over X. $\mathcal{P}(X)$ represents the same set when X is uncountable.

We write J the expected payoff function, *i.e.* the expected gain when playing in a certain way, given a certain setting.

Normal-Form Games (NFGs): We consider here normal-form games, *i.e.* games where all players select one action (Among finite sets of available actions) at the same time, and receive a specific payoff determined by the actions every player took.

In the two-player case, this is akin to specifying two payoff matrices A and B of shape (n, m) - where n is the number of distinct actions available to player 1, and m, that of player 2 -, and the payoff received by player 1 for playing action i when the other player plays action j would be $e_i^t A e_j$, where e_i is the i-th base vector (And similarly, the payoff for player 2 in this case is $e_i^t B e_j$).

Policies (π): We say that players are playing policies, or strategies, when they choose to play a distribution over actions - their policy is their distribution over actions. We name $\bar{\Pi}$ the set of policies, and, for convenience, Π the set of deterministic policies, *i.e.* policies with all their mass concentrated on one action. In an N-player game, Π_i and $\bar{\Pi}_i$ are, respectively, the sets of deterministic policies, and of policies, available to player i. Finally, in an N-player game, we write Π^N and $\bar{\Pi}^N$ the sets of deterministic and stochastic joint policies.

x-sum Games: We say that a game is x-sum if for all joint action a, $\sum_{i=1}^{N} J_i(a) = x$, *i.e.* the sum of rewards of all players is always equal to x, whatever they choose to play. When there exists no such x, we say that a game is *general-sum*.

Value (J, V): The quantity that interests us is $J_i(\pi_i, \pi_{-i})$, which is the expected payoff for player i when it plays policy π_i while the other players play the joint policy π_{-i}. Note that π_{-i} is a joint policy, and can also be written $\pi_{-i} = (\pi_1, ..., \pi_{i-1}, \pi_{i+1}, ..., \pi_N)$. We say that other players play π_{-i} when each player that is not i plays its corresponding component in π_{-i}.

In the two-player case, $J_1(\pi_1, \pi_2) = \pi_1^t A \pi_2$ and $J_2(\pi_2, \pi_1) = \pi_1^t B \pi_2$.

Policy Deviation: We name $\mathcal{U}_{CE}$ the set of policy swaps: $\mathcal{U}_{CE} = \{u : \Pi \to \Pi\}$. $\mathcal{U}_{CCE}$ is the restriction of $\mathcal{U}_{CE}$ to constant functions. Finally, we define policy deviations as functions $f : \bar{\Pi} \to \bar{\Pi}$ such that there exists $u \in \mathcal{U}_{CE}$, $\forall \bar{\pi} \in \bar{\Pi}$, if we decompose $\bar{\pi}$ as a mixture of policies of Π $\bar{\pi} = \sum_\pi \alpha_\pi \pi$, then $f(\bar{\pi}) = \sum_\pi \alpha_\pi u(\pi)$. In the rest of this work, we will assimilate members of $\mathcal{U}_{CE}$ with their policy deviations, and thus will write $u[\pi]$ for members of $\bar{\Pi}$ as well as Π.

2.1.3 Nash Equilibrium

Definition 1 (Nash Equilibrium)**.** A joint-policy π is a Nash equilibrium if it is such that, for all i,

$$\forall \pi' \in \Pi_i,\ J_i(\pi', \pi_{-i}) - J_i(\pi_i, \pi_{-i}) \leq 0$$

To provide intuition: no player has an incentive to stop playing π_i and play another strategy instead. Since no player has an incentive to change its behavior, we can consider this situation to be stable. However, we note that this is a weak notion of stability, a "first-order" one: only per-player deviations are considered. However, if two or more players were to deviate at the same time in a concerted manner, playing according to their Nash may not be in the interest of other players anymore!

Given such a definition, several questions arise: **Q1**. In what type of games do such equilibria exist? **Q2**. Is it possible that several such equilibria exist in a single game?

To answer the first question, we cite a fundamental theorem for Game Theory: the Kakutani Fixed-Point Theorem [91].

Theorem 1 (Kakutani Fixed-Point Theorem)**.** *Take $\mathcal{X}$ a convex compact non-empty subset of $\mathbb{R}^n$ for some finite $n \in \mathbb{N}$, and a function $\phi : \mathcal{X} \to 2^{\mathcal{X}}$ from $\mathcal{X}$ to sets of $\mathcal{X}$. Assume the following properties:*

- *ϕ has a closed graph.*
- *$\phi(x)$ is non-empty, closed and convex for all $x \in \mathcal{X}$.*

Then ϕ admits a fixed-point in $\mathcal{X}$.

Let us see how it is used to prove the following theorem, the Nash theorem [133]. We provide here its proof for completeness:

Theorem 2 (Nash's Theorem)**.** *In any matrix-form game, there always exists a Nash equilibrium.*

Proof. Consider the best-response function ϕ such that

$$\forall \pi \in \Pi,\ \phi(\pi) = \left\{ (\pi'_1, ..., \pi'_N) \in \Pi \mid \forall i,\ \pi'_i \in \arg\max_{\pi'_i \in \Pi} J_i(\pi'_i, \pi_{-i}) \right\}$$

Obviously, $\phi(\pi)$ is never empty, given that Π is a convex compact set. Let $\pi_1, \pi_2 \in \phi(\pi)$. This means that $\forall i,\ J(\pi_{1,i}, \pi_{-i}) = J(\pi_{2,i}, \pi_{-i}) = \max_{\pi'_i \in \Pi_i} J(\pi'_i, \pi_{-i})$.

However, by virtue of J being a matrix multiplication, we have that, $\forall t \in [0, 1]$,

$$J(t\pi_{1,i} + (1-t)\pi_{2,i}, \pi_{-i}) = tJ(\pi_{1,i}, \pi_{-i}) + (1-t)J(\pi_{2,i}, \pi_{-i})$$

Which shows that $\phi(\pi)$ is convex. The closedness of $\phi(\pi)$ comes directly from the continuity of J.

We now prove that the graph of ϕ is closed.

$$graph(\phi) = \{(\pi, \pi') \mid \pi \in \Pi,\ \pi' \in \phi(\pi)\}$$

Let $((\pi_n, \pi'_n)))_n$ be a converging sequence of elements of $graph(\phi)$, and let $\bar{\pi}$ and $\bar{\pi}'$ be the limits of $(\pi_n)_n$ and $(\pi'_n)_n$ respectively.

We note that the function J is continuous with respect to π, since it is a tensor multiplication.

We know that $\bar{\pi}, \bar{\pi}' \in \Pi$ since Π is a convex compact subset of $\mathbb{R}^{\sum_i |\bar{\Pi}_i|}$, and that

$$\forall n,\ J_i(\pi'_{i,n}, \pi_{-i,n}) = \max_{\pi'_i \in \Pi} J_i(\pi'_i, \pi_{-i}),$$

and

$$\max_{\pi'_i \in \Pi} J_i(\pi'_i, \pi_{-i}) = \max_{\pi'_i \in \bar{\Pi}} J_i(\pi'_i, \pi_{-i})$$

since the max is always reached over deterministic policies.

This set is finite, and thus, since the max over a finite set of continuous functions of π is continuous with respect to π, we can take the limit over n and thus have

$$\forall i,\ J_i(\bar{\pi}'_i, \bar{\pi}_{-i}) = \max_{\pi'_i \in \Pi} J_i(\pi'_i, \bar{\pi}_{-i}).$$

Thus $\bar{\pi}' \in \phi(\bar{\pi})$, which means that $(\bar{\pi}, \bar{\pi}') \in graph(\phi)$, which is thus closed!

We can therefore apply the Kakutani Fixed-Point Theorem, and deduce that ϕ has a fixed point, *i.e.* $\exists \pi, \pi \in \phi(\pi)$.

To develop this property: this means that for all i,

$$\forall \pi'_i \in \Pi_i,\ J(\pi'_i, \pi_{-i}) \leq J(\pi_i, \pi_{-i}),$$

which means that π is a Nash equilibrium of the game. □

Remark 1 (Proof extension)**.** *Note that the above proof works for any game with finite states and actions, and continuous payoff function J.*

We can therefore answer **Q1**., Nash equilibria exist in all finite-action games with continuous payoff functions, which includes matrix games! **Q2**. can be answered easily as well. Imagine a game with A identical actions for all players. Then any mixture of these actions is a Nash equilibrium. Outside of this trivial example, let us consider the Diagonal Action game: it is a game with A actions and with diagonal payoff structure: Players get rewarded for all playing the same actions - or get no reward at all if any player chooses a different action from the rest.

Assuming that all diagonal rewards (The rewards for all playing the same actions) are equal, then the game has at least A Nash equilibria - each Nash consisting of players all focusing on one action. In the two-player case, the game has exactly A Nash equilibria, but it has an infinity thereof in an $N > 2$-player game (Consider for example that 2 players play actions that no other player plays. Then no unilateral deviation from any player can make them get > 0 reward. Since all deviations have equal value to their current position, then their current position is a Nash equilibrium; and there can be an infinity of mixtures that will be as well.)

2.1.4 Limitations of Nash Equilibria

Nash Equilibria are a fundamental concept of game theory, and brought about many vital results therein. If anything, they are a vital idea in two-player zero-sum games, where they *are* the optimal solution - provided the other player is considered as an opponent who could behave adversarially, and has unlimited learning possibilities.

However, their usefulness is counteracted in several cases, for different reasons: the general-sum case, the N-player case are cases where Nash equilibria lose some of their meaning and usefulness. Nash equilibria also suffer from an equilibrium selection problem, and a potentially high cost of anarchy.

The general-sum case: Consider the famous Prisoner's Dilemma game, whose payoff matrix is shown in Table 2.1. This matrix is read the following way: each line represents an action for the row player (Player 1), and each column represents an action for the column player (Player 2). Finally, each matrix entry is a tuple of two values. The first value is the payoff for player 1, the row player; and the second value is the payoff for player 2, the column player.

In the prisoner's dilemma, two prisoners are given, separately, a choice. Either choose to cooperate with their fellow prisoner and not denounce each other in the hopes of only escaping with a minor condemnation; or choose to defect, denounce the other prisoner in the hope of being pardoned in exchange for having given law enforcement useful information.

If both prisoners choose to cooperate (With each other!), then they both only get light sentences and get a small reward. If one prisoner chooses to defect while the other cooperates, then he gets a big reward - he walks free! - while the other gets a huge penalty - he is condemned while his

	Defect	Cooperate
Defect	0, 0	5,-2
Cooperate	-2, 5	3, 3

Table 2.1: Prisoner's Dilemma Payoff Matrix

"friend" is walking free. Finally, if both prisoners choose to defect, then they both are thrown in jail - but at least they *both* are, hence the higher payoff to being betrayed.

This game is famously known for its Nash: indeed, the Nash equilibrium of this game is to systematically defect. Nash Equilibria do not, in general, encourage cooperation in general-sum games, and especially not in social dilemmas.

Remark 2 (On the ill-conceivedness of the Prisoner's Dilemma Game)**.** *One could consider that, since Nash equilibria recommend a suboptimal situation (Both players end up defecting, and thus going to jail for a long time, instead of choosing to cooperate and walk with extremely light sentences), they may be the wrong concept for Game Theory; and Social Good in general.*

We would like to challenge this generalization - if it is to be established, it shouldn't be from this example. Indeed, we argue that the Prisoner's Dilemma, as presented here, is intrinsically unrealistic*. Indeed, in real life, we do not betray one another in part because there are consequences to betraying one another*[1]*! Hence, stopping the whole game, making life stop right after the betrayal has happened (Or not) is intrinsically unrealistic. The betrayer may well face heavy consequences for his action!*

We argue that the Iterated Prisoner's Dilemma[2] *is an intrinsically more realistic game - it consists of repeating the Prisoner's Dilemma game an unknown, finite number of times between the same two players. In this game, actions have consequences, and a cooperating player may well punish a defecting player by never choosing to cooperate again!*

Indeed, a "close-to-optimal policy" (Though such an object is difficult to define) in this setting is called Tit-for-Tat*, which consists of doing unto others what they do unto us: start by cooperating, then repeat the other player's last action. If the other player chose to defect, we defect. If he chose to cooperate, we cooperate. This algorithm, although it "loses" to a policy which consistently defects because it initially cooperates, does not lose much (Only loses on one round and never on the others), and is able to cooperate consistently with cooperators. This algorithm of stern altruism (I cooperate with you, unless you don't cooperate with me) is much closer to our vision of "acceptable", and, depending on which population it is confronted to, will be optimal (Given a population of cooperators) or, sometimes, suboptimal (Given a population of defectors). We note that such results have been observed in humans as well (Cooperators, when stuck with defectors, underperform; whereas they overperform when surrounded by other cooperators).*

The N-player case: In the N-player case, the problems with Nash equilibria are many. We choose to mention here the two main ones.

- The computational aspect: Computing a Nash Equilibrium is a PPAD-hard task, a hard class of complexity in general. However, as we will see in Section 2.1.7, there exist straightforward algorithms which minimize a quantity called regret, which, in 2-player 0-sum games, reduces to, in time-average, Nash Equilibria. This property is lost in N-player games, making it much more difficult to find Nash equilibria.

[1]Evolutionary Psychology also makes the argument that the other reason why we do not betray one another, because it "feels bad", is an evolved psychological reaction to encourage cooperation and discourage defection in human groups - in a word, if we feel bad about betraying someone, it may just be because it is bad for us to do so in the long run! On the topic of Evolutionary Psychology, I can only recommend Homo Fabulus. (In French, however, subtitles are available).

[2]On this topic, but also for a biological view of Game Theory, I strongly recommend Dr. Sapolsky's lectures.

- The concept of Nash equilibrium supposes that the other player does everything it can to minimize our payoff - it is a pessimistic, worst-case solution where our opponent does everything possible to beat us. In N-player games, this translates to optimizing for the case where the $N-1$ other players will gang up on us to beat us mercilessly without thinking about doing anything to one another - which one could argue is perhaps slightly too pessimistic. Another issue is also team-based games: optimizing for a Nash would also entail considering that one's teammates are the worst possible teammates while one's opponents are the best possible opponents - once again, perhaps Nash Equilibria aren't the best concepts in N-player games.

The Equilibrium Selection problem: If we are to take back the Diagonal Action game mentioned above, *i.e.* a game where players only receive rewards for playing strictly the same action, we notice that there exists several Nash equilibria in this game (Namely, at least as many as there are actions). However, if two players select two different actions, hence two different Nash equilibria, they will not get any reward, which is suboptimal!

Indeed, Nash equilibria require players to select the same equilibrium for their optimality guarantees to work. Otherwise, they may be as non-optimal as one can make them! However, by definition, Nash equilibria are "uncoordinated" equilibria, where players do *not* synchronize. This is the famous equilibrium selection problem.

The Price of Stability problem: Related to the above problem, the notion of Price of Stability is the measure of how suboptimal a system of N different, independent agents is, compared to the same system where agents would lose their free will and single-mindedly optimize welfare:

$$\text{PoS(Nash)} = \frac{\text{Maximum Welfare of Nash Equilibria}}{\text{Maximum Possible Welfare}}$$

Since agents do not synchronize at all, intuitively - but this has been confirmed as well, see [44] for an example -, their price of stability would be higher than free-willed agents which would be able to coordinate their actions.

We show in the next subsections other concepts of equilibrium which address the above issues. However, we would like to moderate the ideas brought about in this section: it is very easy to show the limitations of an equilibrium concept such as Nash, but it does not mean that it is useless, or "bad" in any way. A Nash equilibrium is what it is, nothing more and nothing less. It is more or less adapted to a given situation, and will give more or less desirable results. However, we are compelled to recognize how powerful Nash equilibria are, and how important they have been in solving zero-sum two-player games [25, 148, 165, 180].

2.1.5 α-Rank

α-Rank is a recently-proposed evolutionary-game-theory-inspired equilibrium concept by Omidshafiei et al. [138] whose strongest benefits are its uniqueness - hence no equilibrium selection problem - and efficient computation in many-player and general-sum games. It was primarily developed to overcome the shortcomings of Elo in ranking strategies by strength. It does so in several ways. First, it captures non-transitive (*i.e.* Rock-Paper-Scissor-type) dynamics, which the Elo rating is insensitive to. Second, it is insensitive to strategy repeats. This matters for the following reason: assume, in a group, that 10 people always play Rock, 1 person plays paper, another plays scissor. Due to the repeating of Rock, the paper player will have a very high Elo rating, despite the fact that all strategies are actually of equivalent strength. Finally, it is computationally less intensive than Nash equilibria, and, perhaps more importantly, it does not suffer from equilibrium selection problem and other issues due to non-uniqueness of Nash equilibria, as the α-Rank solution is *always unique*. For all these reasons, being able to compute the α-Rank solution of a game means being able to gain great insights about its intrinsic dynamics, and its best strategies. Part of this dissertation also wonders about the computational efficiency of using α-Rank to compute good strategies, instead of just understanding the game's intrinsic dynamics.

It has two main versions: the single-population case, adapted to 2-player, symmetric games, where a player may switch from any strategy to any other; and the multi-population case, adapted

to any finite game, where only per-player deviations are considered (There can be no coordinated, simultaneous deviation). We start with the single-population case, then define the multi-population case.

However, in general, the α-Rank distribution is the stationary distribution of a specific Markov chain. In the single-population case, it is over strategies in Π; in the multi-population case, it is over the space of joint strategies Π_N. Its transition probabilities are defined differently in both cases.

To compute this stationary distribution, it is necessary to compute a transition matrix between strategies (single-population model) / joint strategies (multi-population model). These transition matrices define a directed graph, which we call a *response graph*.

It is also parametrized by two variables: α, and m. The latter, m, can be interpreted as controlling the rate of elimination of suboptimal strategies - the higher it is, the lower the probability of switching to less-optimal strategies, while more-optimal strategies' switch probabilities remain relatively unchanged. α represents evolutionary pressure - the higher it is, the more more-optimal strategies' switch probabilities increase, while the more the less-optimal strategies' switch probabilities decrease - in the limit of $\alpha \to \infty$, transitions become deterministic and get positive probability if and only if they increase fitness. That this is the canonical use case of α-Rank: α is taken to infinity.

We note that in the single-population case, Omidshafiei et al. [138] shows that the m term is eliminated in the computations.

Single-population α-Rank

In the single-population case, the game is symmetric, and the α-Rank distribution is over individual policies of Π. It is the stationary distribution of a Markov chain over Π with transition probabilities defined as, for π_1, $\pi_2 \in \Pi$,

$$\mathbb{P}\left(\pi_1 \to \pi_2\right) = \eta \frac{e^{\alpha J_1(\pi_2, \pi_1)}}{e^{\alpha J_1(\pi_1, \pi_2)} + e^{\alpha J_1(\pi_2, \pi_1)}} \tag{2.1}$$

$$\mathbb{P}\left(\pi_1 \to \pi_1\right) = 1 - \sum_{\substack{\pi \in \Pi \\ \pi \neq \pi_1}} \mathbb{P}\left(\pi_1 \to \pi\right) \tag{2.2}$$

where $\eta = \frac{1}{|\Pi|-1}$. Of note is the first term, $\mathbb{P}\left(\pi_1 \to \pi_2\right)$, which bears strong resemblance to the Elo rating [60]. Indeed, the transition probabilities are defined as the probability that π_2 beats π_1, multiplied by the probability of an encounter happening, η. In the Elo ratings model, the former probability is defined as $\frac{Q_2}{Q_1+Q_2}$, where Q_i measures the absolute strength of player i. We see that α-Rank's transition probabilities follow the same intuition, except that it does not use absolute strengths, but relative ones. The idea behind these transition probabilities is clear: the more π_2 beats π_1, the more likely the transition from π_1 to π_2 is to happen.

Multi-population α-Rank

The α-Rank distribution is the stationary distribution of a specific Markov chain over joint strategies, which has transition probabilities at joint strategy π

$$\mathbb{P}\big(\pi \to (\pi_i', \pi_{-i})\big) = \begin{cases} \eta \frac{1 - e^{-\alpha\left(J_i(\pi_i', \pi_{-i}) - J_i(\pi)\right)}}{1 - e^{-\alpha m\left(J_i(\pi_i', \pi_{-i}) - J_i(\pi)\right)}} & \text{if } J_i(\pi_i', \pi_{-i}) \neq J_i(\pi) \\ \frac{\eta}{m} & \text{otherwise}, \end{cases} \tag{2.3}$$

$$\mathbb{P}\left(\pi \to \pi\right) = 1 - \sum_{\substack{i \in [N] \\ \pi_i' \in \Pi_i \setminus \{\pi_i\}}} \mathbb{P}\big(\pi \to (\pi_i', \pi_{-i})\big) \tag{2.4}$$

where $\eta = \frac{1}{\sum_i (|\Pi_i| - 1)}$. Namely, the Markov chain can only change one of its players' strategies at a time. The "softmaxness" of the probabilities ensure that the Markov chain is irreducible for all values of m and α, and thus has a unique stationary distribution.

For two joint strategies π and π', we define

$$\mathbb{P}(\pi \to \pi') = \begin{cases} \mathbb{P}\left((\pi_k, \pi_{-k}) \to (\pi'_k, \pi_{-k})\right) & \text{if } \pi' = (\pi'_k, \pi_{-k}) \\ 0 & \text{otherwise.} \end{cases}$$

With these transition probabilities defined, we can define the Markov chain's transition matrix.

Computing α-Rank

The above transition probabilities yield a transition matrix of the following form

$$\Delta = \begin{pmatrix} 1 - \sum\limits_{\substack{\pi \in \Pi \\ \pi \neq \pi_1}} \mathbb{P}\left(\pi_1 \to \pi\right) & \mathbb{P}\left(\pi_1 \to \pi_2\right) & \dots & \mathbb{P}\left(\pi_1 \to \pi_K\right) \\ \mathbb{P}\left(\pi_2 \to \pi_1\right) & 1 - \sum\limits_{\substack{\pi \in \Pi \\ \pi \neq \pi_2}} \mathbb{P}\left(\pi_2 \to \pi\right) & \dots & \mathbb{P}\left(\pi_2 \to \pi_K\right) \\ \dots & \dots & \dots & \dots \\ \mathbb{P}\left(\pi_K \to \pi_1\right) & \mathbb{P}\left(\pi_K \to \pi_2\right) & \dots & 1 - \sum\limits_{\substack{\pi \in \Pi \\ \pi \neq \pi_K}} \mathbb{P}\left(\pi_K \to \pi\right) \end{pmatrix}.$$

In the single-population case, each row and column correspond to one policy in Π, and $K = |\Pi|$. In the multi-population case, each row and column corresponds to a **joint** strategy, and $K = |\Pi|^N$. Note that in the multipopulation case with many players, most entries in Δ are null because most joint strategies are not within a Hamming distance of 1.

The alpharank distribution σ is defined as the stationary distribution of the Markov chain whose transition matrix is Δ, which can be found by computing the eigenvector of $\Delta \cap \mathbb{R}_+$ of eigenvalue 1. This can be done easily using the power method, since the stationary distribution is the unique distribution $\sigma \in \Delta(\Pi)$ such that

$$\sigma^t \Delta = \sigma^t,$$

and 1 is the highest eigenvalue of Δ, making the power method perfect for finding σ.

Intuition behind α-Rank

α-Rank's intuitive idea is to characterize chains of best-responses, resembling those in the Rock-Paper-Scissors earlier: if I play Rock, you play Paper. I will then switch to Scissors. But that will make you switch to Rock. If we repeat this process infinitely many times, each state's frequency will be $\frac{1}{3}$, which is the α-Rank distribution of Rock-Paper-Scissors when $\alpha = \infty$ and $m > 1$.

The full intuition behind the Markov chain's transition probabilities and hyperparameters is provided in Omidshafiei et al. [138]. They originate from evolutionary game theory dynamics models. Large values of α correspond to high *selection pressure* in the evolutionary model under consideration. Note that the version of α-Rank used throughout this work corresponds to the limiting invariant distribution as $\alpha \to \infty$, under which only strategy profiles which correspond to best-responses can have positive mass.

α-Rank solves the equilibrium selection problem by ensuring the stationary distribution is unique; and the equilibrium computation problem by only requiring the computation of the stationary distribution of a Markov chain. This computation is equivalent to computing the only eigenvector of eigenvalue 1 of the Markov chain's transition matrix, which is polynomially complex in the number of joint strategies of the game.

Notions of Sink Strongly-Connected Components (SSCC):

One key concept of α-Rank is Sink Strong Chain Components (SSCCs). Intuitively, a Sink-Strong Chain Component is an "optimal cycle" in the Markov-chain mentioned above: a fully connected set of states with *no outgoing edges*. While α-Rank's α is finite, there is only one SSCC in the game, which englobes all existing states. However, if $\alpha = \infty$, the game may be divided between disjoint SSCCs. Imagine for example a 3-action, 2-player game where joint actions $(1, 1)$ and $(3, 3)$ offer the highest reward for both players, and all other joint actions are suboptimal. The game therefore has two disjoint SSCCs: there indeed does not exist a best-response-path between $(1, 1)$ and $(3, 3)$, and both joint strategies are "stable" - there exists no outgoing edges leaving from them.

2.1.6 (Coarse) Correlated Equilibrium

Let us imagine that we are the mediator to a game, whatever its type. Our role as a mediator is to help the players find themselves in the best situations possible for them; but also to be listened to. It is therefore in our interest to suggest to each player a course of action that will be the best possible.

However, we also benefit from another advantage: players are only aware of the recommendation we gave them, but not of other players' recommendations. This information asymmetry allows us to develop more complex behaviors than Nash equilibria by taking advantage of the uncertainty one player has over the actions of the other players.

Several optimality principles may be defined from this setting; we choose to focus on two, which will yield correlated, and coarse-correlated equilibria. Intuitively, correlated equilibria correspond to mediators whose recommendations from which there is no incentive to deviate. Coarse-correlated equilibria are rougher mediators, characterized by the fact that players only stand to lose if they decide to completely ignore the mediator's recommendations and consistently play the same policy instead.

These concepts are akin to what can be termed "soft governance": instead of imposing a certain course of actions to coerced agents, (coarse) correlated equilibria coordinate free-willed agents in such a way that the best thing for them is to follow the coordinated instructions. This is "the best of both worlds": centralized instructions, for which everyone is content enough to keep acting.

Let us now formally introduce correlated and coarse-correlated equilibria (CEs and CCEs), and a few of their properties.

Definition 2 (Correlated Equilibrium)**.** We say that a mediator $\rho \in \Delta(\Pi^N)$ is an $\epsilon \geq 0$-correlated equilibrium if

$$\mathbb{E}_{\pi\sim\rho}[J_i(u[\pi_i], \pi_{-i}) - J_i(\pi_i, \pi_{-i})] \leq \epsilon \qquad \forall u \in \mathcal{U}_{CE} \tag{2.5}$$

Whenever $\epsilon = 0$ in Equation 2.5, ρ is a correlated equilibrium.

Definition 3 (Coarse-Correlated Equilibrium)**.** We say that a mediator $\rho \in \Delta(\Pi^N)$ is an $\epsilon \geq 0$-coarse-correlated equilibrium if

$$\mathbb{E}_{\pi\sim\rho}[J_i(u[\pi_i], \pi_{-i}) - J_i(\pi_i, \pi_{-i})] \leq \epsilon \qquad \forall u \in \mathcal{U}_{CCE} \tag{2.6}$$

Whenever $\epsilon = 0$ in Equation 2.6, ρ is a coarse-correlated equilibrium.

As mentioned above intuitively, we see that CEs and CCEs only differ by their deviation-optimality: whereas correlated equilibria are robust to players deviating from each recommendation, coarse-correlated equilibria are only robust to players deviating agnostically from all recommendations (without the ability to choose from which recommendation to deviate).

An important property of these equilibria is their convexity:

Proposition 3 ((C)CE Convexity)**.** *The set of $\epsilon \geq 0$-(coarse) correlated equilibria is convex.*

Proof. Let $\epsilon \geq 0$, ρ_1 and ρ_2 be two ϵ-(coarse) correlated equilibria, and $u \in \mathcal{U}_{\{CE,CCE\}}$. Let $t \in [0, 1]$.

$$\mathbb{E}_{\pi\sim t\rho_1+(1-t)rho_2}[J_i(u[\pi_i],\pi_{-i}) - J_i(\pi_i,\pi_{-i})] =$$
$$t\underbrace{\mathbb{E}_{\pi\sim\rho_1}[J_i(u[\pi_i],\pi_{-i}) - J_i(\pi_i,\pi_{-i})]}_{\leq\epsilon} + (1-t)\underbrace{\mathbb{E}_{\pi\sim\rho_2}[J_i(u[\pi_i],\pi_{-i}) - J_i(\pi_i,\pi_{-i})]}_{\leq\epsilon}$$
$$\leq \epsilon$$

□

A very interesting connection between coarse-correlated equilibria and Nash equilibria exists in 2-player 0-sum games: the marginalization of a coarse-correlated equilibrium is a Nash equilibrium in these games!

Proposition 4 (ϵ-Coarse-Correlated Equilibrium to ϵ-Nash Equilibrium)**.** *Let ρ be an ϵ-coarse-correlated equilibrium in a 2-player, 0-sum game, with $\epsilon \geq 0$. Then the strategies defined, for each player, by $\pi_i = \sum_{\pi\in\Pi_i}\sum_{\pi_{-i}\in\Pi_{-i}}\rho(\pi,\pi_{-i})\pi$ are Nash equilibria.*

Proof. Let ρ be a CCE, and i player 0 or 1.

Then

$$\sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})(J_i(\pi',\pi_{-i}) - J_i(\pi_i,\pi_{-i})) \leq 0 \quad \forall\pi'\in\Pi_i.$$

This is true because all CCE deviations can be seen as policies.

Thus

$$J_i(\pi', \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_{-i}) - J_i(\sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_i, \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_{-i})$$
$$= \sum_{\pi_i,\pi_{-i}}\sum_{\pi'_i,\pi'_{-i}}\rho(\pi'_i,\pi'_{-i})\rho(\pi_i,\pi_{-i})\left(J_i(\pi',\pi_{-i}) - J_i(\pi'_i,\pi_{-i})\right).$$

Since the game is 0-sum, we have that

$$\sum_{\pi_i,\pi_{-i}} -\rho(\pi_i,\pi_{-i})J_i(\pi'_i,\pi_{-i}) = \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})J_{-i}(\pi'_i,\pi_{-i})$$
$$\leq \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})J_{-i}(\pi_i,\pi_{-i})$$
$$\leq \sum_{\pi_i,\pi_{-i}} -\rho(\pi_i,\pi_{-i})J_i(\pi_i,\pi_{-i}).$$

Therefore, plugging the above inequality into the former equation, we get

$$J_i(\pi', \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_{-i}) - J_i(\sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_i, \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_{-i})$$
$$\leq \sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\left(J_i(\pi',\pi_{-i}) - J_i(\pi_i,\pi_{-i})\right)$$
$$\leq \epsilon,$$

i.e. $\sum_{\pi_i,\pi_{-i}}\rho(\pi_i,\pi_{-i})\pi_i$ is an ϵ-Nash equilibrium, which, since it is true for both i, concludes the proof. □

We now define notions of regret, a concept which is intimately linked with deviations and (coarse) correlated equilibria.

2.1.7 Adversarial Regret and its Properties

Imagine that we are tasked with optimizing the sequential choice between $\mathcal{A}$ alternatives - for example, we must, several times in a row, pick how to distribute our money between $\mathcal{A}$ different investments. Every time we do so, we observe how much each investment paid back during a given period, and are allowed to pick another distribution of our wealth over possible investments.

We are interested in minimizing *regret*, a notion of how much more we could have earned, were we to have altered the way we played in a certain way. We will examine two types of regret: *external regret*, and *internal regret* [22], first providing intuition on what they represent, then defining them rigorously.

Intuitively, having *external regret* of ϵ represents the (more or less) embarrassing situation where simply putting all our money on the same investment at every step would yield, for the best investment, ϵ more than our algorithmic approach. As one might have guessed, were we to commercialize our approach, we would like to avoid such situations as much as possible.

Internal regret is a more nuanced idea of optimality: what if, instead of putting my money on a given investment as my algorithm recommended, I chose to put it on another one, at every step - *internal regret* is the difference between the best such change, and the payoff received by following our algorithmic approach.

The returns given by the investments, although supposed to always be finite, are not constrained to anything. In particular, it could well be that an investment that never paid anything off for a large number of steps starts being the most profitable one of the lot. This is where the concept of*adversarial regret minimization*, which we define below, comes from, as the reward could potentially be set by an adversary.

We now provide formal definitions of these concepts. To do this, we take a sequence of payoffs per action $(r_t)_{t=1..T} \in \left(\mathbb{R}^{|\mathcal{A}|}\right)^T$ and distributions over actions $(p_t)_{t=1..T} \in \left(\mathbb{R}^{|\mathcal{A}|}\right)^T$.

Definition 4 (External Regret)**.** We define the external regret of the sequence $(p_t)_{t=1..T}$ by

$$\text{External Regret}\left((p_t)_t\right) = \max_i \sum_t r_t[i] - \langle p_t, r_t \rangle$$

Definition 5 (Internal Regret)**.** We define the internal regret of the sequence $(p_t)_{t=1..T}$ by

$$\text{Internal Regret}\left((p_t)_t\right) = \max_{i,j} \sum_t \left(r_t[i] - r_t[j]\right) p_t[j]$$

We note that there exists another type of regret, *swap regret*: what if, instead of only changing one action to another, we potentially changed all actions at the same time?

Definition 6 (Swap Regret)**.** We define the swap regret of the sequence $(p_t)_{t=1..T}$ by

$$\text{Swap Regret}\left((p_t)_t\right) = \max_{u \in \mathcal{U}_{CE}} \sum_t \langle u[p_t] - p_t, r_t \rangle$$

We notice that *swap regret* is very similar to *internal regret*, and only differs from it by a factor $|\mathcal{A}|\Delta r$ at most, where Δr is the maximum possible reward difference of the process. Having no swap-regret (Or *being no-swap-regret*) therefore implies having no internal-regret, and the converse is true (If there is nothing to gain by moving any policy's mass on another policy, then there can be nothing to gain by moving several at once.)!

Adversarial regret minimization is the setting of minimizing some type of regret, typically internal or external, against a partially-observed sequence of reward functions $(r_t)_t$ chosen by an adversary, of which, at time τ, only $(r_t)_{t=1..\tau-1}$ is known by the regret minimizer. The fact that $(r_t)_t$ is chosen by an adversary means that it can potentially be the worst-case reward sequence for a given algorithm, thereby testing the limits of regret minimizers - which is exactly the case which interests us: we want to make sure our algorithms work as well as possible in the worst possible case.

There exists several well-known external-regret minimizing algorithms, such as the Polynomial Weights Algorithm, presented in Algorithm 1 (And which assumes rewards $> \frac{-1}{\eta}$), or regret matching, in Algorithm 2. These two algorithms are such that External Regret $= \mathcal{O}\left(\sqrt{T}\right)$ (with differing constants). We note that this means that their average regret (regret divided by T, the regret resulting from sampling uniformly rewards and plays) is $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$, and thus tends to 0 / negative values as time grows to infinity.

Algorithm 1 Polynomial Weights(η, T)

Initialize $w_i^1 = 1$, $p_i^1 = \frac{1}{N} \quad \forall i$.
for $t \leq T$ **do**
  Observe reward vector r^{t+1}.
  Set $w_i^{t+1} = w_i^t \left(1 + \eta\, r_i^{t+1}\right)$.
  Set and play $p_i^{t+1} = \frac{w_i^{t+1}}{\sum_j w_j^{t+1}}$.
end for
Return $(p^t)_t$

Algorithm 2 Regret Matching(T)

Initialize $Reg_i^1 = 0$
for $t \leq T$ **do**
  if $\sum_j max(0, Reg_j^t) > 0$ **then**
    $p_i^t = \frac{max(0, Reg_i^t)}{\sum_j max(0, Reg_j^t)} \quad \forall i$
  else
    $p_i^t = \frac{1}{N} \quad \forall i$
  end if
  Play p^t.
  Observe reward vector r^t.
  Compute regret vector $Reg_i^t = Reg_i^{t-1} + r^t[i] - \langle r^t, p^t \rangle$
end for
Return $(p^t)_t$

Blum [22] also presents a way to convert external-regret minimizing algorithms into internal-regret minimizing algorithms.

We now make the following point: when playing in an N-player game, from the point of view of one player, the N-1 others act like an adversary modifying the player's reward. In this setting, it makes sense to use regret-minimizing algorithms, and if all players minimize their internal (external) regret, *they converge towards a (coarse) correlated equilibrium.*

Let us clarify this assertion. Assume that all players follow a regret-minimizing strategy, and all have average regret $\leq \epsilon$. Then uniformly sampling from players' joint strategies yields an ϵ correlated or coarse-correlated equilibrium, depending on the regret type: internal regret minimization yields a correlated equilibrium; external regret, a coarse-correlated equilibrium.

This can be quickly proved by noting that external regret is actually the payoff gain for deviating unilaterally from the above-defined recommender; and swap regret is the payoff gain for swapping policies around, akin to the $\mathcal{U}_{CE}$ deviations (Swap and internal regret being equivalent).

We have therefore provided a way to find correlated and coarse-correlated equilibria in N-player games, by following no-regret minimizers and uniformly sampling their joint policies. We will use these concepts to compute approximate equilibria in Mean-Field games, where their computation with other methods is much more difficult. This approach, of iteratively converging towards no-regret sets (Instead of being able to directly find these in a closed-form manner) rejoins that of learning in Markov Decision Processes (MDPs), where the optimal policy of a given game can

not always be explicitly found, and must be approximated, which is exactly the topic of the next section.

2.2 *Follow the Sweets*: Concepts of Reinforcement Learning

The setting of Reinforcement Learning is intrinsically sequential, and the appropriate concept to model it are Markov Decision Processes (MDPs). A Markov Decision Process is a quadruplet $(\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ a finite set of actions, $P : \mathcal{S}, \mathcal{A} \to \Delta(\mathcal{S})$ (Also noted $P(s_1 \mid s_0, a)$) is a transition function (Usually represented as a matrix), and $r : \mathcal{S}, \mathcal{A} \to \mathbb{R}$ a reward function.

The whole goal of Reinforcement Learning is, for a given $\gamma \in]0, 1[$, to maximize its discounted payoff function

$$J(\pi) = \mathbb{E}_{(s_t, a_t)_t \sim \pi} \left[\sum_t \gamma^t r(s_t, a_t) \right]$$

where $(s_t, a_t)_t \sim \pi$ means that $(s_t, a_t)_t$ come from a process which follows π to choose its actions.

A useful tool in this setting is the Value Function of a given policy, at a given state. It is defined following

$$V^\pi(s_0) = \mathbb{E}_{(s_t, a_t)_t \sim \pi | s_0} \left[\sum_t \gamma^t r(s_t, a_t) \right],$$

the expected discounted value of policy π when starting at state s_0.

We can also write this equation in a recursive fashion:

$$V^\pi(s_0) = \sum_a \pi(s_0, a) \left(r(s_0, a) + \gamma \sum_{s_1} P(s_1 \mid s_0, a) V^\pi(s_1) \right).$$

Taking a vector-based approach (With $R^\pi(s) = \sum_a r(s, a)\pi(s, a)$ and $P^\pi(s_1, s_0) = \sum_a \pi(s_0, a)P(s_1 \mid s_0, a)$)

$$V^\pi = R^\pi + \gamma P^\pi V^\pi,$$

which yields the following (Potentially ill-conditioned if $\gamma = 1$) expression for V^π

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi.$$

We can also define an "augmented" value function, the Q-function, which is the expected value of following a policy π from a given state and after having taken a given action. It is defined following

$$Q^\pi(s_0, a) = r(s_0, a) + \gamma \sum_{s_1, a_1} P(s_1 \mid s_0, a)\pi(s_1, a_1)Q^\pi(s_1, a_1)$$

If we consider the space of (state, action) pairs, then we can also vectorize the above equation. We write $P^\pi(s_1, a_1 \mid s_0, a) = P(s_1 \mid s_0, a)\pi(s_1, a_1)$, R the reward vector, and get the following expression

$$Q^\pi = R + \gamma P^\pi Q^\pi,$$

yielding

$$Q^\pi = (I - \gamma P^\pi)^{-1} R.$$

In the next section, however, we will for completeness quickly examine techniques to find the optimal policy in a close-form fashion via Dynamic Programming. We will then return to reinforcement learning and examine two of the most popular algorithms used there: Q-learning (And its Deep learning variant), and Policy Gradient.

2.2.1 Dynamic Programming

The core idea of Dynamic Programming is that, in some cases, finding the optimal solution for one problem requires finding the optimal solution of sub-problems set in a hierarchical manner.

Let us take the example of finding the shortest path between point A and point B. Assume it goes through point C (Among others). Then we know that our path takes the shortest path between A and C, and C and B. This is directly true: if there existed another shorter path between A and C or C and B, then changing the current path to go through this shorter path would yield an even shortest path, which is impossible.

The consequence is that some types of problems (Notably problems which can be represented as directed acyclic finite graphs) can be solved in an iterative fashion by computing optimal solutions of every intermediate sub-problem. This leads *e.g.* to Dijkstra's algorithm [55].

The form of dynamic programming we are interested in here is Best Response Computation. Indeed, one can compute an exact best response, *i.e.* a policy which maximizes reward, in an exact fashion in MDPs. Of course, although this method always works theoretically, in practice, since it requires full coverage of the MDP, it is not viable when $\mathcal{S}$ or/and $\mathcal{A}$ are very large.

Exact Best-Response computations are used in the specific setting of finite-horizon MDPs: MDPs where there are only a finite amount of steps before termination. Note that, since the setting is supposed to be markovian, this also supposes that one can never return to a former state. We will often use exact best responses in the rest of this work, and therefore present how they are computed in Algorithm 3.

Algorithm 3 Exact Best-Response(State s, Tabular Policy π)

if s is terminal **then**
  Return final reward $r(s)$, π
end if
Set max value $= -\infty$, max action $= -1$.
for All actions $a \in \mathcal{A}$ available at s **do**
  Compute next states s', probabilities $p(s' \mid s, a)$ and $r(s, a)$
  Let average value $= r(s, a)$
  for All successor states s' **do**
    Compute $V_{s'}, \pi_{s'} =$ Exact Best-Response(s')
    average value $+= p(s' \mid s, a)V_{s'}$
    $\pi =$ Merge(π, $\pi(s')$)
  end for
  if average value $>$ max value **then**
    max value $=$ average value
    max action $= a$
  end if
end for
$\pi(s, \text{max action}) = 1.0$
Return max value, π

As mentioned above, computing an exact best response requires going through the whole game tree, which can be extremely costly. Approximate schemes have therefore been developed to compute such best responses in a more approximate and less computationally-demanding fashion. We will start with Q-learning, whose tabular version is actually more computationally demanding than best-response computation, but which, as we will see, can be stochastically approximated.

2.2.2 (Deep) Q-Learning

The core idea behind Q-learning is to learn the Q-function, Q_1, of a policy acting greedily with respect to another Q-function[3] Q_0. In turn, once this new Q-function Q_1 has been learnt, we learn a new Q-function, Q_2, that of a greedy policy with respect to Q_1. This creates a sequence of Q-functions which, eventually and provably converges towards Q^*, the Q-function of the MDP's optimal policy. Note that a policy which is greedy with respect to Q^* is optimal.

The way this is proven is via the use of the improvement operator $\mathcal{T}^*$, defined for Value functions following

$$\mathcal{T}^*V(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s' \mid s,a)V(s')$$

We see that $\mathcal{T}^*$ is γ-contractive for the infinite norm for Value-functions:

$$\begin{aligned}
||\mathcal{T}^*V_1(s) - \mathcal{T}^*V_2(s)||_\infty &= \max_a \left| \gamma \sum_{s'} p(s' \mid s,a) \left(V_1(s') - V_2(s')\right) \right| \\
&\leq \gamma \max_a \sum_{s'} p(s' \mid s,a) \underbrace{|(V_1(s') - Q_2(s'))|}_{\leq ||V_1(s) - V_2(s)||_\infty} \\
&\leq \gamma ||V_1(s) - V_2(s)||_\infty
\end{aligned}$$

Of course, one can always define a Q-function from a value function following

$$Q(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a)V(s').$$

And taking a greedy policy π with respect to Q means selecting, at each state, the action a which verifies

$$a = \arg\max_a r(s,a) + \gamma \sum_{s'} p(s' \mid s,a)V(s'),$$

which means that π will have $\mathcal{T}V$ as a value function. When iterating the process of applying $\mathcal{T}$ on a value function (*i.e.* iteratively computing greedy policy thereupon, and their value functions), we thus apply $\mathcal{T}$ n times. Since $\mathcal{T}$ is γ-contractive, and $\gamma < 1$, this process leads to a fixed point. It is quick to verify that this fixed-point is the optimal policy of the game.

Tabular Q-learning therefore consists of, at each iteration, updating the Q-function. The update is of the form, if Q_t is the former Q-function,

$$Q_{t+1}(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) \max_{a'} Q_t(s',a')$$

which is the Q-function of a greedy policy with respect to Q_t.

Of course, despite exponential convergence speed in the number of iterations, each iteration has the same complexity as dynamic programming (Though Q-learning may treat infinite-horizon problems, something which Dynamic Programming may not directly do): going through the whole state-action space.

The way one usually solves this problem is through stochastic approximation of the operator $\mathcal{T}$, which means approximating the update

$$Q_{t+1}(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) \max_{a'} Q_t(s',a'),$$

by minimizing the difference between Q_{t+1} and the term on the right. If we parameterize Q with parameter vector θ - we suppose that Q is continuously differentiable with respect to θ -, this means finding

[3] A policy which acts greedily with respect to a Q-function is a policy which selects, at every state, the action with the highest Q-value.

$$\arg\min_\theta d\left((Q_\theta(s,a))_{s,a},\quad (r(s,a)+\gamma\sum_{s'}p(s'\mid s,a)\max_{a'}Q_\theta(s',a'))_{s,a}\right),$$

where d is a chosen distance over the state-action space. A first question, quickly answered, is the question of weighting different state-action pairs. Since we do not want to work with the full state space, we will use stochastic approximation of the above loss using sampled state-action pairs, and the natural way to do so is to use the current greedy policy's state distribution (Augmented with some exploration so no state has 0 probability of being visited).

This changes the update rule to

$$\arg\min_\theta \mathbb{E}_{s,a\sim\pi(Q_\theta)}\left[d\left(Q_\theta(s,a),\quad r(s,a)+\gamma\sum_{s'}p(s'\mid s,a)\max_{a'}Q_\theta(s',a')\right)\right],$$

where $\pi(Q_\theta)$ is the greedy policy derived from Q_θ, and d is now a distance on $\mathbb{R}$. We will typically use the squared L2 distance, yielding

$$\arg\min_\theta \mathbb{E}_{s,a\sim\pi(Q_\theta)}\left[\left(Q_\theta(s,a)\quad -\quad r(s,a)-\gamma\sum_{s'}p(s'\mid s,a)\max_{a'}Q_\theta(s',a')\right)^2\right].$$

We can now minimize this loss using *e.g.* gradient descent over randomly sampled batches of data.

However, this setup actually incentivizes Q functions to overestimate Q-values, as demonstrated by Van Hasselt et al. [178]. They suggest using a target Q-function, parameterized by θ', to stabilize the Q-learning update, and updating Q' on different data, or in different ways. In practice, θ' is usually a Polyak-averaged version of θ, or is just set to the current θ every n update steps, and otherwise kept constant.

This provides stability over the targets, while θ is trying to find the correct Q-value estimates for correctly approximating $\mathcal{T}$.

All in all, the final objective becomes

$$\arg\min_\theta \mathbb{E}_{s,a\sim\pi(Q_\theta)}\left[\left(Q_\theta(s,a)\quad -\quad r(s,a)-\gamma\sum_{s'}p(s'\mid s,a)\max_{a'}Q_{\theta'}(s',a')\right)^2\right].$$

When using neural networks to approximate the Q-functions, this algorithm is called Deep Q-learning.

2.2.3 Policy Gradient

Instead of trying to find an exponentially-quick way to converge to an optimal policy, why aren't we simply *maximizing the payoff function directly*?

This is the idea behind Policy Gradient.

If we take a θ-parameterized policy function π_θ, our objective is to find

$$\theta=\arg\max_\theta J(\pi_\theta)$$

We recall that $J(\pi_\theta)$ is the expected payoff for playing π_θ. As a matter of fact, we can link $J(\pi_\theta)$ with V^π

$$J(\pi_\theta)=\mathbb{E}_{s_0}\left[V^{\pi_\theta}(s_0)\right].$$

We introduce d^{π_θ} the γ-discounted state occupancy measure of π_θ,

$$d^{\pi_\theta}(s)=\sum_t\gamma^t\mathbb{P}(s_t=s\mid\pi_\theta),$$

and rewrite J using Q values and assuming for simplicity one unique starting state s_0

$$J(\pi_\theta) = V^{\pi_\theta}(s_0) = \sum_a Q^{\pi_\theta}(s_0, a)\pi_\theta(s_0, a).$$

Differentiating through this expression, we get

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_a \pi_\theta(s_0, a)\nabla_\theta Q^{\pi_\theta}(s_0, a) + Q^{\pi_\theta}(s_0, a)\nabla_\theta \pi_\theta(s_0, a),$$

but

$$\nabla_\theta Q^{\pi_\theta}(s, a) = \gamma \sum_{s'} p(s' \mid s, a)\nabla_\theta V^{\pi_\theta}(s'),$$

thereby leading to

$$\begin{aligned}\nabla_\theta V^{\pi_\theta}(s_0) &= \sum_a Q^{\pi_\theta}(s_0, a)\nabla_\theta \pi_\theta(s_0, a) + \pi_\theta(s_0, a)\gamma \sum_{s'} p(s' \mid s, a)\nabla_\theta V^{\pi_\theta}(s') \\ &= \sum_s d^{\pi_\theta}(s)\nabla_\theta \pi_\theta(s, a)Q^{\pi_\theta}(s, a),\end{aligned}$$

which gives us the policy gradient formula, used as the root of so many great algorithms [150] for Reinforcement Learning breakthroughs! ...

Well, not really.

It turns out that the γ term in d^{π_θ} is ignored most of the time, as shown by Nota and Thomas [137], and the state occupancy measure is used instead! This of course biases results and removes many convergence guarantees. The purpose of this paragraph is to attract attention to the difference between the *discounted* state distribution (Which is **not** the outcome of sampling encountered states), and the state distribution (Which **is** the outcome of sampling encountered states).

We have shown some ways to learn optimal policies in single-player games. However, multiplayer games introduce a whole new array of complexity. The next question therefore provides an answer to the questions, how can we learn optimal policies in multiplayer games - and what even are optimal policies in multiplayer games?

2.3 *Learning to Play*: Learning in Games

The topic of learning in games is notoriously harder when there are several learning entities at once: since other players alter the observed environment as they play (by typically modifying the reward and transition functions), the whole learning process becomes non-stationary.

In N-player general-sum games, another question becomes: what does "learning" and "solving" a game mean? We have chosen, during this thesis, to consider that "solving" a game means "being able to reach a given equilibrium", with the equilibrium being specified by the user.

Solving two-player zero-sum games is eased (Though in no way trivialized) by the fortunate fact that, in this setting, the marginalization of a coarse-correlated equilibrium yields a Nash equilibrium. Stated differently, this means that, if one takes a no-external-regret algorithm and computes its average policy (the average of all policies it played over time), then this average will converge to a Nash equilibrium. These methods constitute regret-based methods. Other methods take advantage of special properties of 2-player 0-sum games - it is possible to modify them enough that the modification becomes very easy to solve, all the while remaining close to the true game. Another set of methods uses search, and stops searching using learnt value functions. Finally, another set of methods iterates best-responses: by continually maximizing value against a given, moving, objective, the average policy eventually becomes optimal.

2.3.1 Regret-minimization-based Methods

The main regret-minimization method is called CounterFactual Regret minimization (CFR). Its high effectiveness in small games has sparked a high number of works attempting to accelerate it, and to scale it. Unfortunately, though the method was indeed significantly sped-up, its complexity is still too high for large games.

CFR's main idea rests upon the fact that minimizing regret is an operation that can be done *locally at every state*. Local regret R_s is defined at every state s, for every available action a, by $R_s(a) = \sum_{t=1}^{T} Q_t(s,a) - \langle Q_t(s,\cdot), \pi_t(s,\cdot)\rangle$, where t is the learning time, $Q_t(s,a)$ is the value of taking action a at state s at time t, and π_t is the policy played at time t by the algorithm. Intuitively, local regret is the regret for having taken a given action instead of another - this is akin to running a bandit problem per state. Local regret can be linked to global regret (*i.e.* external or internal regret) via a *CFR theorem* which proves that minimizing local regrets leads to minimizing global regret.

Unfortunately, despite many attempts at scaling up this algorithm via the introduction of function approximation [31, 71, 170, 174], none have been able to scale up to very large games outside of abstraction-based methods [32], which are dimensionality-reduction methods. These methods have managed to solve Poker, but have not been used on other games, which may be due to a high amount of handcrafting in their scaling-up.

2.3.2 Search-based Methods

Search methods became world-famous when DeepMind's AlphaGo [165] beat the world champion of Go in 2016. AlphaGo was then refined into AlphaZero [166]. Both approaches rely on learning-aided-search, an extremely successful approach in perfect-information games.

The main idea behind AlphaZero is to combine search with a learnt policy function. While playing, the algorithm searches for an optimal action to take. To do so, it explores the game tree by more or less following its policy function, and keeps a count of which action it has selected in which states. Many different searches are run from the current state, and an action policy is generated from action counts.

Despite no convergence proofs (though advances such as [70] might begin to provide some answers), this algorithm has reached grandmaster-level in Chess, Go and Shogi, consistently beating top humans.

Yet, there exists a category of games which Alphazero and AlphaZero-like methods cannot solve: imperfect-information games. Indeed, Alphazero only works in games where one observes the full game state - Alphazero always knows that a given piece is a rook, a bishop or a king. In contrast, in games such as Stratego, a piece could be anything, from the highest to the lowest level. Search-based methods therefore need to enumerate every possible true state of the game, and run search on those. Since these may number in the 10^{100} or more, this operation is impossible in practice. Player of Games [161] is an attempt at solving this problem by combining search with regret-minimization methods, and has provided quite some success, at the cost of heavy compute requirements.

2.3.3 Regularization-based methods

An old type of method has resurfaced in recent years: regularization methods. While it has been known for a long time that it was easy to find the Nash of an entropy-regularized game, and that, as the regularization waned, the sequence of these Nash equilibria continuously converged to the true Nash of the game [84], such methods are known to be quite unstable when the regularization goes under a game-dependent threshold.

Friction FoReL, introduced by Perolat et al. [146], keeps the idea of computing the equilibria of regularized games for which equilibrium computation is straightforward. However, instead of vanishing the regularization, said regularization is parametrized by a policy - which can be the former game's Nash. The algorithm produces a sequence of games regularized by the former game's

Nash equilibrium, and it is proven that this sequence converges to the true Nash of the game - and, empirically, it does so *stably*. The method has been used to compute a human-level AI in Stratego [148], demonstrating its ability to scale to extremely complex and large games.

However, it is extremely unclear how to adapt this type of methods to reach other types of equilibria, or whether their great convergence properties extend to N-player games.

2.3.4 Iterated Best-Responses-based methods

The following methods, one of which will be at the center of this work, are based on the idea that averaging a sequence of best responses, computed in the "right" way, leads to equilibria. The "right" way being up to debate, several algorithms came to be.

The first, Fictitious Play [156], is also the simplest: the sequence of best-responses is such that the new best-response at time $t+1$ is computed against its own average at time t.

Double Oracle [117] attempts to be more refined - instead of computing a best-response against an average, which would mean putting mass on policies which may not be very good, why not compute a best-response against the Nash, the "best possible" distribution, of the set of policies we have computed so far?

Finally, PSRO [98] generalizes both algorithms by allowing one to change the above distribution: Nash (which yields Double Oracle), uniform (Fictitious Play), or any other solvers may be used. The initial paper was focused on stochastic solvers, which would help best-response computation steps explore new spaces of the policy space; however, this dissertation was interested in using new solvers to compute new equilibria.

Fictitious Play

As stated above, Fictitious Play [156] is a two-step algorithm for 2-player, 0-sum games. From an initial policy $\bar{\pi}_0 = \pi_0 \in \bar{\Pi}$, one step of the symmetrized algorithm at iteration t consists of

1. Computing $\pi_{t+1}^{BR} = BR(\bar{\pi}_t)$,
2. Computing the average $\bar{\pi}_{t+1} = \frac{1}{t+1}\pi_{t+1}^{BR} + \frac{t}{t+1}\bar{\pi}_t$.

The policy $\bar{\pi}$ defined above converges towards Nash equilibria in 2-player, 0-sum games. However, and perhaps surprisingly, *it is not no-regret* [79, 189], only its continuous version is [182]: it is not because Fictitious Play averages over a no-regret sequence, as one would have expected, that it converges to a Nash equilibrium, but because of the intrinsic link between link Fictitious Play's average dynamics and 2-player 0-sum games' innate structures.

Double Oracle

Double Oracle can be seen as an attempt to speed up Fictitious Play. Instead of uniformly averaging over policies as Fictitious Play does, Double Oracle computes the Nash equilibrium of the restricted set of policies it has computed. To do so, it requires access to a payoff matrix registering how all discovered policies of a given player fare against those of the other player. It then runs a Nash equilibrium solver on this partial game-representing matrix, and uses this Nash distribution to mix policies optimally, as shown in Algorithm 4.

Policy Space Response Oracle

Introduced by Lanctot et al. [98], Policy Space Response Oracle (PSRO) can be seen as a generalization of Double Oracle [117] and Fictitious Play. The algorithm, which we present in a simplified, symmetric form in Algorithm 5, is intrinsically many-player oriented, and its convergence proof works in N-player, general-sum games.

However, this generality is counterbalanced by its convergence speed in the worst case: in games where the Nash has full support, PSRO has to potentially be iterated once for every deterministic strategy of every player to represent it. This quantity is exponential in the number of states and

Algorithm 4 Double Oracle

Require: Π^0 initial policy pool and σ^0 distribution over Π^0.
$N = 0$
while $\Pi^N \neq \Pi^{N-1}$ **do**
  $N = N + 1$
  Compute $\pi^* = \arg\max_{\pi'} \mathbb{E}_{\pi \sim \sigma^{N-1}}[J_i(\pi'_i, \pi_{-i})]$
  $\Pi^N = \Pi^{N-1} \cup \{\pi^*\}$
  Compute payoff matrix $J^N[i, j] = J(\Pi_i^N, \Pi_j^N)$, where π_i^N, π_j^N are policies of Π^N.
  Compute Nash equilibrium σ^N over J^N.
end while

actions; hence making PSRO an unwieldy algorithm in the worst case. However, on small and medium games, it empirically performs really well, and modifications of the algorithm such as those used for Capture the Flag [89], or the AlphaStar league [181], have contributed to major breakthroughs in Multiagent Reinforcement Learning.

As the name suggests, PSRO introduces the notion of *response oracles*. A response oracle is a function which takes a restricted game pool Π_N and a distribution ρ, and outputs a subset of Π. These are traditionally Best Responses operators, either exact or RL-derived; but this dissertation will introduce new types thereof.

Algorithm 5 Policy Space Response Oracle (PSRO)

Require: Π^0 initial policy pool and σ^0 distribution over Π^0.
$N = 0$
while $\Pi^N \neq \Pi^{N-1}$ **do**
  $N = N + 1$
  Compute $\pi^* = \arg\max_{\pi'} \mathbb{E}_{\pi \sim \sigma^{N-1}}[J_i(\pi'_i, \pi_{-i})]$
  $\Pi^N = \Pi^{N-1} \cup \{\pi^*\}$
  Compute payoff matrix $J^N[i, j] = J(\Pi_i^N, \Pi_j^N)$, where π_i^N, π_j^N are policies of Π^N.
  Compute new distribution σ^N over J^N.
end while

The *restricted game* derived from Π is a Normal Form Game defined by considering each joint policy in Π as a joint action. Choosing an action means playing its corresponding joint policy; the return for each action is its corresponding joint policy's expected per-player return. We see that PSRO is extremely close to Double Oracle, the only striking difference lying in its allowing for any meta-solver, not just Nash equilibria.

Theorem 5 (Convergence to Nash [98])**.** *When σ is the Nash equilibrium of the restricted game derived from Π^N (*i.e. *PSRO is Double Oracle), PSRO converges towards the Nash equilibrium of the game.*

Proof. This proof entirely rests upon the finiteness of the sets of deterministic policies of the game.

PSRO must necessarily terminate, since there is a finite number of deterministic policies in the game.

At termination, the σ is such that there is no new best-response outside of Π which improves value against it - which means that at least one best-response against σ exists within Π. This means that best-responses cannot improve on the value of σ, since σ is a Nash of the restricted game. This means that σ is a Nash equilibrium of the true game. □

Importantly, note that PSRO's convergence is very brutal, as, before its last iterate, there is no guarantee that its distance from the true equilibrium, even its value-based distance (exploitability) will be any close to optimal. Yet at its last iterate, and it always reaches its last iterate eventually, it has found the equilibrium. Note that this is still converging, according to the mathematical

definition of convergence: PSRO's distance to its limit reaches 0 eventually, and never goes away from it after having reached it.

In the introductory paper of PSRO [98], a variant of the replicator dynamics [115, 176], called the Projected Replicator Dynamics (PRD), has been used as an approximate Nash meta-solver.

Algorithm 6 Projected Replicator Dynamics (One player, one step)

Require: δ time discretization, and $\gamma > 0$ exploration component, Q_t payoff vector of current player at time t, T number of steps, π_t current policy.

1: Set $\Delta\pi = \pi \odot (Q_t - \pi^t Q_t)$

2: Compute $\pi_{t+1} = \arg\min_{\pi' \in \Delta^{\frac{\gamma}{K+1}}} \|\pi' - (\pi_t + \delta\Delta\pi)\|$

3: **return** π_{t+1}.

where $\Delta^{\frac{\gamma}{K+1}} = \{\pi \mid \forall k,\ \pi_k \geq \frac{\gamma}{K+1}\}$.

There are no convergence guarantees on PRD; however, it is hypothesized that taking the average of the policies returned by Algorithm 6, provided that all players run PRD at the same, yields an approximation of a Nash equilibrium.

2.4 *E Pluribus Unum*: Concepts of Mean-Field Games

We have so far analyzed learning in games with N players, with N varying from 1 to any finite number of players. However, as this number increases, games become increasingly difficult to solve: combinatorial effects arise; if one wishes to, for example, compute the payoff matrix of a 10 player game with 10 actions for each player, one will have to compute an object of size 10^{10} for each player, to take into account every different action combination! This is of course not practically feasible in situations with 100s of players, let alone millions or billions.

Computing equilibria at such a scale requires a change of framework. Following the steps of statistical physics, if we can consider that all players are interchangeable, and only their states matter, *i.e.* if the game is symmetric, then we can make the assumption that players are infinite - and only examine one quantity, their distribution. The game defined by one representative player (an average player, who will indicate whether at some state, some agents could be tempted to act in such or such way) playing against the state distribution of an infinite population, on which this agent has **no** impact, is called a Mean-Field Game, a notion co-introduced by [86, 102].

In Mean-Field games, all agents are assumed to be independent decision makers, who only impact one another via their state distributions' impact on reward and dynamic functions. As such, we are interested in finding game-theoretic equilibria in Mean-Field games, that is, policies for which agents never have an incentive to deviate.

Rigorously, we define a Mean-Field Game as a quintuplet $(\mathcal{S}, \mathcal{A}, r, p, \mathcal{T})$ where

- $\mathcal{S}$ is the set of states.
- $\mathcal{A}$ is the set of actions.
- $r : \mathcal{S}, \mathcal{A}, \Delta(\mathcal{S}) \to \mathbb{R}$ is the reward function, which depends on $\mu \in \Delta(\mathcal{S})$ the population distribution.
- $p : \mathcal{S}, \mathcal{A}, \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ is the dynamics function, which also depends on μ.
- $\mathcal{T}$ is the set of times.

We write $\mathcal{M}$ the state distribution flow of the infinite population, and we note $\mu^\pi \in \mathcal{M}$ the state distribution of a population where every agent plays π. The evolution equation of μ^π is defined as

$$\mu^\pi_{t+1}(x) = \sum_{x_t \in \mathcal{X}} \sum_{a \in \mathcal{A}} p(x \mid x_t, a, \mu^\pi_t)\pi(x_t, a)\mu^\pi_t(x_t). \tag{2.7}$$

Our quantity of interest will be the function $J : \bar{\Pi}, \Delta(\mathcal{S}) \to \mathbb{R}$ the expected payoff for an agent playing π when the population is distributed following μ.

2.4.1 Equilibria and Main Properties

In Mean-Field Games, the typical problem of interest is finding a policy $\pi : \mathcal{S}, \mathcal{T} \to \Delta(\mathcal{A})$ which is a Nash equilibrium, *i.e.* no agent has an incentive to play another policy. We provide a formal definition below.

Definition 7 (ϵ-Mean-Field Nash Equilibrium)**.** A policy $\pi \in \bar{\Pi}$ is an ϵ-Mean-Field Nash equilibrium if

$$J(\pi', \mu^{\pi}) - J(\pi, \mu^{\pi}) \leq \epsilon \quad \forall \pi \in \Pi.$$

A policy π is a Mean-Field Nash equilibrium whenever the above inequality is true for $\epsilon = 0$.

Several algorithms exist for learning Mean-Field Nash equilibria in the monotonic case, which are adaptations of N-player algorithms. We can cite two, which we will use as benchmarks later in this work, Mean-Field Fictitious Play [149] and Mean-Field Online Mirror Descent [147]. We will only provide general properties of these algorithms: they both converge to Nash equilibria in *monotonic games*, a property introduced by Lasry [102] which encourages agents to avoid crowded states.

Definition 8 (Monotonicity)**.** A game is said to be monotonic when

$$J(\pi, \mu^{\pi}) - J(\pi', \mu^{\pi}) \leq J(\pi, \mu^{\pi'}) - J(\pi', \mu^{\pi'}) \qquad \forall \pi, \pi' \in \Pi$$

Another way to write this property is to consider the reward vector $r^{\pi}(\mu) : \mathcal{S} \to \mathbb{R}$ where $r(\mu)(s, a) = r(s, a, \mu)$ for all $s \in \mathcal{S}$; and augment the vectors μ with actions: $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is now a distribution over states-action. The monotonicity property then becomes

$$\langle r(\mu^{\pi}) - r(\mu^{\pi'}),\, \mu^{\pi} - \mu^{\pi'} \rangle \leq 0 \qquad \forall \pi, \pi' \in \Pi$$

We see that monotonicity is a restrictive property, which leaves room for more general Nash-converging algorithms to emerge.

Campi and Fisher [34] recently introduced a notion of Mean-Field correlated equilibria. Their central idea is that the strategy recommender jointly samples both a policy *and* a Mean-Field state distribution flow, with the added constraint that the distribution over policies conditioned on μ induces μ. Namely, if ρ is the joint distribution over Π and $\mathcal{M}$, we have

$$\mu\left(\sum_{\pi \in \Pi} \frac{\rho(\pi, \mu)}{\sum_{\pi' \in \Pi} \rho(\pi, \mu)} \pi\right) = \mu\,, \qquad \forall \mu \in \mathcal{M},\, \sum_{\pi \in \Pi} \rho(\pi, \mu) > 0,$$

where $\mu(\sum_i \alpha_i \pi_i)$ is the mean-field flow resulting from a population sampling policies $(\pi_i)_i$ with probability $(\alpha_i)_i$ at the start of the game, and playing it until the end.

We now introduce the definition of a Mean-Field correlated equilibrium in [34]:

Definition 9 (Campi-Fisher [34] correlated equilibrium)**.** $\rho \in \Delta(\Pi \times \mathcal{M})$ is a correlated equilibrium if

$$\mathbb{E}_{\pi,\, \mu \sim \rho}\left[J(u(\pi), \mu) - J(\pi, \mu)\right] \leq 0 \qquad \forall u : \Pi \to \Pi.$$

Their paper explores convergence properties towards correlated equilibria, showing that sequences of correlated equilibria in N-player games converge to a Mean-Field correlated equilibrium; and showing that Mean-Field correlated equilibria are asymptotically optimal in N-player games, as N tends to infinity, without however providing an optimality rate.

2.4.2 Algorithms

A few algorithms have already been introduced to solve Mean-Field games. Of them, two are striking by their simplicity and proximity with Reinforcement Learning or Multiagent Reinforcement Learning; and by their only convergence requirement being monotonicity, which is in contrast with other existing algorithms which require strong contraction properties.

These two are Mean-Field Online Mirror Descent [147], and Mean-Field Fictitious Play [36, 149].

Mean-Field Online Mirror Descent: the algorithm requires estimating, at each of its steps, the current policy's Q function. It then accumulates this Q-function's output, and uses this accumulation to compute a soft policy. The process is given in Algorithm 7. If the game is monotone, it converges to its unique Nash equilibrium.

Algorithm 7 Mean-Field Online Mirror Descent

Require: learning rate $\eta > 0$, Γ the gradient of the convex-conjugate of a strongly convex function.

$t = 0$
$y_0 = 0$.
while $t > 0$ **do**
  Compute $\pi_t = \Gamma(y_t)$.
  Compute Q^{π_t} the Q-value of π_t *when the whole population plays* π_t.
  Compute $y_{t+1} = y_t + \eta Q^{\pi_t}$.
  $t = t + 1$.
end while
return π_t

Mean-Field Fictitious Play: the algorithm requires computing a best-response at each timestep to the current Mean-Field state distribution, and averaging each former best-response's Mean-Field state distribution. The process is given in Algorithm 8. If the game is monotone, it converges to its unique Nash equilibrium.

Algorithm 8 Mean-Field Fictitious Play

Require: Initial state distribution μ_0.
$t = 1$, $\mu_1 = \mu_0$.
while $t > 0$ **do**
  Compute $\pi^{BR} = \arg\max_{\pi \in \Pi} J(\pi, \mu_t)$.
  Compute $\mu^{\pi^{BR}}$.
  Update $\mu_{t+1} = \frac{t}{t+1}\mu_t + \frac{1}{t+1}\mu^{\pi^{BR}}$.
  $t = t + 1$.
end while
return $\pi_t = \pi(\mu_t)$

Chapter 3

Break the Traditions: Beyond Nash - 2 players - 0 sum

This chapter investigates how PSRO can be adapted to converge to *as many types of equilibria as possible* in *as many games as possible*. We write PSRO(σ, g) the version of PSRO using σ as an optimal distribution, and g as a response oracle.

We first investigate how PSRO can be modified to converge towards α-Rank-optimal distributions, then how it can be modified to converge towards (coarse) correlated equilibria, both in N-player general-sum games. We then generalize this approach to all equilibria of a certain form, in all games. Finally, we investigate limitations to such approaches, which lead us to the next chapter of this thesis.

3.1 Computing α-Rank-optimal strategies in N-player games

PSRO(Nash, BR) converges to a Nash equilibrium in two-player zero-sum games [117], and McMahan et al. [117]'s argument can directly be extended to N-player general-sum games. However, given the computational complexity of Nash equilibria and their other limitations outlined in Section 2.1.4, α-Rank appears to be a promising meta-solver candidate as it applies to N-player general-sum games, has no equilibrium selection problem, and has desirable computational properties; if only to evaluate the most important strategies in a given game.

However, it remains unclear how to compute α-Rank in extensive-form games; and even in normal-form games, multi-population α-Rank requires the manipulation of exponential-size matrices in the number of actions - though these matrices will be $1 - \frac{N(|\Pi|-1)-1}{|\Pi|^N}$-sparse -, thus requiring simplification.

In this context, PSRO, which slowly grows the space of policies it considers, sounds like an ideal candidate to compute α-Rank on large normal-form games and on extensive-form games, as it (a) provides a straightforward method to compute normal-form equilibria for extensive-form games, and (b) typically won't require game-theoretic solvers to be run on the full game, but only on partial, hopefully much slower slices.

However, open questions remain regarding convergence guarantees of PSRO when using α-Rank. The first question we ask ourselves is whether it suffices to replace the Nash metasolver with an α-Rank one in PSRO, still using best-responses. The second one is whether it is enough to compute a Nash equilibrium and then run α-Rank over policies it has found. In case none of these were to be enough, how should one alter PSRO to converge to the α-Rank distribution?

We summarize our results in Table 3.1, giving a full exposition below: We first start by verifying, in Section 3.1.1, whether using standard PSRO with an α-Rank solver instead of a Nash solver allows one to get to a game's SSCC. Finding a counterexample, we show that maximizing value can prevent one from finding the game's SSCC. A new objective, defined in Section 3.1.2, PBR,

Game type	$\mathcal{M}$	$\mathcal{O}$	Converges to α-Rank?
SP	α-Rank	BR	✗ (example 3)
SP	α-Rank	PBR	✓ (Sub-SSCC,† proposition 8)
MP	α-Rank	BR	✗ (example 4)
MP	α-Rank	PBR	✓ (With novelty-bound oracle,† proposition 6)
SP / MP	Uniform or Nash	BR	✗ (Examples 1 and 2)

Table 3.1: Theory overview. SP and MP, resp., denote single and multi-population games. BR and PBR, resp., denote best response and preference-based best response.

which consists in maximizing the number of opponents against which one wins, is thus proposed, which, under certain conditions, will systematically discover at least one SSCC of the true game.

3.1.1 The difficulty of Converging to α-Rank-optimal Distributions

Before doing any alteration to the original PSRO algorithm, we should first verify one thing: when using the vanilla algorithm *with no alteration*, at convergence, does the PSRO pool always contain an α-Rank-optimal cycle of the game? How about when using the uniform distribution instead of the Nash distribution as meta-solver? If so, then we do not need to adapt the algorithm to get convergence guarantees to α-Rank: running it until convergence to the Nash of the game should also yield the α-Rank-optimal distribution thereof.

However, we provide two counterexamples showing that vanilla PSRO does not always reach α-Rank-optimal cycles of the true game, thus closing this avenue of proof.

	A	B	X
A	0	1	ε
B	1	0	$-\varepsilon$
X	$-\varepsilon$	ε	0

(a) Example 1 payoff matrix.

	A	B
A	-1	1
B	1	-1
X	$-\varepsilon$	$-\varepsilon/2$

(b) Example 2 payoff matrix.

Table 3.2: Illustrative games used to analyze the behavior of PSRO in example 1. Here, $0 < \varepsilon \ll 1$. The first game is symmetric, whilst the second is zero-sum. Both tables specify the payoff to Player 1 under each strategy profile.

Examples 1 and 2 show that PSRO(Nash) does not always find policies on which respectively single-population and multi-population α-Rank puts mass.

Example 1. *Consider the two-player symmetric game specified in table 3.2a. The sink strongly-connected component of the single-population response graph (and hence the α-Rank distribution) contains all three strategies, but all NE are supported on $\{A, B\}$ only, and the best response to a strategy supported on $\{A, B\}$ is another strategy supported on $\{A, B\}$. Thus, the single-population variant of PSRO, using either $\{Nash, Uniform\}$ as metasolver, and with initial strategies contained in $\{A, B\}$ will terminate before discovering strategy X; the full α-Rank distribution will thus* not *be recovered.*

Example 2. *Consider the two-player zero-sum game specified in table 3.2b. All strategy profiles receive non-zero probability in the multi-population α-Rank distribution. However, the Nash equilibrium over the game restricted to actions A, B for each player has a unique Nash equilibrium of $(1/2, 1/2)$. Player 1's best response to this Nash is to play some mixture of A and B, and therefore strategy X is not recovered by PSRO(Nash, BR) in this case, and so the full α-Rank distribution will thus* not *be recovered.*

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	ε	ε	ε	ε	0

Table 3.3: Symmetric zero-sum game used to analyze the behavior of PSRO in example 3. Here, $0 < \varepsilon \ll 1$ and $\phi \gg 1$.

Another straightforward attempt to establish convergence to α-Rank might involve running PSRO to convergence (until the oracle returns a strategy already in the convex hull of the known strategies), using α-Rank as the meta-solver, and a standard best response oracle. However, Example 3 shows that this will not work in general for the single-population case. Figures 3.1 and 3.2 illustrate step-by-step what is described in Example 3.

Example 3. *Consider the symmetric zero-sum game specified in table 3.3. As X is the sole sink component of the game's response graph (as illustrated in fig. 3.1a), the single-population α-Rank distribution for this game puts unit mass on X. We now show that a PSRO algorithm that computes best responses to the α-Rank distribution over the current strategy set need not recover strategy X, by computing directly the strategy sets of the algorithm initialized with the set $\{C\}$.*

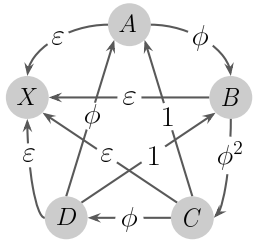
1. *The initial strategy space consists only of the strategy C; the best response against C is D.*
2. *The α-Rank distribution over $\{C, D\}$ puts all mass on D; the best response against D is A.*
3. *The α-Rank distribution over $\{C, D, A\}$ puts all mass on A; the best response against A is B.*
4. *The α-Rank distribution over $\{C, D, A, B\}$ puts mass $(1/3, 1/3, 1/6, 1/6)$ on (A, B, C, D) respectively. For ϕ sufficiently large, the payoff that C receives against B dominates all others, and since B has higher mass than C in the α-Rank distribution, the best response is C.*

Thus, PSRO(α-Rank, BR) leads to the algorithm terminating with strategy set $\{A, B, C, D\}$ and not discovering strategy X in the sink strongly-connected component.

This conclusion also holds in the multi-population case, as the following counterexample shows.

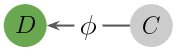
Example 4. *Consider the game in table 3.3, treating it now as a multi-population problem. We verify that the multi-population α-Rank distributions obtained by PSRO with initial strategy sets consisting solely of C for each player are: (i) a Dirac delta at the joint strategy (C, C), leading to best responses of D for both players; (ii) a Dirac delta at (D, D) leading to best responses of A for both players; (iii) a Dirac delta at (A, A), leading to best responses of B for both players; and finally (iv) a distribution over joint strategies of the 4×4 subgame induced by strategies A, B, C, D that leads to a best response* not *equal to X; thus, the full α-Rank distribution is again* **not** *recovered.*

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	ε	ε	ε	ε	0



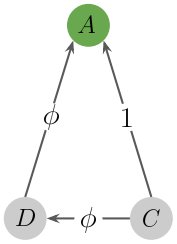
(a) Overview. The left table represents the payoff table of the game, and the graph on the right represents the full game's response graph, with values over directed edges indicating the payoff gained by deviating from one strategy to another. The table should be read thus: the row player chooses a deviation, when everyone else plays what the column player plays. We see that the α-Rank distribution is focused on X, as all arrows point to X.

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	$\boldsymbol{\phi}$	0	$-\varepsilon$
X	ε	ε	ε	ε	0



(b) Consider an initial strategy space consisting only of the strategy C; the best response against C is D.

	A	B	C	D	X
A	0	$-\phi$	1	$\boldsymbol{\phi}$	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	ε	ε	ε	ε	0



(c) The α-Rank distribution over $\{C, D\}$ puts all mass on D; the best response against D is A.

Figure 3.1: Example 3 with oracle $\mathcal{O} = \text{BR}$. In each step above, the α-Rank support is highlighted by the light green box of the payoff table, and the BR strategy against it in bold, dark green. Continued in Figure 3.2

3.1.2 A New Response Oracle

The previous examples indicate that the use of standard best responses in PSRO may be the root cause of the incompatibility between PSRO and the α-Rank solution concept.

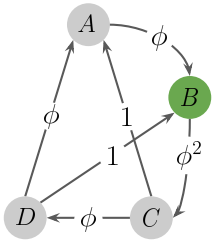
To go around this limitation, we introduce a new oracle, the *Preference-based Best Response (PBR) oracle*, which is more closely aligned with the dynamics defining α-Rank, and which enables us to establish desired PSRO guarantees with respect to α-Rank.

Single-population case We first consider the allegedly simpler single-population case.

Given an N-strategy population Π_N and corresponding meta-solver distribution $\sigma \in \Delta(\Pi_N)$, a PBR oracle is defined as any function satisfying

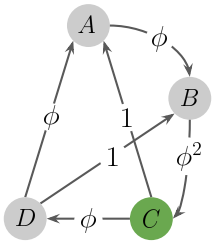
$$\text{PBR}\big(\sigma, \Pi_N\big) \subseteq \arg\max_{\pi \in \Pi} \sum_{\pi_i \in \Pi_N} \sigma_i \mathbb{1}\left[J_1(\pi, \pi_i) > J_2(\pi, \pi_i)\right] , \tag{3.1}$$

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	$\boldsymbol{\phi}$	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	ε	ε	ε	ε	0



(a) The α-Rank distribution over $\{C, D, A\}$ puts all mass on A; the best response against A is B.

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	$\boldsymbol{\phi^2}$	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	ε	ε	ε	ε	0



(b) The α-Rank distribution over $\{C, D, A, B\}$ puts mass $(1/3, 1/3, 1/6, 1/6)$ on (A, B, C, D) respectively. For ϕ sufficiently large, the payoff that C receives against B dominates all others, and since B has higher mass than C in the α-Rank distribution, the best response is C.

Figure 3.2: Second part of Figure 3.1.

where the arg max returns the *set* of policies optimizing the objective, and the optimization is over pure strategies in the underlying game. The intuition for the definition of PBR is that we would like the oracle to return strategies that will receive high mass under α-Rank when added to the population; objective 3.1 essentially encodes the probability flux that the vertex corresponding to σ would receive in the random walk over the α-Rank response graph.

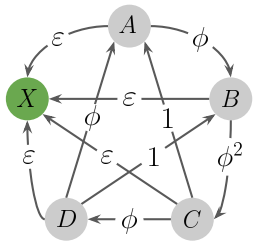
We demonstrate below that the use of the PBR resolves the issue highlighted in example 3. Figure 3.3 provides, just like in Example 3, an accompanying visual - though note that only the last step has been represented here. The former steps are assumed to be the worst possible case when X is only found at the last step of the algorithm, *i.e.* where the former steps are the same as in Figure 3.1. Indeed, since α-Rank's distribution is always fully concentrated on one strategy until the last step, it is possible that X, despite being a possible solution of PBR at every step, will not have been returned by the oracle until the last step.

Example 5. *Steps 1 to 3 of correspond exactly to those of example 3. In step 4, the α-Rank distribution over $\{C, D, A, B\}$ puts mass $(1/3, 1/3, 1/6, 1/6)$ on (A, B, C, D) respectively. A beats C and D, thus its PBR score is $1/6 + 1/6 = 1/3$. B beats A and D, thus its PBR score is $1/3 + 1/6 = 1/2$. C beats B, its PBR score is thus $1/3$. D beats C, its PBR score is thus $1/6$. Finally, X beats every other strategy, and its PBR score is thus 1. Thus, there is only one strategy maximizing PBR, X, which is then chosen, thereby recovering the SSCC of the game and the correct α-Rank distribution at the next timestep.*

Multi-population case In the multi-population case, consider a population of N strategy profiles Π_N and corresponding meta-solver distribution σ. Several meta-SSCCs may exist in the multi-population α-Rank response graph. When this happens to be the case, we run the PBR oracle for each meta-SSCC separately, as follows: Suppose there are ℓ meta-SSCCs, and denote by $\sigma^{(\ell)}$ the distribution σ restricted to the ℓ^{th} meta-SSCC, for all $1 \leq \ell \leq L$. The PBR for player k on the ℓ^{th} meta-SSCC is then defined by

$$\text{PBR}^k\left(\sigma^{(\ell)}, \Pi_N\right) \subseteq \arg\max_\pi \sum_i \sigma_i^{(\ell)} \mathbb{1}\left[J_k(\pi, \pi_i^{-k}) > J_k(\pi_i^k, \pi_i^{-k})\right] . \tag{3.2}$$

	A	B	C	D	X
A	0	$-\phi$	1	ϕ	$-\varepsilon$
B	ϕ	0	$-\phi^2$	1	$-\varepsilon$
C	-1	ϕ^2	0	$-\phi$	$-\varepsilon$
D	$-\phi$	-1	ϕ	0	$-\varepsilon$
X	$\boldsymbol{\varepsilon}$	$\boldsymbol{\varepsilon}$	$\boldsymbol{\varepsilon}$	$\boldsymbol{\varepsilon}$	0



(e) The α-Rank distribution over $\{C, D, A, B\}$ puts mass $(1/3, 1/3, 1/6, 1/6)$ on (A, B, C, D) respectively. A beats C and D, and therefore its PBR score is $1/3$. B beats A and D, therefore its PBR score is $1/2$. C beats B, its PBR score is therefore $1/3$. D beats C, its PBR score is therefore $1/6$. Finally, X beats every other strategy, and its PBR score is thus 1. There is only one strategy maximizing PBR, X, which is then chosen, and the SSCC of the game, recovered.

Figure 3.3: example 3 with oracle $\mathcal{O} = \text{PBR}$. Steps a to a are not shown as they are identical to their analogs in fig. 3.1.

Thus, the PBR oracle generates one new strategy for each player for every meta-SSCC in the α-Rank response graph; we return this full set of strategies and append to the policy space accordingly. Intuitively, this leads to a *diversification* of strategies introduced by the oracle, as each new strategy need only perform well against a subset of prior strategies. This hints at interesting links with the recently-introduced concept of rectified-Nash BR [13], which also attempts to improve diversity in PSRO, albeit only in two-player zero-sum games.

We henceforth denote PSRO(α-Rank, PBR) as α-PSRO for brevity. Now that we have defined a new algorithm, we would like to find relevant metrics to estimate how far from convergence one is at any given step, and characterize convergence quality. We introduce two such metrics, the first one, α-Conv, being an analog of NashConv for α-Rank and measuring how close to an SSCC our current pool is; the second one, PCS-Score, measures the quality of the population discovered by α-PSRO.

α-PSRO: Algorithm and Metrics

With the notation introduced in the former section, we define α-Conv, a metric akin to exploitability or NashConv to measure convergence of PSRO to the α-Rank optimal distribution.

α-Conv: Single population case: We start by defining the single-population version of PBR-Score, given by

$$\text{PBR-Score}(\pi, \sigma, \Pi_N) = \sum_i \sigma_i \mathbb{1}\left[J_1(\pi, \pi_i) > J_2(\pi, \pi_i)\right].$$

The single-population α-Conv is then defined as

$$\alpha\text{-Conv}(\pi, \sigma, \Pi_N) = \max_{\pi \in \Pi} \text{PBR-Score}(\pi, \sigma, \Pi_N) - \max_{\pi_k \in \Pi_N} \text{PBR-Score}(\pi_k, \sigma, \Pi_N) \tag{3.3}$$

where we note that $\max_{\pi \in \Pi}$ is taken over the set of pure strategies of the underlying game Π.

α-Conv: Multi-population case: We define PBR Score as

$$\text{PBR-Score}^k(\pi', \sigma, \Pi_N) = \sum_{\pi \in \Pi} \sigma(\pi) \mathbb{1}\left[J_k(\pi', \pi_{-k}) > J^k(\pi_k, \pi_{-k})\right],$$

and

$$\alpha\text{-Conv} = \textstyle\sum_{k \in \mathcal{N}} \max_{\pi \in \Pi^N} \text{PBR-Score}^k(\pi, \sigma, \Pi_N) - \max_{\pi \in \Pi} \text{PBR-Score}^k(\pi, \sigma, \Pi_N), \tag{3.4}$$

where $\max_{\pi \in \Pi}$ is taken over the pure strategies of the underlying game.

PCS Score: Unfortunately, in the multi-population case, a PBR-Score of 0 does not necessarily imply α-partial convergence, as we show later in Proposition 9, the demonstration of which shows an example with PBR-Score of 0, and yet a non-convergence to the true SSCC of the game. We thus introduce a further measure, PCS-Score, defined by

$$\text{PCS-Score} = \frac{\#\text{ of } \alpha\text{-PSRO strategy profiles in the underlying game's SSCCs}}{\#\text{ of } \alpha\text{-PSRO strategy profiles in meta-SSCCs}}$$

which assesses the quality of the α-PSRO population by measuring the percentage of *truly* optimal strategies among the ones that have currently been identified as optimal by PSRO.

These metrics defined, we now turn to the question of practical computation.

Computing α-Conv and PCS-Score: Algorithms 9 and 10 provide pseudocode to compute PBR and PBR-Score in simple games - note that they compute the multipopulation version of PBR.

PCS-Score is computed by pre-computing the full game's SSCC, and computing the proportion of currently selected strategies in the empirical game that also belongs to the full game's SSCC. It is therefore only exactly computable in small games.

Note that the PBR-Score and PCS-Score are useful measures for assessing the quality of convergence in our examples, in a manner analogous to NashConv. The computation of these scores is, however, not tractable in general games. Notably, this is also the case for NashConv (as it requires computation of player-wise best responses, which can be problematic even in moderately-sized games). Despite this, these scores remain a useful way to empirically verify the convergence characteristics in small games where they can be tractably computed.

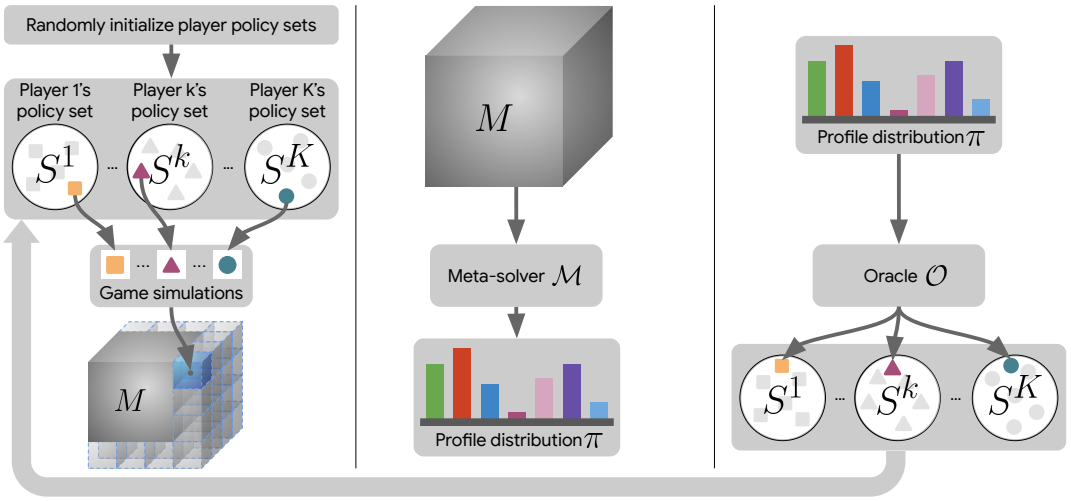
Algorithm 9 PBR Score(Strategy π, Payoff Tensor, Current Player Id, Joint Strategies, Joint Strategy Probability)

New strategy score = 0
for Joint strategy J, Joint probability P in Joint Strategies, Joint Strategy Probability **do**
  New strategy = J
  New strategy[Current Player Id] = π
  New strategy payoff = Payoff Tensor[New Strategy]
  Old strategy payoff = Payoff Tensor[J]
  New strategy score += P * (New Strategy Payoff > Old Strategy Payoff)
end for
Return New strategy score

Algorithm 10 PBR(Payoff Tensor list LM, Joint Strategies per player PJ, Alpharank Probability per Joint Strategy PA, Current Player)

$\max_{PBR} = 0$
$\max_{strat} = None$
for Strategy π available to Current Player among all possible strategies **do**
  score = PBR Score(π, LM[Current Player Id], Current Player Id, PJ, PA)
  if score > $\max_{PBR}$ **then**
    $\max_{PBR}$ = score
    $\max_{strat} = \pi$
  end if
end for
Return $\max_{PBR}, \max_{strat}$

We have defined computable metrics to characterize convergence. Armed with this knowledge, we can now provide answers to our current central question: Does α-PSRO converge to α-Rank SSCCs?



(a) Complete: compute missing payoff tensor $\boldsymbol{M}$ entries via game simulations.

(b) Solve: given the updated payoff tensor $\boldsymbol{M}$, calculate meta-strategy σ via meta-solver $\mathcal{M}$.

(c) Expand: append a new policy to each player's policy space using the oracle $\mathcal{O}$.

Figure 3.4: Overview of PSRO($\mathcal{M}$, $\mathcal{O}$) algorithm phases.

Algorithm 11 PSRO($\mathcal{M}$, $\mathcal{O}$)

1: Initialize the players' policy set $\Pi = \prod_k \Pi_k$ via random policies
2: **for** iteration $\in \{1, 2, \cdots\}$ **do**
3: Update payoff tensor M for new policy profiles in Π via game simulations {Figure 3.4a}
4: Compute the meta-strategy σ using meta-solver $\mathcal{M}(M)$ {Figure 3.4b}
5: Expand the policy space for each player $k \in \mathcal{N}$ via $\Pi_k \leftarrow \Pi_k \cup \mathcal{O}^k(\sigma)$ {Figure 3.4c}
6: **end for**

3.1.3 α-PSRO: Theory, Practice, and Connections to Nash

We study, in this section, the theoretical and practical properties of PSRO(α-Rank, PBR), or α-PSRO for brevity. We first start with a theoretical convergence study, to then move on to classes of games for which we can relate the PBR objective with RL, to finally explore relationships between α-Rank and classical equilibria.

Theoretical properties

We consider that α-PSRO has converged if no new strategy has been returned by PBR for any player at the end of an iteration. We note that converging towards an α-Rank-optimal strategic cycle is not necessarily equivalent to converging towards the full cycle - some strategic subcycles may be uncaptured by the algorithm -, and introduce a new definition of convergence to capture this.

Definition 10. A PSRO algorithm is said to converge **α-fully** (resp., **α-partially**) to an SSCC of the underlying game if its strategy population contains the full SSCC (resp., a sub-cycle of the SSCC, denoted a 'sub-SSCC') after convergence.

We also define the notion of **novelty-bound** oracles, which are oracles unable to return an already-discovered policy, and which return nothing when no unknown policy increases value. More formally,

Definition 11. An oracle $\mathcal{O}$ is said to be **novelty-bound** if $\mathcal{O}(\Pi_N, J, \Pi) \subseteq \Pi \setminus \Pi_N$. In case $\Pi \setminus \Pi_N = \emptyset$, the algorithm terminates.

In particular, the novelty-bound version of the PBR oracle is given by restricting the arg max appearing in eq. (3.2) to only be over strategies not already present in the population, yielding the Novelty-Bound PBR oracle:

$$\mathrm{PBR}\big(\sigma, \Pi_N\big) \subseteq \underset{\pi \in \Pi \setminus \Pi_N}{\arg\max} \sum_{\pi_i \in \Pi_N} \sigma_i \mathbb{1}\left[J_1(\pi, \pi_i) > J_2(\pi, \pi_i)\right] , \tag{3.5}$$

These definitions enable the following results for α-PSRO in the single- and multi-population cases:

Proposition 6. *If at any point the population of α-PSRO contains a member of an SSCC of the game, then α-PSRO will at least α-partially converge to that SSCC.*

Proof. Suppose that a member of one of the underlying game's SSCCs appears in the α-PSRO population. This member will induce its own meta-SSCC in the meta-game's response graph. At least one of the members of the underlying game's corresponding SSCC will thus always have positive probability under the α-Rank distribution for the meta-game, and the PBR oracle for this meta-SSCC will always return a member of the underlying game's SSCC.

If the PBR oracle returns a member of the underlying SSCC already in the PSRO population, we claim that the corresponding meta-SSCC already contains a cycle of the underlying SSCC, and has thus α-partially converged - to that cycle. To see this, note that if the meta-SSCC does not contain a cycle, it must be a singleton. Either this singleton is equal to the full SSCC of the underlying game (in which we have α-fully converged), or it is not, in which case the PBR oracle must return a new strategy from the underlying SSCC, contradicting our assumption that it has terminated. □

Proposition 6 states that, if the α-PSRO population somehow encounters a member of an SSCC, it will at least capture an optimal cycle thereof. However, whether such an encounter is possible has not yet been proven, and, in general, requires one more condition on the oracle, novelty-boundedness, as we see below:

Proposition 7. *If we constrain the PBR oracle used in α-PSRO to be novelty-bound, then α-PSRO will α-fully converge to at least one SSCC of the game.*

Proof. Suppose that α-PSRO has converged, and consider a meta-SSCC. Since α-PSRO has converged, it follows that each strategy profile of the meta-SSCC is an element of an SSCC of the underlying game. Any strategy profile in this SSCC which is not in the meta-SSCC will obtain a positive value for the PBR objective, and since α-PSRO has converged, there can be no such strategy profile. Thus, the meta-SSCC contains every strategy profile contained within the corresponding SSCC of the underlying game, and therefore conclude that α-PSRO α-fully converges to an SSCC of the underlying game. □

Provided our oracle is novelty-bound, then α-PSRO will always converge towards an SSCC of the game.

Stronger guarantees exist for two-players symmetric (i.e., single-population) games, which do not require novelty-boundedness for convergence towards an SSCC.

Proposition 8. *In the single-population case, there is a unique SSCC, and single-population α-PSRO always converges α-partially to it.*

Proof. The uniqueness of the SSCC follows from the fact that in the single-population case, the response graph is fully-connected. Suppose at termination of α-PSRO, the α-PSRO population contains no strategy within the SSCC, and let s be a strategy in the SSCC. We claim that s attains a higher value for the objective defining the PBR oracle than any strategy in the α-PSRO

population, which contradicts the fact that α-PSRO has terminated. To complete this argument, we note that by virtue of s being in the SSCC, we have $J_1(s, s') > J_1(s', s)$ for all s' outside the SSCC, and in particular for all $s' \in S$, thus the PBR objective for s is 1. In contrast, for any $s_i \in S$, the PBR objective for s_i is upper-bounded by $1 - \sigma_i$. If $\sigma_i > 0$, then this shows s_i is not selected by the oracle, since the objective value is lower than that of s. If $\sigma_i = 0$, then the objective value for s_i is 0, and so an SSCC member will always have a maximal PBR score of 1 against a population not composed of any SSCC member, and all members of that population have < 1 PBR scores. Consequently, single-population α-PSRO cannot terminate before it has encountered an SSCC member. By proposition 6, the proposition is therefore proven. □

One can wonder whether the novelty-boundedness condition is actually necessary for multipopulation α-PSRO to converge. Proposition 9 unfortunately shows that without novelty-boundedness, there exist games where α-PSRO will not converge to any SSCC of the game. These have primarily to do with the difficulty of joint exploration when several players have different rewards and incentives to explore the policy space.

Proposition 9. *(Multi-population) Without a novelty-bound oracle, there exist games for which α-PSRO does not converge α-partially to any SSCC.*

Proof. We exhibit a specific counterexample to the claim. Consider the three-player, three-strategy game with response graph illustrated in fig. 3.5a; note that we do not enumerate all strategy profiles not appearing in the SSCC for space and clarity reasons. The sequence of updates undertaken by α-PSRO in this game is illustrated in figs. 3.5b to 3.5f; whilst the singleton strategy profile $(3, 2, 3)$ forms the unique SSCC for this game, α-PSRO terminates before reaching it, which concludes the proof. The steps taken by the algorithm are described below; again, we do not enumerate all strategy profiles not appearing in the SSCC for space and clarity reasons.

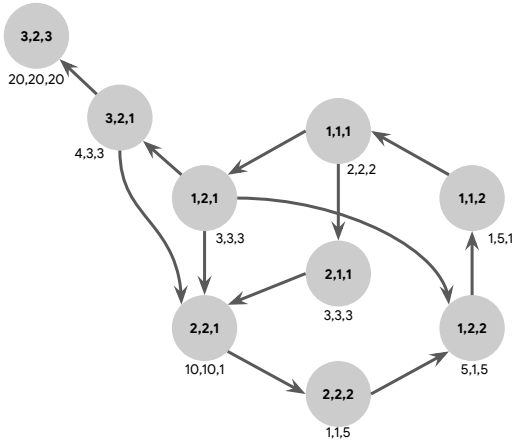
1. Begin with strategies $[[2], [1], [1]]$ in the α-PSRO population (Player 1 only has access to strategy 2, Players 2 and 3 only have access to strategy 1)
2. The PBR to (2,1,1) for player 2 is 2, and no other player has a PBR on this round. We add 2 to the strategy space of player 2, which changes the space of available joint strategies to $[(2, 1, 1), (2, 2, 1)]$.
3. α-Rank puts all its mass on (2,2,1). The PBR to (2,2,1) for player 3 is 2, and no other player has a PBR on this round. We add strategy 2 to player 3's strategy space, which changes the space of available joint strategies to $[(2, 1, 1), (2, 2, 1), (2, 2, 2)]$.
4. α-Rank puts all its mass on (2,2,2). The PBR to (2,2,2) for player 1 is 1, and no other player has a PBR on this round. We add strategy 1 to player 1's strategy space, which changes the space of available joint strategies to $[(1, 1, 1), (1, 1, 2), (1, 2, 1), (1, 2, 2), (2, 1, 1), (2, 2, 1), (2, 2, 2)]$.
5. Define σ as the α-Rank probabilities of the meta-game. Player 1 playing strategy 2 has a PBR score of $\sigma((1, 1, 1)) + \sigma((1, 2, 1))$, and the same player playing strategy 3 has a PBR score of $\sigma((1, 2, 1))$, which is lower than the PBR Score of playing strategy 2. No other player has a valid PBR for this round, and therefore, α-PSRO terminates.

□

In the above example, pictured in fig. 3.5, a relatively weak joint strategy (Strategy (3,2,1)) bars agents from finding the optimal joint strategy of the game (Strategy (3,2,3)) : getting to this joint strategy requires coordinated changes between agents, and is therefore closely related to the common problem of *Action/Equilibrium Shadowing* mentioned in [113].

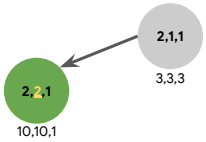
Intuitively, the lack of convergence without a novelty-bound oracle can occur due to intransitivities in the game (i.e., cycles in the game can trap the oracle).

An example demonstrating this issue is shown in fig. 3.5, with an accompanying step-by-step walkthrough in the proof of Proposition 9. Specifically, SSCCs may be hidden by "intermediate" strategies that, while not receiving as high a payoff as current population-pool members, can actually lead to well-performing strategies outside the population. As these "intermediate" strategies are avoided, SSCCs are consequently not found. Note also that this is related to the common problem of action/equilibrium shadowing, as detailed in Matignon et al. [113].



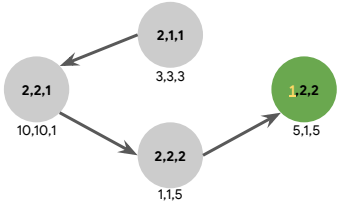
(a) Full game response graph.

(b) α-PSRO Step 1.

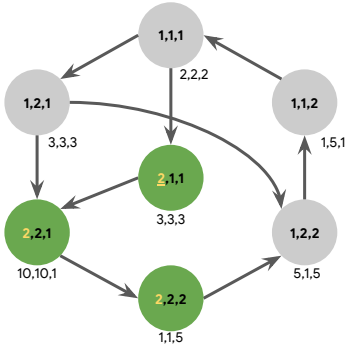


(c) α-PSRO Step 2.

(d) α-PSRO Step 3.



(e) α-PSRO Step 4.



(f) α-PSRO Step 5.

Figure 3.5: The three-player, three-strategy game serving as a counterexample in the proof of Proposition 9. Strategy profiles are illustrated by gray circles, with payoffs listed beneath. All strategy profiles not pictured are assumed to be dominated, and are therefore irrelevant in determining whether α-PSRO reaches an SSCC for this game. Highlighted in green are the joint strategies which include the PBR output, with the exact strategy highlighted in yellow.

In section 3.1.4, we further investigate convergence behavior beyond the conditions studied above. In practice, we demonstrate that despite the negative result of Proposition 9, α-PSRO does significantly increase the probability of converging to an SSCC, in contrast to PSRO(Nash, BR). Overall, we have shown that for general-sum multi-player games, it is possible to give theoretical guarantees for a version of PSRO driven by α-Rank in several circumstances. By contrast, using exact NE in PSRO is intractable in general. In prior work [98], this motivated the use of approximate Nash solvers generally based on the simulation of dynamical systems or regret minimization algorithms, both of which generally require specification of several hyperparameters (e.g., simulation iterations, window sizes for computing time-average policies, and entropy-injection rates), and a greater computational burden than α-Rank to carry out the simulation in the first place.

We now know which conditions are required for α-PSRO to converge. However, these conditions rely on being able to compute the PBR objective, and novelty-boundedness, yet the PBR oracle is difficult to compute in complex games. The next section explores classes of games where one can use RL techniques to compute the PBR oracle.

PBR Oracle and Relationship with RL:

Recall from Algorithm 5 that the BR oracle inherently solves a single-player optimization problem, permitting use of a single-agent RL algorithm as a BR approximator, a property important for practical use.

As we can notice in section 3.1.1, however, there exist games where the BR and PBR objectives are seemingly incompatible, preventing the use of standard, reward-maximizing RL agents for PBR approximation.

While exact PBR is computable in small-scale (e.g., normal-form) games, we consider more general games classes where PBR can also be approximated using RL, *i.e.* games where the RL objective is *compatible* with the PBR objective. More formally,

Definition 12 (Compatibility)**.** Objective $\mathcal{A}$ is said to be *compatible* with objective $\mathcal{B}$ if any solution to $\mathcal{A}$ is a solution to $\mathcal{B}$.

We have the following compatibility results:

Proposition 10. *A constant-sum game is denoted as* ***win-loss*** *if $J_k(s) \in \{0, 1\}$ for all $k \in [K]$ and $s \in S$. BR is compatible with PBR in win-loss games in the two-player single-population case.*

Proof. We overload the best-response objective as follows:

$$J_1(\pi, \sigma) = \sum_{\pi' \in \Pi} \sigma(\pi') J_1(\pi, \pi') \tag{3.6}$$

$$= \sum_{\pi' \in \Pi} \sigma(\pi') \mathbb{1}[J_1(\pi, \pi') > J_2(\pi, \pi')] \,. \tag{3.7}$$

Noting that the final line is the single-population PBR objective, the proof is concluded. □

Proposition 11. *A symmetric two-player game is denoted* ***payoff-monotonic*** *if there exists a function $f : \Pi \to \mathbb{R}$ and a non-decreasing function $\sigma : \mathbb{R} \to \mathbb{R}$ such that $J_1(\pi, \pi') = \sigma(f(\pi) - f(\pi'))$. BR is compatible with PBR in payoff-monotonic games in the single-population case.*

Proof. Rewriting the objectives given that the game is payoff-monotonic, we have that the value-based objective becomes

$$\sum_{k=1}^{K} \sigma_k J_1(\pi, \pi_k) = \sum_{k=1}^{K} \sigma_k \sigma(f(\pi) - f(\pi_k)) \,, \tag{3.8}$$

where we pay close attention to the overloading of σ: σ_k refers to the probability of playing π_k, whereas $\sigma(x)$ refers to the monotonic function from $\mathbb{R}$ to $\mathbb{R}$.

Given the fact that the only condition we have on σ is its non-decreasing character, this objective does not reduce to maximizing $f(\pi)$ in the general case.

The objective for PBR is

$$\sum_{k=1}^{K} \sigma_k \mathbb{1}[J_1(\pi, \pi_k) > J_2(\pi, \pi_k)] = \sum_{k=1}^{K} \sigma_k \mathbb{1}[\sigma(f(\pi) - f(\pi_k)) > \sigma(f(\pi_k) - f(\pi))] \tag{3.9}$$

Since σ is non-decreasing,

$$\sigma(f(\pi) - f(\pi_k)) > \sigma(f(\pi_k) - f(\pi)) \;\Rightarrow\; f(\pi) > f(\pi_k)$$

and conversely,

$$f(\pi) > f(\pi_k) \;\Rightarrow\; \sigma(f(\pi) - f(\pi_k)) \geq \sigma(f(\pi_k) - f(\pi))$$

Without loss of generality, we reorder the strategies such that if $i < k$, $f(\pi_i) \leq f(\pi_k)$.

Let π_v maximize the value objective. Therefore, by payoff-monotonicity, π_v maximizes $\sigma(f(\pi) - f(\pi_K))$. Three possibilities then ensue.

If there exists π such that

$$\sigma(f(\pi) - f(\pi_K)) > \sigma(f(\pi_K) - f(\pi))$$

then

$$\sigma(f(\pi_v) - f(\pi_K)) > \sigma(f(\pi_K) - f(\pi_v))$$

since π_v maximizes $\sigma(f(\pi) - f(\pi_K))$ and σ is non-decreasing. Consequently π_v maximizes the PBR objective. Indeed, let us remark that for all $k \leq K$, we have that

$$\sigma(f(\pi_v) - f(\pi_k)) > \sigma(f(\pi_k) - f(\pi_v))$$

since

$$\sigma(f(\pi_v) - f(\pi_k)) \geq \sigma(f(\pi_v) - f(\pi_K)) > \sigma(f(\pi_K) - f(\pi_v)) \geq \sigma(f(\pi_k) - f(\pi_v)).$$

Otherwise, if there does not exist any policy π such that $\sigma(f(\pi) - f(\pi_K)) > \sigma(f(\pi_K) - f(\pi))$, that is, for all π,

$$\sigma(f(\pi) - f(\pi_K)) \leq \sigma(f(\pi_K) - f(\pi)).$$

Since π_K is a possible solution to the value objective,

$$\sigma(f(\pi_v) - f(\pi_K)) = \sigma(f(\pi_K) - f(\pi_v)).$$

Let n be the integer such that

$$\pi_n = \arg\max\{f(\pi_k), \pi_k \in \text{ Population } \mid \exists \pi \text{ s.t. } \sigma(f(\pi) - f(\pi_k)) > \sigma(f(\pi_k) - f(\pi))\}.$$

If π_n **exists**, then we have that for all π_i such that $f(\pi_i) > f(\pi_n)$,

$$\sigma(f(\pi_v) - f(\pi_i)) = \sigma(f(\pi_i) - f(\pi_v)).$$

The PBR objective is

$$\sum_{k=1}^{K} \sigma_k \mathbb{1}[\sigma(f(\pi) - f(\pi_k)) > \sigma(f(\pi_k) - f(s))],$$

which, according to our assumptions, is equivalent to

$$\sum_{k=1}^{n} \sigma_k \mathbb{1}[\sigma(f(\pi) - f(\pi_k)) > \sigma(f(\pi_k) - f(\pi))].$$

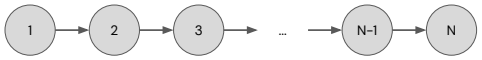
We know that for all $i \leq n$, $\sigma(f(\pi_v) - f(\pi_i)) > \sigma(f(\pi_i) - f(\pi_v))$, and therefore, π_v maximizes the PBR objective.

Finally, if π_n does not exist, then any policy is solution to the PBR objective, and therefore π_v is. □

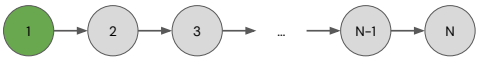
Payoff-monotonic games include real-world games such as Games of Skills [47]. Since the games derived from Elo ratings are payoff-monotonic, we can use Elo ratings as a concrete implementation of a payoff-monotonic game: in this case, σ is a sigmoid, and f represents the Elo rating of a given player.

A toy example illustrating Proposition 11 is shown in fig. 3.6. The setting is that of a payoff-monotonic game where every strategy is assigned a number. Strategies are then dominated by all strategies with higher number than theirs. We compute BR and PBR on an initial population composed of one strategy that we choose to be dominated by every other strategy. Any strategy dominating the current population is a valid solution for PBR, as represented in fig. 3.6c; whereas, if we consider that the game is payoff-monotonic with σ a strictly increasing function, only one strategy maximizes Best Response, strategy N – and it is thus the only solution of BR, as shown in fig. 3.6d.

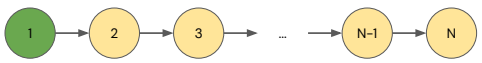
As we can see, the solution of BR is part of the possible solutions of PBR, demonstrating the result of proposition 11: BR is compatible with PBR in payoff-monotonic games.



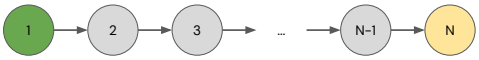
(a) Full game response graph of a payoff-monotonic game.



(b) Starting population (green).



(c) Current population (green) and possible solutions of PBR (yellow)



(d) Current population (green) and possible solution of BR (yellow).

Figure 3.6: Toy example of compatibility between PBR and BR: The solution returned by BR is one of the possible solutions of PBR.

Finally, we wonder about connections between PBR and the field of Preference-Based RL [187]. Preference-Based RL aims at maximizing, not a reward, but the preferences of a user: the only information received by the algorithm purports to a preference from the user, *i.e.* given two policies, the user selects the better one. The algorithm then attempts to find a strategy which dominates all other strategies for this induced (partial) order. This is also the objective of PBR, it attempts to find a strategy which dominates all others for the partial order defined by winning. We demonstrate that, under certain conditions, there are strong connections between the PBR objective defined above and the broader field of preference-based RL.

Proposition 12. *Consider symmetric win-loss games where outcomes between deterministic strategies are deterministic. A preference-based RL agent (i.e., an agent aiming to maximize its probability of winning against a distribution σ of strategies $\{\pi_1, \dots, \pi_N\}$) optimizes exactly the PBR objective eq.* (3.1).

Proof. Commencing with the above preference-based RL objective, we calculate as follows,

$$\arg\max_{\pi} \mathbb{P}\left(\pi \text{ beats } \sum_{i=1}^{N} \sigma_i \pi_i\right) = \arg\max_{\pi} \sum_{i=1}^{N} \sigma_i \mathbb{P}(\pi \text{ beats } \pi_i) \tag{3.10}$$

$$= \arg\max_{\pi} \sum_{i=1}^{N} \sigma_i \mathbb{1}[\pi \text{ receives a positive expected payoff against } \pi_i] \tag{3.11}$$

with the final equality whenever game outcomes between two deterministic strategies are deterministic. Note that this is precisely the PBR objective eq. (3.1). □

Given this insight, we believe an important subject of future work will involve the use of preference-based RL algorithms in implementing the PBR oracle for more general classes of games.

α-Rank and classical equilibria

We conclude this section with some indicative results of the relationship between α-Rank and NE.

We first explore a few cases where Nash equilibria and α-Rank have non-zero intersection:

Proposition 13. *For symmetric two-player zero-sum games where off-diagonal payoffs have equal magnitude, all NE have support contained within that of the single-population α-Rank distribution.*

Proof. In the single-population case, the support of the α-Rank distribution is simply the (unique) sink strongly-connected component of the response graph (uniqueness follows from the fact that the response graph, viewed as an undirected graph, is fully-connected). We will now argue that for a strategy π^* in the sink strongly-connected component and a strategy π_o outside the sink strongly-connected component, we have

$$\sum_{\pi \in \Pi} \sigma(\pi) J_1(\pi^*, \pi) > \sum_{\pi \in \Pi} \sigma(\pi) J_1(\pi_o, \pi)\,, \tag{3.12}$$

This inequality states that when an opponent plays according to σ, the expected payoff to the row player is greater if they defect to π^* whenever they would have played π_o. This implies that if a supposed symmetric Nash equilibrium contains a strategy π_o outside the sink strongly-connected component in its support, then it could receive higher reward by playing π^* instead, which contradicts the fact that it is an NE. We show eq. (3.12) by proving a stronger result — namely, that π^* dominates π_o as strategies. Firstly, since π^* is the sink strongly-connected component and π_o is not, π^* beats π_o, and so $J_1(\pi^*, \pi_o) > J_1(\pi^*, \pi^*) = J_1(\pi_o, \pi_o) > J_1(\pi_o, \pi^*)$. Next, if $\pi_i \notin \{\pi^*, \pi_o\}$ is in the sink strongly-connected component, then π_i beats π_o, and so $J_1(\pi^*, \pi_i) > J_1(\pi_o, \pi_i)$ if π^* beats π_i, and $J_1(\pi^*, \pi_i) = J_1(\pi_o, \pi_i)$ otherwise. Finally, if $\pi_i \neq \pi^*, \pi_o$ is not in the sink strongly-connected component, then $J_1(\pi^*, \pi_i) = J_1(\pi_o, \pi_i)$ is π_o beats π_i, and $J_1(\pi^*, \pi_i) > J_1(\pi_o, \pi_i)$ otherwise. Thus, eq. (3.12) is proven, and the result follows. □

Proposition 14. *In all symmetric two-player zero-sum games, there exists an NE with support contained within that of the α-Rank distribution.*

Proof. Consider the restriction of the game to the strategies contained in the sink strongly-connected component of the original game. Let σ be an NE for this restricted game, and consider this as a distribution over all strategies in the original game (putting 0 mass on strategies outside the sink component). We argue that this is an NE for the full game, and the statement follows. To see this, note that since any strategy outside the sink strongly-connected component receives a non-positive payoff when playing against a strategy in the sink strongly-connected component, and that for at least one strategy in the sink strongly-connected component, this payoff is negative. Considering the payoffs available to the row player when the column player plays according to π, we observe that the expected payoff for any strategy outside the sink strongly-connected component is negative, since every strategy in the sink strongly-connected component beats the strategy outside the component. The payoff when defecting to a strategy in the sink strongly-connected component must be non-positive, since π is an NE for the restricted game. □

For more general games, the link between α-Rank and Nash equilibria will likely require a more complex description. Indeed, we provide below a counterexample where α-Rank and Nash supports are fully disjoint.

The Game of Chicken Consider the Game of Chicken in the multipopulation case.

This game has three Nash equilibria: Two pure, (D,C) and (C,D), and one mixed, where the population plays Dare with probability $\frac{1}{3}$. Nevertheless, α-rank only puts weight on (C,D) and (D,C), effectively not putting weight on the full mixed-Nash support.

We also explore the relationship between α-Rank's support and coarse correlated equilibria's support, finding the following counterexample where both supports are fully disjoint:

	D	C
D	$(0, 0)$	$(7, 2)$
C	$(2, 7)$	$(6, 6)$

Table 3.4: Game of Chicken payoff table

Prisoner's Dilemma Consider the Prisoner's Dilemma in the multi-population case.

	D	C
D	$(0, 0)$	$(3, -1)$
C	$(-1, 3)$	$(2, 2)$

Table 3.5: Prisoner's Dilemma payoff table

This game has coarse correlated equilibria that include (C,D), (D,C) and (C,C) in their support; nevertheless, α-Rank only puts weight on (D,D), effectively being fully disjoint from the support of the coarse correlated equilibria.

Now that we have fully described our algorithm and its asymptotic behavior, we finally turn to its empirical behavior.

3.1.4 Evaluation

We conduct evaluations on games of increasing complexity, extending beyond prior PSRO applications that have focused on two-player zero-sum games.

We first describe experimental procedures.

Experimental procedures

We run experiments on two main domains. On the one hand, Kuhn and Leduc poker, two simplified versions of Poker; on the other hand, randomly-generated normal-form games. We will describe all these games and how we solved them in this section.

Kuhn and Leduc Poker: K-player Kuhn poker is played with a deck of $K + 1$ cards. Each player starts with 2 chips and 1 face-down card, and antes 1 chip to play. Players either bet (raise/call) or fold iteratively, until each player is either in (has contributed equally to the pot) or has folded. Amongst the remaining players, the one with the highest-ranked card wins the pot.

Leduc Poker, in comparison, has a significantly larger state space. Players in Leduc have unlimited chips, receive 1 face-down card, ante 1 chip to play, with subsequent bets limited to 2 and 4 chips in rounds 1 and 2. A maximum of two raises are allowed in each round, and a public card is revealed before the second round.

The code backend for the Poker experiments used OpenSpiel [99]. Specifically, we used OpenSpiel's Kuhn and Leduc poker implementations, and exact best responses were computed by traversing the game tree following Algorithm 3. 100 game simulations were used to estimate the payoff matrix for each possible strategy pair.

Although the underlying Kuhn and Leduc poker games are stochastic (due to random initial card deals), the associated meta-games are essentially deterministic (as, given enough game simulations, the mean payoffs are fixed). The subsequent PSRO updates are, thus, also deterministic.

Despite this, we report averages over 2 runs per PSRO metasolver, primarily to capture stochasticity due to differences in machine-specific rounding errors that occur due to the distributed computational platforms we run these experiments on.

For experiments involving α-Rank, we conduct a full sweep over the ranking-intensity parameter, α, following each iteration of α-PSRO. We implemented a version of α-Rank (building on the OpenSpiel implementation that used a sparse representation for the underlying transition matrix, enabling scaling-up to the large-scale NFG results presented in the experiments.

For experiments involving the projected replicator dynamics (PRD), we used uniformly-initialized meta-distributions, running PRD for $5e4$ iterations, using a step-size of $\texttt{dt} = 1e-3$, and exploration parameter $\gamma = 1e-10$. Time-averaged distributions were computed over the entire trajectory.

Normal Form Games Generation Algorithms 12 to 14 provide an overview of the procedure we use to randomly-generate normal-form games for the oracle comparisons visualized in fig. 3.7.

Algorithm 12 GenerateTransitive(Actions, Players, $\text{mean}_{\text{value}} = [0.0, 1.0]$, $\text{mean}_{\text{probability}} = [0.5, 0.5]$, $\text{var} = 0.1$)

$\mathcal{T} = []$
for Player k **do**
  Initialize $f_k = [0] * \text{Actions}$
  for Action $a \leq$ Actions **do**
    Randomly sample mean μ from $\text{mean}_{\text{value}}$ according to $\text{mean}_{\text{probability}}$
    $f_k[a] \sim \mathcal{N}(\mu, \text{var})$
  end for
end for
for Player k **do**
  $\mathcal{T}[k] = f_k - \frac{1}{|\text{Players}|-1} \sum_{i \neq k} f_i$
end for
Return $\mathcal{T}$

Algorithm 13 GenerateCyclic(Actions, Players, $\text{var} = 0.4$)

$\mathcal{C} = []$
for Player k **do**
  Initialize $C[k] \sim \mathcal{N}(0, \text{var})$, $\text{Shape}(C[k]) = (\text{Actions}_{\text{First Player}}, \ldots, \text{Actions}_{\text{Last Player}})$
end for
for Player k **do**
  $\text{Sum} = \sum_{\text{Actions } a_i \text{ of all player } i \neq k} C[k][a_1, \ldots, a_{k-1}, :, a_{k+1}, ...]$
  $\text{Shape}(\text{Sum}) = (1, \ldots, 1, \text{Actions}_{\text{Player } k}, 1, \ldots, 1)$
  $C[k] = C[k] - \text{Sum}$
end for
Return $\mathcal{C}$

Algorithm 14 General Normal Form Games Generation(Actions, Players)

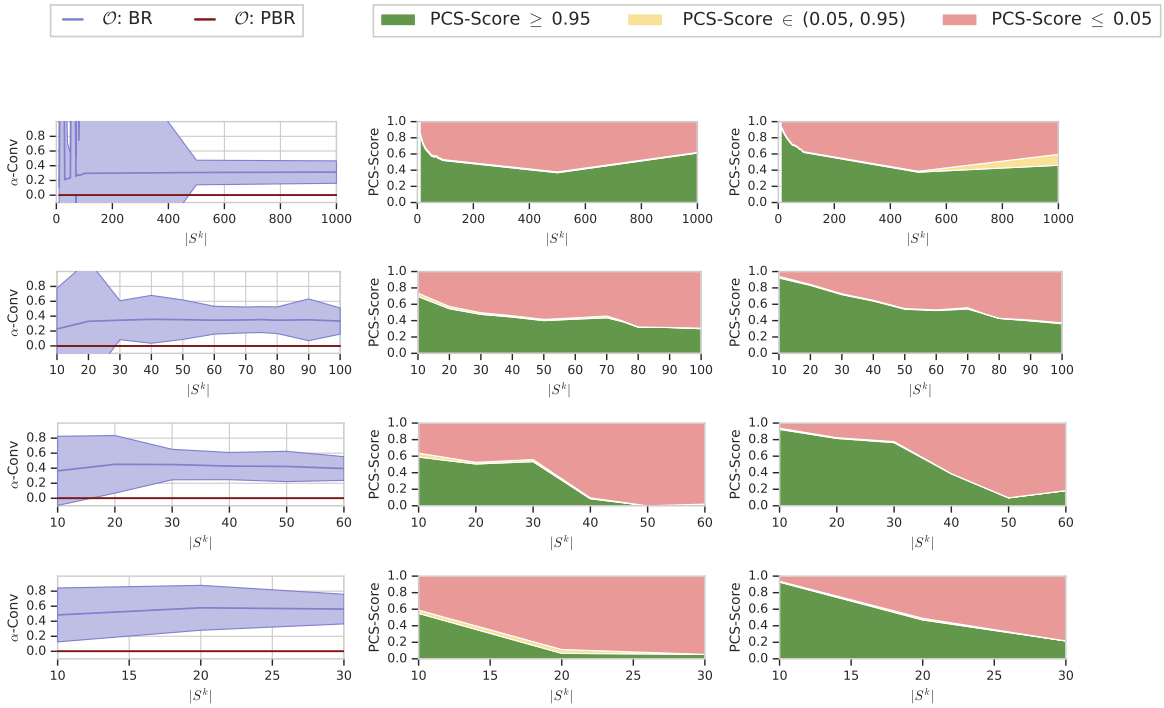
Generate matrix lists $\mathcal{T}$ = GenerateTransitive(Actions, Players), $\mathcal{C}$ = GenerateCyclic(Actions, Players)
Return $[\mathcal{T}[k] + \mathcal{C}[k]$ for Player k$]$

These implementation details specified, we can go straight to the results.

Experimental results

We first start to analyze the algorithm's behavior on normal-form games to evaluate oracle differences, to then shift towards the more complex poker games to evaluate meta-solver differences.

Oracle comparisons



(a) α-Conv (b) PCS-Score for BR (c) PCS-Score for PBR

Figure 3.7: Oracle comparisons for randomly-generated normal-form games with varying player strategy space sizes $|S^k|$. The rows, in order, correspond to 2- to 5-player games.

We evaluate the performance of the BR and PBR oracles in games where PBR can be exactly computed. We consider randomly generated, K-player, general-sum games with increasing strategy space sizes, $|S^k|$. Figure 3.7 reports these results for the 2- to 5-player instances. The asymmetric nature of these games, in combination with the number of players and strategies involved, makes them inherently very-large in scale. For example, the largest game we consider involves 5 players with 30 strategies each, making for a total of more than 24 million strategy profiles in total. For each combination of K and $|S^k|$, we generate 10^6 random games. We conduct 10 trials per game, in each trial running the BR and PBR oracles starting from a random strategy in the corresponding response graph, then iteratively expanding the population space until convergence. Importantly, this implies that the starting strategy may not even be in an SSCC.

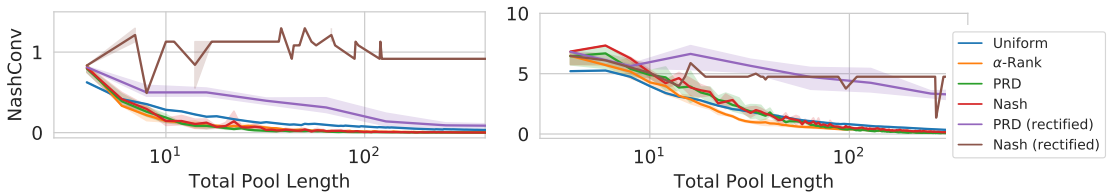
Figure 3.7 plots both α-Conv and PCS-Score for both oracles, demonstrating that PBR outperforms BR in the sense that it captures more of the game SSCCs.

The PCS-Score here is typically either (a) greater than 95%, or (b) less than 5%, and otherwise rarely between 5% to 95%.

For all values of $|S^k|$, PBR consistently discovers a larger proportion of the α-Rank support in contrast to BR, serving as useful validation of the theoretical results of section 3.1.3.

Meta-solver comparisons

We consider next the standard benchmarks of Kuhn and Leduc poker. We first consider two-player instances of these poker domains, permitting use of an exact Nash meta-solver. Figure 3.8 compares the NashConv of PSRO($\mathcal{M}$, BR) for various meta-solver $\mathcal{M}$ choices. Note that the x axis of Figure 3.8 and Figure 3.9 is the Total Pool Length (The sum of the length of each player's pool in PSRO) instead of the number of iterations of PSRO, since Rectified solvers can add more than one



(a) Kuhn poker. (b) Leduc poker.

Figure 3.8: Results for 2-player poker domains.

policy to the pool at each PSRO iteration (Possibly doubling pool size at every PSRO iteration). It is therefore more pertinent to compare exploitabilities at the same pool sizes rather than at the same number of PSRO iterations.

In Kuhn poker (fig. 3.8a), the α-Rank, Nash, and the Projected Replicator Dynamics (PRD) meta-solvers converge essentially at the same rate towards zero NashConv, in contrast to the slower rate of the Uniform meta-solver, the very slow rate of the Rectified PRD solver, and the seemingly constant NashConv of the Rectified Nash solver. Given how surprising this result seemed to be, we analyzed the dynamics in Section 3.1.4 and determined that the algorithm's stagnation is indeed due to rectified Nash, and not to an eventual implementation error or numerical problems.

As noted in Lanctot et al. [98], PSRO(Uniform, BR) corresponds to Fictitious Play [27] and is thus guaranteed to find an NE in two-player zero-sum games. Its slower convergence rate is explained by the assignment of uniform mass across all policies $s \in S$, implying that PSRO essentially wastes resources on training the oracle to beat even poor-performing strategies. While α-Rank does not seek to find an approximation of Nash, it nonetheless reduces the NashConv yielding competitive results in comparison to an exact-Nash solver in these instances. Notably, the similar performance of α-Rank and Nash serves as empirical evidence that α-Rank can be applied competitively even in the two-player zero-sum setting, while also showing great promise to be deployed in broader settings where Nash is no longer tractable.

We next consider significantly larger variants of Kuhn and Leduc Poker involving more than two players, extending beyond the reach of prior PSRO results [98]. Figure 3.9 visualizes the NashConv of PSRO using the various meta-solvers (with the exception of an exact Nash solver, due to its intractability in these instances). In all instances of Kuhn Poker, α-Rank and PRD show competitive convergence rates. In 3-player Leduc poker, however, α-Rank shows fastest convergence, with Uniform following throughout most of training and PRD eventually reaching a similar NashConv. Several key insights can be made here. First, computation of an approximate Nash via PRD involves simulation of the associated replicator dynamics, which can be chaotic [145] even in two-player two-strategy games, making it challenging to determine when PRD has suitably converged. Second, the addition of the projection step in PRD severs its connection with NE; the theoretical properties of PRD were left open in Lanctot et al. [98], leaving it without any guarantees. These limitations go beyond theoretical, manifesting in practice, e.g., in fig. 3.9d, where PRD is outperformed by even the uniform meta-solver for many iterations. Given these issues, we take a first (and informal) step towards analyzing PRD in section 3.1.4. For α-Rank, by contrast, we both establish theoretical properties, and face no simulation-related challenges as its computation involves solving of a linear system, even in the general-sum many-player case [138], thus establishing it as a favorable and general PSRO meta-solver.

MuJoCo Soccer

While the key objective of this chapter is to take a step in establishing a theoretically-grounded framework for PSRO-based training of agents in many-player settings, an exciting question regards the behaviors of the proposed α-Rank-based PSRO algorithm in complex domains where function-approximation-based policies need to be relied upon. In this section, we take a first step towards conducting this investigation in the MuJoCo soccer domain introduced in Liu et al. [105]. We remark that our results, albeit interesting, are primarily intended to lay the foundation for use of

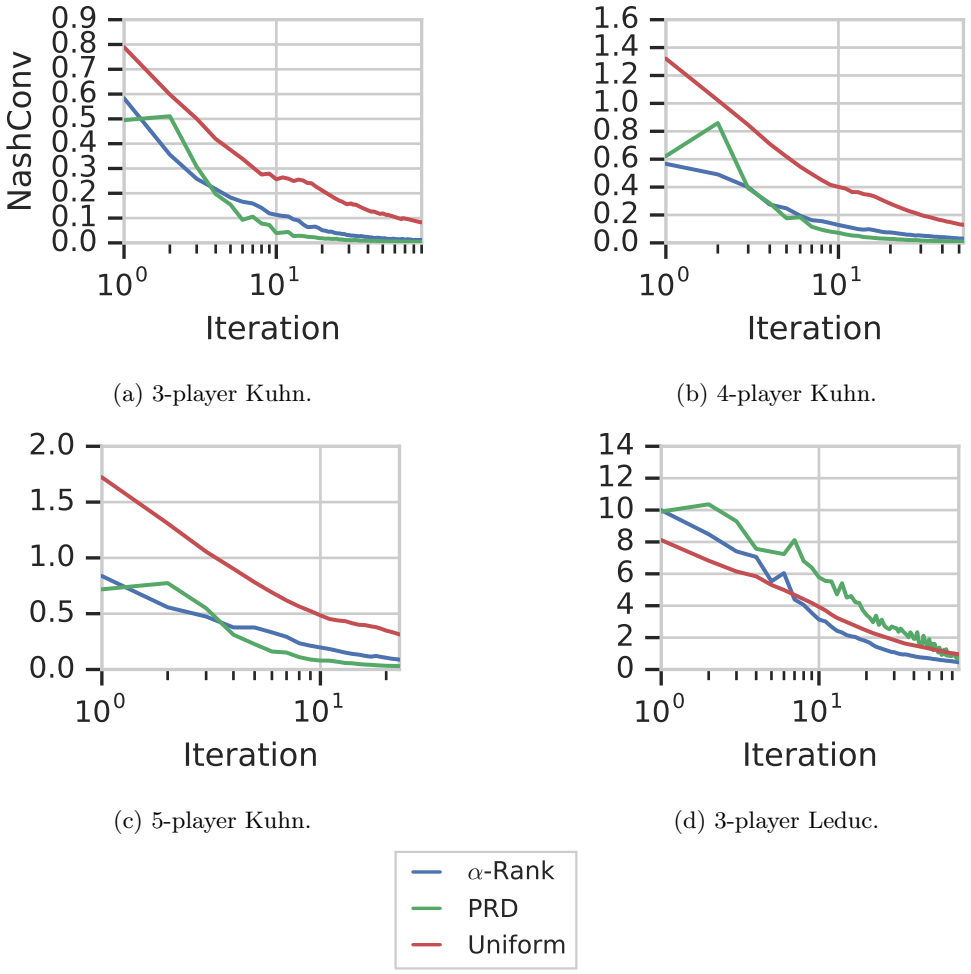


Figure 3.9: Results for poker domains with more than 2 players.

α-Rank as a meta-solver in complex many agent domains where RL agents serve as useful oracles, warranting additional research and analysis to make conclusive insights.

We conduct two sets of initial experiments. The first set of experiments compares the performance of PSRO(α-Rank, RL) against PSRO(Uniform, RL) in games of 3 vs. 3 MuJoCo soccer, and the second set compares PSRO(α-Rank, RL) against a self-play in 2 vs. 2 games.

Training Procedure

For each of the PSRO variants considered, we compose a hierarchical training procedure composed of two levels. At the low-level, which focuses on simulations of the underlying MuJoCo soccer game itself, we consider a collection of 32 reinforcement learners (which we call agents) that are all trained at the same time, as in Liu et al. [105]. We compose teams corresponding to multiple clones of agent per team (yielding homogeneous teams, in contrast to [105], which evaluates teams of heterogeneous agents) and evaluate all pairwise team match-ups. Note that this yields a 2-"player" meta-game (where each "player" is actually a team, i.e., a team-vs.-team setting), with payoffs corresponding to the average win-rates of each team when pitted against each other.

The payoff matrix is estimated by simulating matches between different teams composed of frozen policies that have been added to the pool. The number of simulations per entry is adaptive based on the empirical uncertainty observed on the pairwise match outcomes. In practice, we

observed an average of 10 to 100 simulations per entry, with fewer simulations used for meta-payoffs with higher certainty. For the final evaluation matrix reported in Figure 3.10, which was computed after the conclusion of PSRO-based training, 100 simulations were used per entry. Additionally, instead of adding one policy per PSRO iteration per player we add three (which corresponds to the 10% best RL agents).

Several additional modifications were made to standard PSRO to help with the inherently more difficult nature of Deep Reinforcement Learning training:

- Agent performance, used to choose which agents out of the 32 to add to the pool, is measured by the α-Rank-average for PSRO(α-Rank, RL) and Nash-average for PSRO(Uniform, RL) of agents in the (Agents, Pool) versus (Agents, Pool) game.
- Each oracle step in PSRO is composed of 1 billion learning steps of the agents. After each step, the top 10% of agents (the 3 best agents) are added to the pool, and training of the 32 agents continues;
- We use a 50% probability of training using self-play (the other 50% training against the distribution of the pool of agents).

Results

In the first set of experiments, we train the PSRO(α-Rank, RL) and PSRO(Uniform, RL) agents independently (i.e., the two populations never interact with one another). Following training, we compare the effective performance of these two PSRO variants by pitting their 8 best trained agents against one another, and recording the average win rates. These results are reported in Figure 3.10 for games involving teams of 3 vs. 3. It is evident from these results that PSRO(α-Rank, RL) significantly outperforms PSRO(Uniform, RL). This is clear from the colorbar on the far right of Figure 3.10, which visualizes the post-training alpharank distribution over the payoff matrix of the metagame composed of both training pipelines.

In the second set of experiments, we compare α-PSRO-based training to self-play-based training. This provides a means of gauging the performance improvement solely due to PSRO; these results are reported in fig. 3.11 for games involving teams of 2 vs. 2.

We conclude by remarking that these results, although interesting, primarily are intended to lay the foundation for use of α-Rank as a meta-solver in complex many-agent domains where RL agents serve as useful oracles; additionally, more extensive research and analysis is necessary to make these results conclusive in domains such as MuJoCo soccer.

Notes on Rectified Nash performance

This section provides additional insights into the surprising Rectified Nash behavior detailed in section 3.1.4. We begin with an important disclaimer that Rectified Nash was developed solely with symmetric games in mind. As Kuhn Poker and Leduc Poker are not symmetric games, they lie beyond the theoretical scope of Rectified Nash. Nevertheless, comparing the performance of rectified and non-rectified approaches from an empirical perspective yields insights, which may be useful for future investigations that seek to potentially extend and apply rectified training approaches to more general games.

As noted earlier, the poor performance of PSRO using Rectified Nash in fig. 3.8 is initially surprising as it indicates premature convergence to a high-NashConv distribution over the players' policy pools. Investigating this further led to a counterintuitive result for the domains evaluated: Rectified Nash was, roughly speaking, not increasing the overall diversity of behavioral policies added to each player's population pool. In certain regards, it even prevented diversity from emerging.

To more concretely pinpoint the issues, we detail below the first 3 iterations of PSRO(Rectified Nash, BR) in Kuhn Poker. Payoff matrices at each PSRO iteration are included in tables 3.6a to 3.6c. For clarity, we also include the 5 best responses trained by Rectified Nash and the policies

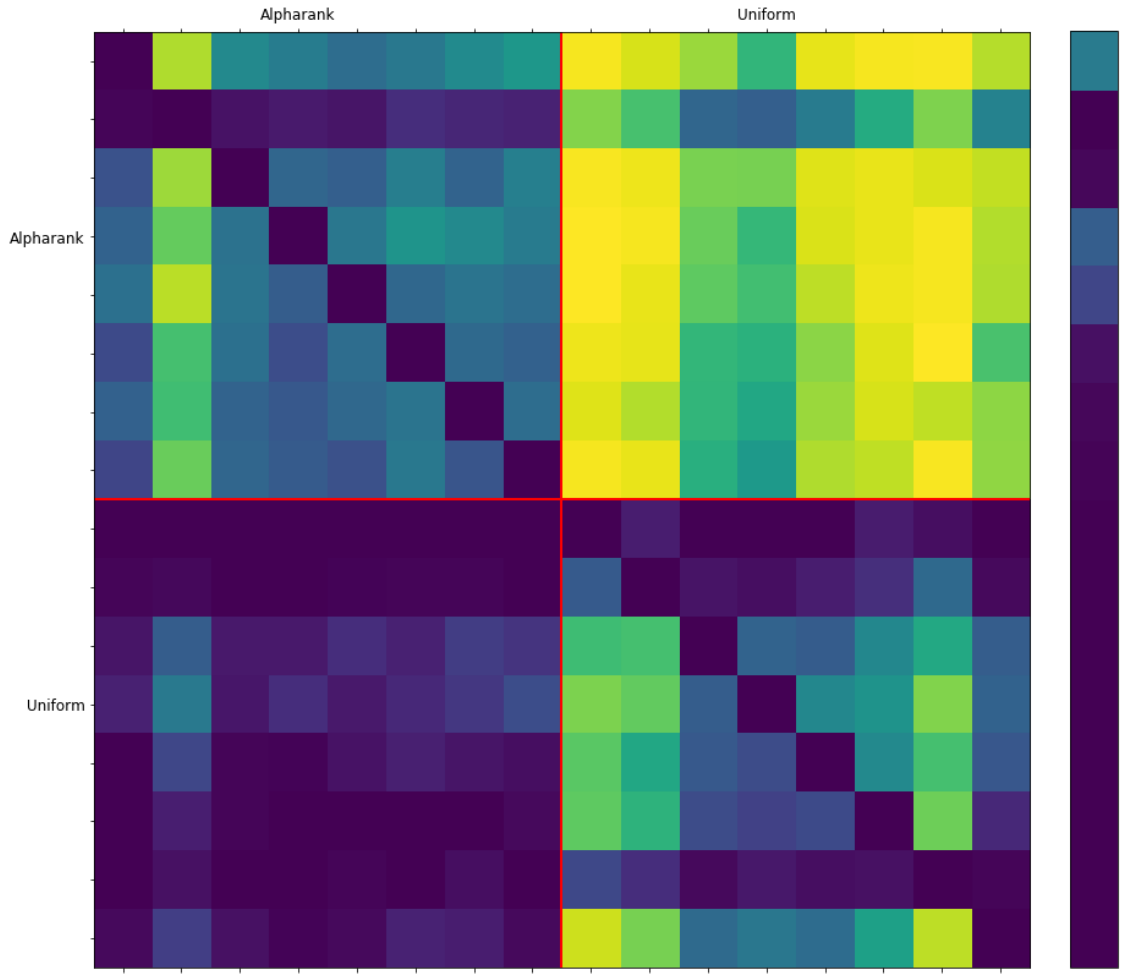


Figure 3.10: α-PSRO versus PSRO(Uniform, BR) in the MuJoCo Soccer domain. The left figure is the matrix representing the probability of winning for α-PSRO and PSRO(Uniform, BR)'s best 8 agents. The right bar represents the α-Rank distribution over the meta-game induced by these agents. Yellow represents high probabilities, dark-blue represents low probabilities. The diagonal is taken to be 0.

they were trained against, in their order of discovery: 2 policies for Player 1 (in fig. 3.13) and 3 policies for Player 2 (in fig. 3.14).

1. Iteration 0: both players start with uniform random policies.
2. Iteration 1:
    - Player 1 trains a best response against Player 2's uniform random policy; its policy set is now the original uniform policy, and the newly-computed best response.
    - Player 2 trains a best response against Player 1's uniform random policy; its policy set is now the original uniform policy, and the newly-computed best response.
    - Player 2's best response beats both of Player 1's policies.
    - Payoff values are represented in table 3.6a.
3. Iteration 2:

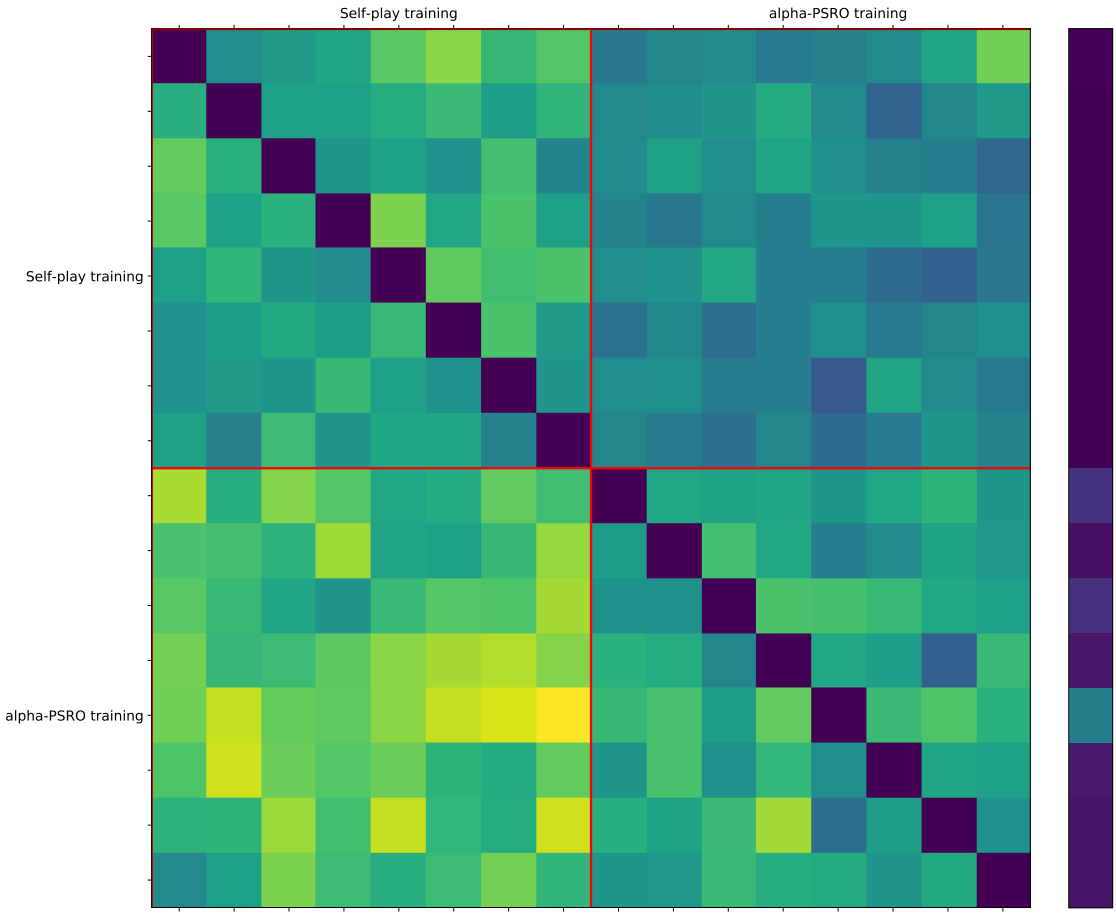


Figure 3.11: α-PSRO training pipeline vs. training pipeline without PSRO.

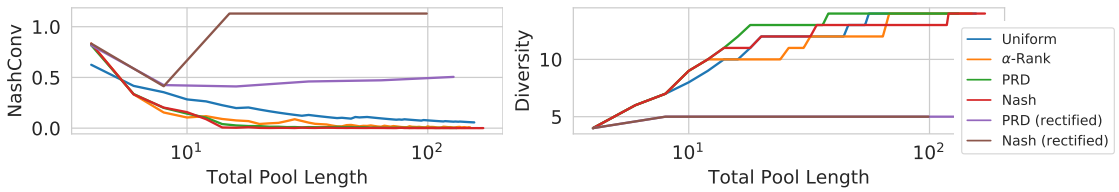
- By Rectified Nash rules, Player 1 only trains policies against policies it beats; i.e., only against Player 2's random policy, and thus it adds the same policy as in iteration 1 to its pool.
- Player 2 trains a best response against the Nash mixture of Player 1's first best response and random policy. This policy also beats all policies of player 1.
- Payoff values are represented in table 3.6b.

4. Iteration 3:

   - Player 1 only trains best responses against Player 2's random policy.
   - Player 2 only trains best responses against the Nash of Player 1's two unique policies. This yields the same policies for player 2 as those previously added to its pool (i.e., a loop occurs).
   - Payoff values are represented in table 3.6c

5. Rectified Nash has looped.

As noted above, Rectified Nash loops at iteration 3, producing already-existing best responses against Player 1's policies. Player 1 is, therefore, constrained to never being able to train best responses against any other policy than Player 2's random policy. In turn, this prevents Player 2 from training additional novel policies, and puts the game in a deadlocked state.



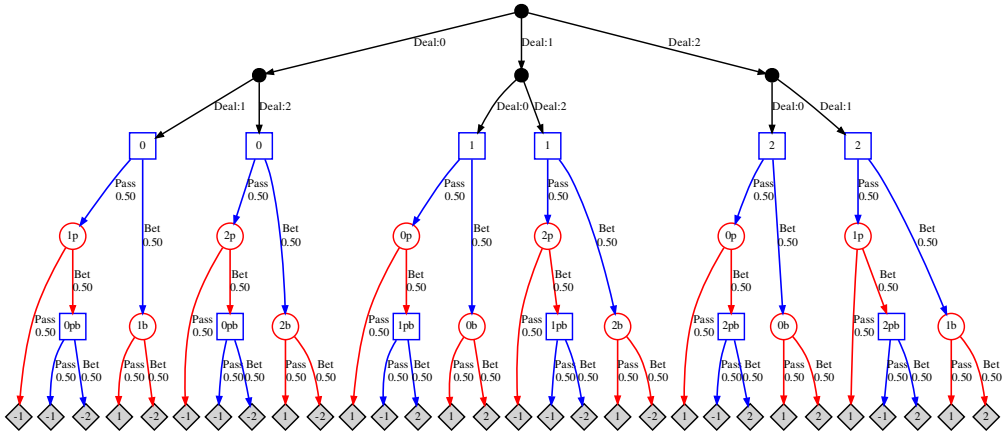
(a) Kuhn poker Exploitability.

(b) Kuhn poker Diversity.

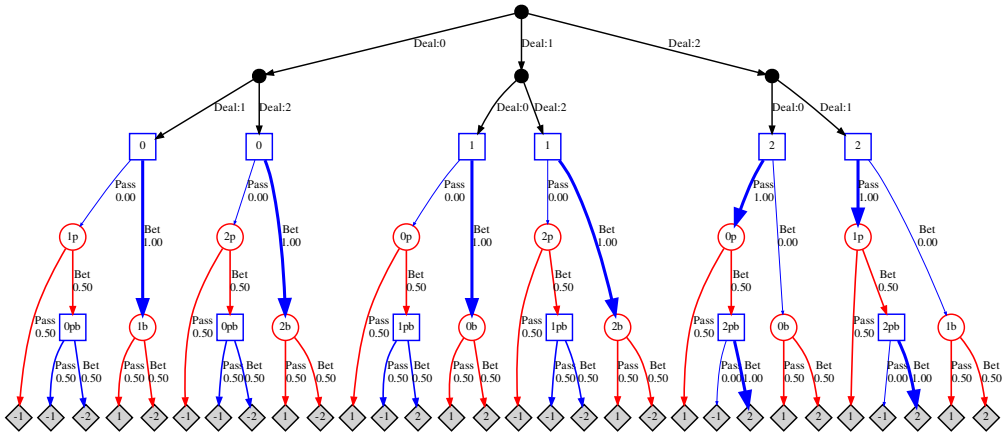
Figure 3.12: Policy Exploitability and Diversity in 2-player Kuhn for a given seed and 100 simulations per payoff entry. Diversity is measured by computing the number of unique policies computed by PSRO.

Noise in the payoff matrices may lead to different best responses against the Nash Mixture of policies, effectively increasing diversity. However, this effect did not seem to manifest in our experiments. To more clearly illustrate this, we introduce a means of evaluating the policy pool diversity by counting the number of unique policies in the pool. Specifically, given that Kuhn poker is a finite state game, comparing policies is straightforward, and only amounts to comparing each policy's output on all states of the games. If two policies have exactly the same output on all the game's states, they are equal; otherwise, they are distinct. We plot in fig. 3.12 the policy diversity of each meta-solver, where we observe that both Rectified Nash and Rectified PRD discover a total of 5 different policies. We have nevertheless noticed that in a few rare seeds, when using low number of simulations per payoff entry (Around 10), Rectified Nash was able to converge to low exploitability scores, suggesting a relationship between payoff noise, uncertainty and convergence of Rectified Nash whose investigation we leave for future work. We also leave the investigation of the relationship between Policy Diversity and Exploitability for future work, though note that there appears to be a clear correlation between both. Overall, these results demonstrate that the Rectified Nash solver fails to discover as many unique policies as the other solvers, thereby plateauing at a low NashConv.

Finally, regarding Rectified PRD, which performs better in terms of NashConv when compared to Rectified Nash, we suspect that payoff noise *in combination* with the intrinsic noise of PRD, plays a key factor - but those two are not enough to deterministically make Rectified PRD converge to 0 exploitability, since in the seed that generated fig. 3.12, it actually does not (though it indeed converges in Figure 3.8). We conjecture this noisier behavior may enable Rectified PRD to free itself from deadlocks more easily, and thus discover more policies on average. A more detailed analysis of Rectified PRD is left as future work.



(a) Initial (uniform) policies.



(b) Player 1's first best response indicated in blue, and the policy it best-responded against in red.

Figure 3.13: Game tree with both players' policies visualized. Player 1 decision nodes and action probabilities indicated, respectively, by the blue square nodes and blue arrows. Player 2's are likewise shown via the red counterparts.

Finally, before closing this section, we would like to investigate potential similarities between one used meta-solver, Projected Replicator Dynamics (PRD), and α-Rank.

Towards Theoretical Guarantees for the Projected Replicator Dynamics

Computing Nash equilibria is intractable for general games and can suffer from a selection problem [49]; therefore, it quickly becomes computationally intractable to employ an exact Nash meta-solver in the inner loop of a PSRO algorithm. To get around this, Lanctot et al. [98] use regret minimization algorithms to attain an approximate correlated equilibrium (which is guaranteed to be an approximate Nash equilibrium under certain conditions on the underlying game, such as two-player zero-sum). A dynamical system from evolutionary game theory that also converges to

$$\begin{bmatrix} 0.1014 & -0.4287 \\ 0.4903 & -0.1794 \end{bmatrix}$$

(a) Iteration 1.

$$\begin{bmatrix} 0.1014 & -0.4287 & -0.2461 & -0.2284 \\ 0.4903 & -0.1794 & -0.4988 & -0.5228 \\ 0.5169 & -0.1726 & -0.4946 & -0.5 \\ 0.5024 & -0.1832 & -0.4901 & -0.5066 \end{bmatrix}$$

(b) Iteration 2.

$$\begin{bmatrix} 0.1014 & -0.4287 & -0.2461 & -0.2284 & -0.264 & -0.2602 & -0.2505 \\ 0.4903 & -0.1794 & -0.4988 & -0.5228 & -0.5015 & -0.5501 & -0.5159 \\ 0.5169 & -0.1726 & -0.4946 & -0.5 & -0.5261 & -0.5279 & -0.4979 \\ 0.5024 & -0.1832 & -0.4901 & -0.5066 & -0.5069 & -0.4901 & -0.5033 \\ 0.4893 & -0.1968 & -0.5084 & -0.4901 & -0.5015 & -0.4883 & -0.4796 \\ 0.4841 & -0.1496 & -0.4892 & -0.491 & -0.4724 & -0.4781 & -0.5087 \\ 0.5179 & -0.1769 & -0.503 & -0.521 & -0.4991 & -0.4739 & -0.4649 \\ 0.4959 & -0.1613 & -0.5123 & -0.518 & -0.5126 & -0.5039 & -0.4853 \end{bmatrix}$$

(c) Iteration 3.

Table 3.6: PSRO(Rectified Nash, BR) evaluated on 2-player Kuhn Poker. Player 1's payoff matrix shown for each respective training iteration.

equilibria under certain conditions is the *replicator dynamics* [20, 46, 162, 176], which defines a dynamical system over distributions of strategies $(\sigma_\pi^k(t) \mid k \in [K], \pi \in \Pi_k)$, given by

$$\dot{\sigma}_\pi^k(t) = \sigma_\pi^k(t) \left[J_k(\pi, \sigma^{-k}(t)) - J_k(\sigma^k(t)) \right] , \quad \text{for all } k \in [K], \pi \in \Pi_k , \tag{3.13}$$

with an arbitrary initial condition. Lanctot et al. [98] introduced a variant of replicator dynamics, termed *projected replicator dynamics* (PRD), which projects the flow of the system so that each distribution $\sigma^k(t)$ lies in the set $\Delta_{\Pi_k}^\gamma = \{\sigma \in \Delta_{\Pi_k} \mid \sigma_s \geq \frac{\gamma}{|\Pi_k|+1}, \forall \pi \in \Pi_k\}$; see, e.g., Nagurney and Zhang [129] for properties of such projected dynamical systems. This heuristically enforces additional "exploration" relative to standard replicator dynamics, and was observed to provide strong empirical results when used as a meta-solver within PSRO. However, the introduction of projection potentially severs the connection between replicator dynamics and Nash equilibria, and the theoretical game-theoretic properties of PRD were left open in Lanctot et al. [98].

Here, we take a first step towards investigating theoretical guarantees for PRD. Specifically, we highlight a possible connection between α-Rank, the calculation of which requires no simulation, and a constrained variant of PRD, which we denote the 'single-mutation PRD' (or s-PRD), leaving formal investigation of this connection for future work.

Specifically, s-PRD is a dynamical system over distributions $(\sigma_\pi^k(t) | k \in [K], \pi \in \Pi_k)$ that follows the replicator dynamics (equation 3.13), with initial condition restricted so that each σ_0^k lies on the 1-skeleton $\Delta_{\Pi_k}^{(1)} = \{\sigma \in \Delta_{\Pi_k} \mid \sum_{\pi \in \Pi_k} \mathbb{1}_{\sigma_s \neq 0} \leq 2\}$. Further, whenever a strategy distribution σ_t^k enters a δ-corner of the simplex, defined by $\Delta_{\Pi_k}^{[\delta]} = \{\sigma \in \Delta_{\Pi_k}^{(1)} \mid \exists \pi \in \Pi_k \text{ s.t. } \sigma_s \geq 1 - \delta\}$, the

non-zero element of $\sigma^k(t)$ with mass at most δ is replaced with a uniformly randomly chosen strategy after a random time distributed according to $\text{Exp}(\mu)$, for some small $\mu > 0$. This concludes the description of s-PRD. We note at this stage that s-PRD defines, essentially, a dynamical system on the 1-skeleton (or edges) of the simplex, with random mutations towards a uniformly-sampled randomly strategy profile s at the simplex vertices. At a high-level, this bears a close resemblance to the finite-population α-Rank dynamics defined in Omidshafiei et al. [138]; moreover, we note that the connection between s-PRD and true α-Rank dynamics becomes even more evident when taking into account the correspondence between the standard replicator dynamics and α-Rank that is noted in Omidshafiei et al. [138, Theorem 2.1.4].

We conclude by noting a major limitation of both s-PRD and PRD, which can limit their practical applicability even assuming a game-theoretic grounding can be proven for either. Specifically, with all such solvers, simulation of a dynamical system is required to obtain an approximate equilibrium, which may be costly in itself. Moreover, their dynamics can be chaotic even for simple instances of two-player two-strategy games [145]. In practice, the combination of these two limitations may completely shatter the convergence properties of these algorithms in practice, in the sense that the question of *how long to wait until convergence* becomes increasingly difficult (and computationally expensive) to answer. By contrast, α-Rank does not rely on such simulations, thereby avoiding these empirical issues.

We conclude by remarking again that, albeit informal, these results indicate a much stronger theoretical connection between α-Rank and standard PRD that may warrant future investigation.

3.2 Computing (Coarse) Correlated Equilibria in N-player Games

It is well understood how to find correlated or coarse-correlated equilibria in N-player, general-sum normal-form games. One runs a linear-program on the payoff matrices, and the correlated equilibrium in question is returned after a more or less long waiting time. However, it is unclear how to compute those in extensive-form games. In particular, although one may want to "flatten" the extensive-form game into a normal-form game, doing so yields an exponentially-large normal-form game, the exponentiality being in the size of the extensive-form game , which can be prohibitive for most solvers.

Fortunately, PSRO offers the opportunity to work with parts of this normal-form game. The hopes are twofold:

- If there exists an equilibrium with small support, then the hope is that PSRO would find it before needing to fully reproduce the full normal-form game.
- The hope is that PSRO ends up producing a reasonably good equilibrium approximation far before having fully reproduced the full normal-form game.

Fulfilling our hopes, this is indeed what happens in most empirically-tested games for Nash equilibria, and α-Rank. We thus wonder whether this method can also be adapted to converge towards (coarse) correlated equilibria, which is the topic of this section.

3.2.1 Adapting PSRO to (Coarse) Correlated Equilibria

We provide an adaptation of PSRO to (coarse) correlated equilibria in algorithm 15, describing all its components and properties in following sections.

Algorithm 15 Joint PSRO(BR)

$\Pi^0 = (\Pi_1^0, ..., \Pi_n^0) = (\{\pi_1^0\}, ..., \{\pi_n^0\})$
$J^0 =$ Payoff Estimation(Π^0)
$\sigma^0 =$ meta-solver(J^0)
for $t \leftarrow \{1, ...\}$ **do**
  for $p \leftarrow \{1, ..., n\}$ **do**
    $\Pi_p^t = \Pi_p^{t-1} \cup BR_p(\Pi^{0:t-1}, \sigma^{t-1})$
  end for
  $J^t =$ Payoff Estimation(Π^t)
  $\sigma^t =$ meta-solver(J^t)
  if $\Pi^t = \Pi^{t-1}$ **then**
    Break.
  end if
end for
return $\Pi^{0:t}$, σ^t

Best Response Operators

At iteration $t+1$, each set $\Pi_p^{0:t}$ can be expanded using either a CCE or CE best response (BR) operator. The type of BR operator used determines the type of equilibrium that JPSRO converges to, as we describe in section 3.2.1.

JPSRO(CCE) : At each iteration there is a single Best-Response objective for each player, which expands the player policy set, $\Pi_p^{0:t+1} = \Pi_p^{0:t} \cup BR_{CCE,p}(\Pi^t, \sigma)$, where

$$BR_{CCE,p}(\Pi^t, \sigma) = \arg\max_{\pi_p^* \in \Pi_p} \mathbb{E}_{\pi \sim \sigma}[J_p(\pi_p^*, \pi_{-p})].$$

The CCE BR attempts to exploit the joint distribution with the responder's own policy preferences marginalized out, resulting in a joint policy distribution over the *other* players' policies. This means that a player is best responding to a weighted mixture of up to $\otimes - p|\Pi_p^t|$ joint opponent policies. This is an upper bound because σ is often sparse.

JPSRO(CE): There is a BR for each possible recommendation a player can get, $\Pi_p^{t+1} = \Pi_p^{0:t} \cup BR_{CE,p}(\Pi^t, \sigma)$, where

$$BR_{CE,p}(\Pi^t, \sigma) = \cup_{\pi_p \in \Pi_p, \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) > 0} \arg\max_{\pi_p^* \in \Pi_p} \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) J_p(\pi_p^*, \pi_{-p}).$$

The CE BR attempts to exploit each policy conditional "slice". In practice, we only calculate a BR for positive support policies (similar to Rectified Nash [14]. Computing the arg max of the BRs can be achieved through RL or exactly traversing the game tree. Similarly each BR is responding to a weighted mixture of up to $\otimes - p|\Pi_p^t|$ joint opponent policies.

Notice that if the distribution is factorizable (like a Nash equilibrium), then the CE Best Response is equal for all player policies, and furthermore is equal to the CCE Best Response, illuminating the connection to PSRO's original Best Response operator.

The best response is independent of the best responding player's policy. We can compute the arg max in a number of ways. Two common ways are exact best response, and reinforcement learning.

Exact Best Response: Maintain exact tabular policies and compute a best response against the joint policies for each player, through maximizing value by traversing the game tree. We employ this approach in this work to allow us to compare meta-solvers without introducing noise from approximate BRs. This method is only suitable for small games, or when using only deterministic policies.

RL: In this setting, the learning algorithms train against randomly sampled joint-policies according to σ, and do standard value maximization. Both on-policy (such as Policy Gradient) and off-policy (such as Q-Learning) are suitable learning algorithms. Function approximation may also be used. This approach has been used extensively in PSRO before.

Meta-Solvers:

Many of the traditional PSRO solvers are factorizable solutions. Equivalently, their joint probabilities can be marginalized without losing any information. In previous work joint solvers have been used [125], however the authors marginalized the distributions so they could be used in classic PSRO.

Uniform: This solver places equal probability mass over each policy it has found so far. PSRO using a uniform distribution is also known as Fictitious Self Play (FSP) [83]. A key advantage of this approach is that it is not necessary to compute the meta-game to obtain this distribution. It is proven to slowly converge in the two-player, constant-sum setting.

Nash Equilibrium (NE): The well known solution concept [132], when used in PSRO is called Double Oracle (DO) [117]. This is difficult to compute for n-player, general-sum, and is equivalent to CE in two-player, constant-sum so we did not benchmark against this meta-solver.

Projected Replicator Dynamics (PRD): An evolutionary method of approximating NE, introduced in [98].

There are a number of solvers which produce full joint distributions. We describe some we think are relevant here. Note that all factorizable solutions mentioned previously can be trivially promoted to full distributions.

α-Rank: A solution concept based on the stationary distribution of a Markov chain [138]. α-Rank has been studied before in the context of PSRO [125], however the authors marginalized over the distribution.

Maximum Welfare (C)CE (MW(C)CE): A non-unique linear formulation that maximizes the sum of payoffs over all players. In the case where there are multiple (C)CEs with maximum welfare we can define a maximum entropy version to spread weight, MEMW(C)CE, and a random version to select one at random, RMW(C)CE. We use the latter as a meta-solver baseline in experiments.

Random Vertex (C)CE (RV(C)CE): A linear formulation. In our implementation we formulate the standard linear (C)CE problem and randomly sample a linear cost function from the unit ball. Note that this selects a random vertex on the (C)CE polytope and is not sampling from within the polytope volume or elsewhere on the polytope surface.

Maximum Entropy (C)CE (ME(C)CE): A unique nonlinear convex formulation that maximizes the Shannon Entropy of the resulting distribution [141]. We do not evaluate this solution concept in this work due to computational difficulties when scaling to large payoff tensors, however we expect its performance to be similar to MG(C)CE.

Maximum Gini (C)CE (MGCE): A unique quadratic convex formulation that maximizes the Gini Impurity (a form of Tsallis Entropy), introduced in this work.

Random Dirichlet: Sample a distribution randomly from a Dirichlet distribution with $\alpha = 1$. This has not been used in the literature before but we believe acts as a good (naive) baseline against RVCE.

Random Joint: Sample a single joint policy from the set. This has not been used in the literature before either but we believe acts as a good (naive) baseline against RV(C)CE.

We propose that (C)CEs are good candidates as meta-solvers. They are more tractable than NEs and can enable coordination to maximize payoff between cooperative agents. In particular we propose three flavours of equilibrium meta-solvers. Firstly, greedy (such as MW(C)CE), which select highest payoff equilibria, and attempt to improve further upon them. Secondly, maximum entropy (such as MG(C)CE) attempts to be robust against many policies through spreading weight. Finally, random samplers (such as RV(C)CE) attempt to explore by probing the extreme points of equilibria. Note that these meta-solvers search through the equilibrium subspace, not the full policy space, and this restriction is a powerful way of achieving convergence. Note that since CEs $\subseteq$ CCEs, one can also use CE meta-solvers with JPSRO(CCE).

Convergence to Equilibria

We provide two convergence proofs for JPSRO. Firstly, when using CCE meta-solvers with a CCE best response operator, which we refer to as JPSRO(CCE), and secondly when using CE meta-solvers with a CE best response operator, which we refer to as JPSRO(CE). Note that, in order to ignore possibly undefined values of $\sigma_t(\pi_{-p}|\pi_p)$, we use the formulation of correlated equilibria using joint probabilities instead of conditional ones. The definitions being equivalent, the conclusions are as well. Note that we also assume that $\forall p, t, |\text{BR}_p^t| > 0, \forall \pi_p \text{ st. } \sigma_t(\pi_p) > 0, |\text{BR}_p^t(\pi_p)| > 0$, i.e. every time a best response should be computed, it is. We also discuss a relaxation of these conditions, and why it is useful, later in this section.

Proof of convergence of JPSRO(CCE)

Theorem 15 (CCE Convergence)**.** *When using a CCE meta-solver and CCE best response in JPSRO(CCE) the mixed joint policy converges to a CCE under the meta-solver distribution.*

We recall the definition of coarse correlated equilibria. For joint probability σ, joint policy set $\Pi = \otimes_p \Pi_p$ where Π_p is the set of valid policies of player p and $\otimes$ is the Cartesian product, and payoff function G, such that $J_p(\sigma)$ is the payoff of player p when all player play according to σ, a Coarse Correlated Equilibrium is a joint distribution σ over Π such that, for any player p and any policy π'_p of player p,

$$\sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi) \tag{3.14}$$

In other words, a CCE is a distribution from which no player has an incentive to unilaterally deviate *before* being assigned their action. From this definition of CCEs, we derive the definition of CCEGap, which measures the above gap over all players

$$\text{CCEGap}(\sigma) = \sum_p \left\lfloor \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \right\rfloor_+$$

where $\lfloor x \rfloor_+ = max(0, x)$, this $\lfloor \rfloor_+$ term being necessary because the gap is potentially negative, as one can see from Equation 3.14. From this definition, we introduce the following lemma:

Lemma 16 (Game CCE and CCEGap)**.** *We have the following equivalence:*

1. *σ is a CCE of the game*
2. *CCEGap(σ) = 0*

Proof. Let us first prove (i) $\rightarrow$ (ii). Suppose σ is a CCE. Then for any player p and any policy π'_p of player p,

$$\sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi)$$

therefore, by subtracting the right hand-term and taking the maximum over $\pi'_p \in \Pi_p$,

$$\max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \leq 0$$

and so

$$\left\lfloor \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(J_p(\pi'_p, \pi_{-p} - J_p(\pi)) \right\rfloor_+ = 0$$

Summing this last inequality over all players yields (ii).

Let us now prove (ii) $\rightarrow$ (i). Suppose that σ is such that $\text{CCEGap}(\sigma) = 0$. Then, for all p,

$$\max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \leq 0 \tag{3.15}$$

For all $\pi''_p \in \Pi_p$ we have

$$\sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi''_p, \pi_{-p}) \leq \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi'_p, \pi_{-p})$$

and therefore, by subtracting $\sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi)$ and using Equation 3.15,

$$\sum_{\pi \in \Pi} \sigma(\pi)(J_p(\pi''_p, \pi_{-p}) - J_p(\pi)) \leq 0$$

Rearranging the terms yields the proof. □

The context of JPSRO motivates us to expand and overload the definition CCEGap. Let us denote by Π^* the policies of the extensive form game, and by $\Pi^{0:t}$ all the policies found by JPSRO by iteration t. We immediately have, for all t, $\Pi^{0:t} \subset \Pi^*$. We expand CCEGap via, for all t,

$$\text{CCEGap}(\sigma, \Pi^*, \Pi^{0:t}) = \quad \sum_p \left\lfloor \max_{\pi^*_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma(\pi)(J_p(\pi^*_p, \pi_{-p}) - J_p(\pi)) \right\rfloor_+$$

The only difference is the search space of π^*_p, which now lives within Π^*, while the policies used in the sum live in $\Pi^{0:t}$. It is nevertheless easy to see that this new definition characterizes CCEs of Π^* (and not of $\Pi^{0:t}$), albeit a restricted class, since $\Pi^{0:t} \subset \Pi^*$ and one can expand σ to be zero over $\Pi^* \setminus \Pi^{0:t}$. Let us now prove Theorem 15.

Proof. To prove that JPSRO with a CCE meta-solver, JPSRO(CCE), converges to a CCE, we need only prove one thing: that JPSRO(CCE) is unable to produce new policies if and only if it has reached a CCE of the extensive form game. Provided this is true, and since all games have a finite number of deterministic policies, we have that JPSRO(CCE) necessarily cannot produce new policies forever, and therefore eventually can only produce already-discovered policies.

Note that the joint distribution σ_t of JPSRO(CCE) is by construction a CCE over $\Pi^{0:t}$ for all t (when using a CCE meta-solver). It is nevertheless not necessarily a CCE of Π^*.

Let us now suppose that JPSRO(CCE) has not produced any new policy for any player at iteration t. Given the JPSRO(CCE) formulation, we can therefore restrict the search space of policies from Π^* to $\Pi^{0:t}$ in the CCEGap max term, since the max of the expression is reached in $\Pi^{0:t}$, and we thus rewrite the CCEGap definition:

$$\sum_p \left\lfloor \max_{\pi'_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \right\rfloor_+ = \sum_p \left\lfloor \max_{\pi'_p \in \Pi^{0:t}_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \right\rfloor_+$$

But since σ_t is a CCE over $\Pi^{0:t}$, the second term is null. Therefore, $\text{CCEGap}(\sigma, \Pi^*, \Pi^{0:t}) = 0$, and according to Lemma 16, σ_t is therefore a CCE over Π^*, which concludes the proof. □

Proof of convergence of JPSRO(CE)

Theorem 17 (CE Convergence)**.** *When using a CE meta-solver and CE best response in JPSRO(CE) the mixed joint policy converges to a CE under the meta-solver distribution.*

We recall the definition of correlated equilibria. Keeping the same notations as above, a correlated equilibrium is a joint distribution σ over Π such that, for any player p and any policies π_p, π'_p of player p,

$$\sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi'_p,\pi_{-p}) \leq \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi_p,\pi_{-p})$$

In other words, a CE is a distribution from which no player has an incentive to unilaterally deviate even *after* having been assigned their action. They are therefore stronger than CCEs, and the result CEs $\subseteq$ CCEs easily follows from the above inequality. From this definition of CEs, we derive the definition of CEGap, which measures the above gap over all players.

$$\text{CEGap}(\sigma) = \sum_{p,\pi_p\in\Pi_p} \left\lfloor \max_{\pi'_p} \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \right\rfloor_+$$

From this definition, we conclude the following lemma:

Lemma 18 (Game CE and CEGap)**.** *We have the following equivalence:*

1. *σ is a CE of the game*
2. *CEGap(σ) = 0*

Proof. Let us first prove (i) $\rightarrow$ (ii). Let σ be a CE of the game. Therefore, for all p, for all $\pi_p, \pi'_p \in \Pi_p$,

$$\sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi'_p,\pi_{-p}) \leq \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi_p,\pi_{-p})$$

therefore

$$\sum_{\substack{\pi_{-p}\in\\ \Pi_{-p}}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \leq 0$$

which is true for all $\pi'_p \in \Pi_p$, so also true for the max over them

$$\max_{\pi'_p\in\Pi_p} \sum_{\substack{\pi_{-p}\in\\ \Pi_{-p}}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \leq 0$$

$$\left\lfloor \max_{\pi'_p\in\Pi_{-p}} \sum_{\substack{\pi_{-p}\in\\ \Pi_{-p}}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \right\rfloor_+ = 0$$

Therefore (i) $\rightarrow$ (ii).

Let us now suppose that σ is such that CEGap(σ) = 0. Thus

$$\sum_{p,\pi_p\in\Pi_p^{0:t}+} \left\lfloor \max_{\pi'_p} \sum_{\substack{\pi_{-p}\in\\ \Pi_{-p}}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \right\rfloor_+ = 0$$

Given the presence of the positivity operator $\lfloor . \rfloor_+$, we deduce that for all p, for all $\pi_p, \pi'_p \in \Pi_p^{0:t}$,

$$\sum_{\substack{\pi_{-p}\in\\ \Pi_{-p}}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \leq 0$$

We therefore deduce

$$\sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi'_p,\pi_{-p}) \leq \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})J_p(\pi_p,\pi_{-p})$$

which concludes the proof. □

Once again, the CEGap definition is extended

$$\text{CEGap}(\sigma,\Pi^*,\Pi^{0:t}) = \sum_{p,\pi_p\in\Pi_p^{0:t}} \left\lfloor \max_{\pi_p^*\in\Pi_p^*} \sum_{\pi_{-p}\in\Pi_{-p}^t} \sigma(\pi_p,\pi_{-p})(J_p(\pi_p^*,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \right\rfloor_+$$

It is once again easy to see that $\text{CEGap}(\sigma,\Pi^*,\Pi^{0:t})$ characterizes CEs of Π^*.

This lemma proven, we prove Theorem 17.

Proof. Once again, it is sufficient to prove that JPSRO(CE) stops producing new policies if and only if it has reached a CE of the extensive form game, the rest of the argument being supplied by the finiteness of the game forcing JPSRO(CE) to eventually stop producing new policies.

Let us now suppose that JPSRO(CE) has not produced any new policy for any new player at iteration t. This means that for all $\pi_p \in \Pi_p^t$,

$$\max_{\pi_p^*\in\Pi_p^*} \sum_{\substack{\pi_{-p}\in \\ \Pi_{-p}^{0:t}}} \sigma(\pi_p,\pi_{-p})J_p(\pi_p^*,\pi_{-p}) = \max_{\pi'_p\in\Pi_p^t} \sum_{\substack{\pi_{-p}\in \\ \Pi_{-p}^{0:t}}} \sigma(\pi_p,\pi_{-p})J_p(\pi'_p,\pi_{-p}).$$

We subtract $\sum_{\pi_{-p}\in\Pi_{-p}^t} \sigma(\pi_p,\pi_{-p})J_p(\pi_p,\pi_{-p})$ to both expressions, apply $\lfloor . \rfloor_+$ and sum over $\pi_p \in \Pi_p^t$ and p, and finally apply the fact that σ is a CE of the restricted game to obtain that

$$\text{CEGap}(\sigma,\Pi^*,\Pi^{0:t}) = \sum_{p,\pi_p\in\Pi_p} \left\lfloor \max_{\pi'_p\in\Pi_p^t} \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p,\pi_{-p})(J_p(\pi'_p,\pi_{-p}) - J_p(\pi_p,\pi_{-p})) \right\rfloor_+ = 0$$

which, by extension, is also true for the CEGap over the extensive form game. By Lemma 18, σ is therefore a CE of the extensive form game, which concludes the proof. □

Relaxation on Proof Requirements Our definition of Best Responses (BRs) is that they are functions that return a set of policies which maximize their value against a given objective. There are two reasons to add a set of policies. Firstly, the max of a given objective can be reached at different points, thus returning a set of policies enables us to potentially include them all. Secondly, using sets also enables us to potentially set some of the BR outputs to $\emptyset$. Concretely, this means that no policy is computed by the BR in that case, which saves compute time and memory. The proofs shown so far rely on each BR having cardinality greater than or equal to 1, which means that one should compute at least one new policy every time the BR operator is called. We can relax this condition into the following conditions, which we prove are sufficient (but not necessary) for convergence.

CCE-Condition:

$$\forall T > 0, p, \exists t > T, |\text{BR}_p^t| \geq 1$$

i.e. each player receives an infinity of best responses.

CE-Condition:

$$\forall T > 0, p, \pi_p, \exists t > T, \text{ either } \forall t' \geq t, \sigma_{t'}(\pi_p) = 0 \text{ or } |\text{BR}_p^t(\pi_p)| \geq 1$$

i.e. any policy of any player is either never selected by the CE meta-solver after some time, or is considered for a best response an infinite number of times.

Solver-Condition: $\forall t, \forall t' \geq t$, if $\forall p, \forall \pi_p \in \Pi_p^{0:t'}, \pi_p \in \Pi_p^{0:t}$, then $\forall \pi \in \Pi^{0:t}$ (or $\pi \in \Pi^{t'}$), $\sigma_t(\pi) = \sigma_{t'}(\pi)$: if no new policy has been added to the pool between t and t', the amount of mass granted to each policy by the solver does not change, i.e. repeating policies does not affect solver outputs, and the solver's outputs are constant given the same pools.

These conditions are sufficient for convergence:

Theorem 19 (Relaxed CCE-Convergence)**.** *When using a CCE meta-solver and CCE best response in JPSRO(CCE), under CCE-Condition and Solver-Condition, the mixed joint policy converges to a CCE under the meta-solver distribution.*

Proof. Let us suppose CCE-Condition and Solver-Condition. We have that JPSRO(CCE) will necessarily be able to produce new policies until it reaches a CCE. Let us prove this: while $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$, JPSRO(CCE) is able to add at least one new policy to its pool. Indeed, let $t > 0$ be such that $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$. Then there exists at least one p such that

$$\max_{\pi'_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) > 0.$$

Let us select one of these p with minimal $t' \geq t, |\text{BR}_p^t| \geq 1$, i.e. the first best response with positive CCEGap to be added to the pool after and including t. t' exists because we suppose CCE-Condition. Let us suppose that no new policies have been added to the pool between t and t'. Then, since no new best response has been added to the pool between t and t', $\sigma_t = \sigma_{t'}$ since we suppose Solver-Condition, and therefore $\forall \pi' \in \text{BR}_p^{t'}$,

$$\sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) > 0.$$

We have that necessarily, $\text{BR}_p^{t'} \cap \Pi_p^{0:t} = \emptyset$, as otherwise σ_t would not be a CCE of $\Pi^{0:t}$: indeed, since σ_t is a CCE of $\Pi^{0:t}$, $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) = 0$, and thus $\forall p, \pi'_p \in \Pi_p^{0:t}$,

$$\sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(J_p(\pi'_p, \pi_{-p}) - J_p(\pi)) \leq 0,$$

thus new best responses can be added to the pool. We therefore have that $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$ implies that at least one new policy can be found by JPSRO.

Thus a new best response can always be added, and will always be added since we have CCE-Condition, to the pool while σ_t is not a CCE of the extensive form game. Therefore, if JPSRO(CCE) is unable to add any new policy to the pool (which has to be verified over all players, or measured through CCEGap), then it must be at a CCE, which concludes the proof. □

Theorem 20 (Relaxed CE-Convergence)**.** *When using a CE meta-solver and CE best response in JPSRO(CE), under CE-Condition and Solver-Condition, the mixed joint policy converges to a CE under the meta-solver distribution.*

Proof. Let us suppose CE-Condition and Solver-Condition. We have that JPSRO(CE) will necessarily be able to produce new policies until it reaches a CE. Let us prove this: while $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$, JPSRO(CE) is able to add at least one new policy to its pool. Indeed, let $t > 0$ be such that $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$. Then there exists at least one p, π_p st. $\sigma_t(\pi_p) > 0$ such that

$$\max_{\pi'_p \in \Pi^*_p} \sum_{\pi_{-p} \in \Pi^t_{-p}} \sigma_t(\pi_p, \pi_{-p})(J_p(\pi'_p, \pi_{-p}) - J_p(\pi_p, \pi_{-p})) > 0.$$

By CE-Condition, we have that either new policies have been added to the pool before any such p, π_p has been selected, or that there exists t' such that $t' \geq t, |\text{BR}_p^t(\pi_p)| \geq 1$. Indeed, if no new best response has been added to the pool by $t' \geq t$, the Solver-Condition implies that for all

these p, π_p st. $\sigma_t(\pi_p) > 0$, we also have $\sigma_{t'}(\pi_p) > 0$, hence there exists t', $|\text{BR}_p^t(\pi_p)| > 1$. Let us select the minimal t' over all p, π_p such that $\text{CEGap}_p(\sigma_t, \Pi^*, \Pi^{0:t})(\pi_p) > 0$.

Let us suppose that no new policies have been added to the pool between t and t'. Then, since no new best response has been added to the pool between t and t', $\sigma_t = \sigma_{t'}$ since we suppose Solver-Condition, and therefore $\forall \pi' \in \text{BR}_p^{t'}(\pi_p), \sum_{\pi_{-p} \in \Pi_{-p}^t} \sigma_t(\pi_p, \pi_{-p})(J_p(\pi'_p, \pi_{-p}) - J_p(\pi_p, \pi_{-p})) > 0$. We have that necessarily, $\text{BR}_p^{t'}(\pi_p) \cap \Pi_p^{0:t} = \emptyset$, as otherwise σ_t would not be a CE of $\Pi^{0:t}$: indeed, since σ_t is a CE of $\Pi^{0:t}$, $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) = 0$, and thus $\forall p, \pi_p \in \Pi_p^{0:t}, \pi'_p \in \Pi_p^{0:t}$,

$$\sum_{\pi_{-p} \in \Pi_{-p}^{0:t}} \sigma_t(\pi_p, \pi_{-p})(J_p(\pi'_p, \pi_{-p}) - J_p(\pi_p, \pi_{-p})) \leq 0.$$

Thus new best responses can be added to the pool. We therefore have that $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$ implies that at least one new policy can be found by JPSRO.

Thus a new best response can always be added, and will always be added since we have CE-Condition, to the pool while σ_t is not a CE of the extensive form game. Therefore, if JPSRO(CE) is unable to add any new policy to the pool (which has to be verified over all players, or measured through CEGap), then it must be at a CE, which concludes the proof. □

Discussion on Relaxation

These relaxed conditions matter especially for JPSRO(CE), which has potentially exponential complexity in term of number of policies to keep (if the solver spreads mass on all policies at each iteration, then the number of policies in each players' pools at iteration t is $\geq 1 + \sum_{k=1}^{t} 2^k = 2^{t+1} - 1$).

Given that the policies produced for one player at the same iteration are potentially similar (even identical), a number of modifications could be imagined to keep JPSRO(CE) tractable. For example: a) randomly select only one π_p from which to best respond for each player, b) only compute a best response for one randomly chosen π_p, or c) compute all BRs, but only add the BR with the largest gap to the pool.

It could make sense to randomly select only one π_p from which to best respond for each player, at each iteration, or even to only compute a best response for one randomly chosen π_p for one randomly-chosen p at each iteration.

Note that it is necessary to impose a condition on the solver (although an alternate Solver-Condition could be formulated). To illustrate this, let us imagine modes between the best response chooser and the solver. Namely, let us imagine a two-player game, for which on even t, in JPSRO(CCE), the best response operator only computes one best response for player 1 (and on odd t, the best response is computed only for player 2). Let us also infer that the current restricted game has two CCEs. The first of these (CCE1) is not "expandable" for player 1, but is for player 2 (i.e. the best response for player 1 is already in the pool, but player 2's best response is not). The second (CCE2) is expandable for player 1, but not for player 2. If the CCE solver outputs CCE1 on even t, and CCE2 on odd t, then the algorithm never produces new policies, and therefore never converges.

Of course, the conditions provided are sufficient, but not necessary, and in the case where best response and meta-solver outputs' randomizations are decorrelated, it makes intuitive sense that the algorithm should also converge with probability 1, which one can prove with a more involved argument.

Evaluation

While the concept of JPSRO is straightforward, careful attention needs to be made around c) defining evaluation metrics, and d) establishing convergence. We discuss these in detail in this section.

On Metagame Estimation: There are two strategies for estimating the meta-game (a normal form payoff tensor populated by the returns of all the policies); exact sampling and empirical sampling.

Exact Sampling: The exact return is computed for each player by traversing the entire game tree. This is only suitable for small games, or when using deterministic policies that cannot reach the majority of the game tree.

Empirical Sampling: For larger games, or situations where the policy cannot be easily queried (for example when using a policy that depends on internal state like an LSTM) we may have to estimate the return through sampling.

In this work we used exact sampling so we could conduct an exact study into the performance of different meta-solvers without introducing noise from other sources. However, the authors believe this approach can be scaled with empirical sampling, as has been achieved with PSRO.

On Quantifying Convergence: Measuring convergence to NE (NE Gap, Lanctot et al. [98]) is suitable in two-player, constant-sum games. However, it is not rich enough in cooperative settings. We propose to measure convergence to (C)CEs using (C)CE-Gaps. A gap of zero implies convergence to an equilibrium.

We also measure the expected value obtained by each player, because convergence to an equilibrium does not imply a high value, and we are ultimately interested in high-value equilibria. Both gap and value metrics need to be evaluated under a meta-distribution. Using the same distribution as the meta-solver may be unsuitable because meta-solvers do not necessarily result in equilibria, may be random, or may maximize entropy. Therefore we may also want to evaluate under other distributions such as MW(C)CE, because it constitutes an equilibrium and maximizes value.

A final relevant measurement is the number of unique policies found over time. The goal of a meta-solver is to expand policy space by proposing a joint policy / joint policies to best-respond to. Failure to find novel policies at an acceptable rate could be evidence of suboptimal performance. Not all novel policies are useful, so caution should be exercised when interpreting this metric. When using a (C)CE meta-solver, a positive (C)CE gap is positive indicates the existence of at least one novel BR policy.

Value: This describes the undiscounted return for each player at the root state of a game when following a joint policy, mixed under a joint distribution.

$$
\begin{aligned}
V_p(\sigma) &= \sum_{\pi \in \Pi} \sigma(\pi) J_p(\pi) = \mathop{\mathbb{E}}_{\pi \sim \sigma} \left[J_p(\pi) \right] \\
V_p(\sigma(\cdot \,|\pi_p)) &= \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_{-p}|\pi_p) J_p(\pi_p, \pi_{-p}) \\
&= \mathop{\mathbb{E}}_{\substack{\pi_{-p} \sim \\ \sigma(\cdot|\pi_p)}} \left[J_p(\pi_p, \pi_{-p}) \right]
\end{aligned}
$$

NE Gap: This quantity describes how close joint policies are to an NE (referred to as NashConv in [98]) under σ. This is only defined for marginal distributions over policies.

$$
\begin{aligned}
\text{NEGap}_p(\sigma) &= \sum_{\pi \in \Pi} \sigma(\pi) J_p(\text{BR}_p, \pi_{-p}) - V_p(\sigma) \\
&= \mathop{\mathbb{E}}_{\pi \sim \sigma} \left[J_p(\text{BR}_p, \pi_{-p}) \right] - V_p(\sigma) \\
\text{NEGap}(\sigma) &= \sum_p \text{NEGap}_p(\sigma)
\end{aligned} \tag{3.16}
$$

CCE Gap: This quantity describes how close joint policies are to a coarse correlated equilibrium (CCE) under σ. The origins of this metric can be deduced from studying the CCE BR

operator.

$$\begin{aligned}\text{CCEGap}_p(\sigma) &= \left\lfloor \sum_{\pi \in \Pi} \sigma(\pi) J_p(\text{BR}_p, \pi_{-p}) - V_p(\sigma) \right\rfloor_+ \\ &= \left\lfloor \mathop{\mathbb{E}}_{\pi \sim \sigma} \left[J_p(\text{BR}_p, \pi_{-p}) \right] - V_p(\sigma) \right\rfloor_+ \\ \text{CCEGap}(\sigma) &= \sum_p \text{CCEGap}_p(\sigma)\end{aligned}$$

Where $\lfloor x \rfloor_+ = max(0, x)$, is the non-negative operator. Note that it is possible for a best response over all joint strategies to have lower value than playing according to the joint distribution for a given player (because a BR is blind to the best responding player's correlation with the opponent policies, and deviating from this correlation can hurt performance).

CE Gap: This quantity describes how close joint policies are to a correlated equilibrium (CE) under σ.

$$\begin{aligned}\text{CEGap}_p(\sigma, \pi_p) &= \left\lfloor \sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_{-p}|\pi_p) J_p(\text{BR}_p(\pi_p), \pi_{-p}) - V_p(\sigma(\cdot\,|\pi_p)) \right\rfloor_+ \\ &= \left\lfloor \mathop{\mathbb{E}}_{\substack{\pi_{-p} \sim \\ \sigma(\cdot|\pi_p)}} \left[J_p(\text{BR}_p(\pi_p), \pi_{-p}) \right] - V_p(\sigma(\cdot\,|\pi_p)) \right\rfloor_+ \\ \text{CEGap}_p(\sigma) &= \sum_{\pi_p \in \Pi_p} \sigma(\pi_p) \text{CEGap}_p(\sigma, \pi_p) \\ \text{CEGap}(\sigma) &= \sum_p \text{CEGap}_p(\sigma)\end{aligned}$$

Unique Policy: Each iteration of JPSRO(CCE) produces n new policies (one for each player), and JPSRO(CE) produces up to the number of policies found so far. These are best responses to the joint mixture of existing policies, however, they are not guaranteed to be distinct from previous policies that have been found. The number of unique policies found so far could be a good indicator of how efficiently a meta-solver is producing new policies.

Games: We study several games with JPSRO; Kuhn Poker, Trade Comm, and Sheriff. These cover three-player, general-sum, and common-payoff games. Implementations of all the games are available in OpenSpiel [100].

Kuhn Poker: A simplified n-player, zero-sum, sequential, imperfect information version of poker. It consists of $n + 1$ playing cards. In each round of the game, every player remaining *antes* one chip. One card is dealt to each player. Each player has two choices, *bet* one chip or *check*. If a player bets other players have the option to *call* or *fold*. Out of the players that bet, the one with the highest card wins. If all players check, the player with the highest card wins. The original two-player game is described in [94]. An n-player extension is described in [97]. Additional information about the game (such as equilibrium) can be found in [85].

Trade Comm: A simple two-player, common-payoff trading game [167]. In this game each player (in secret) receives one of I different items. The first player can then make one of I utterances to the second agent, and vice versa. Then each agent chooses one of I^2 trades in private, if the trade is compatible both agents receive 1 reward, otherwise both receive 0. The goal of the agents is therefore to find a bijection between the items and utterances and the trade proposal. There are I^4 deterministic policies per player, and good learning algorithms will be

able to search over these policies. Because the game is common-payoff, it is very transitive, and has many dominated strategies, however there are multiple strategies with equal payoff, and therefore many equilibria in partially explored policy space. It is for this reason many learning algorithms get stuck exploiting sub-optimal policies they have already found.

Sheriff: A simplified two-player, general-sum version of the board game Sheriff of Nottingham [62]. The game consists of a smuggler, who is motivated to import contraband without getting caught, and a sheriff, who is motivated to either find contraband or accept bribes. The players negotiate a bribe over several rounds after which the bribe is accepted or rejected. If the sheriff finds contraband, the smuggler pays a fine, otherwise if no contraband is found the sheriff must pay compensation to the smuggler. The smuggler also gets value from smuggling goods. The game has different optimal values for NFCCE, EFCCE, EFCE, and NFCE solutions concepts.

Results: We evaluate a number of (C)CE meta-solvers in JPSRO on pure competition, pure cooperation, and general-sum games (Section 3.2.1). All games used are available in OpenSpiel [100]. More thorough descriptions of the games used can be found at the end of that section. We use an exact BR oracle, and exactly evaluate policies in the meta-game by traversing the game tree to precisely isolate the meta-solver's contribution to the algorithm.

We compare against common meta-solver including uniform, α-Rank [125, 138], Projected Replicator Dynamics (PRD) [98] which is an NE approximator, and random vertex (coarse) correlated equilibrium (RV(C)CE) which randomly selects a solution on the vertices of (C)CE polytope. We also include a random joint and random Dirichlet solvers as baselines. We treat the solutions to the meta-solvers as full joint distributions. Random solvers were evaluated with five seeds and we plot the mean. When evaluating, we measure equilibrium gaps under their own meta-solver distribution and MW(C)CE to provide a consistent and value maximizing comparison. Experiments were run for up to 6 hours, after which they were terminated.

Kuhn Poker [94, 97, 169] is a zero-sum poker game with only two actions per player. The two-player variant is solvable with PSRO, however the three-player version benefits from JPSRO. The results in Figure 3.15a show rapid convergence to equilibrium.

Trade Comm is a two-player, common-payoff trading game, where players attempt to coordinate on a compatible trade. This game is difficult because it requires searching over a large number of policies to find a compatible mapping, and can easily fall into a sub-optimal equilibrium. Figure 3.15b shows a remarkable dominance of CCE meta-solvers. It is clear that traditional PSRO meta-solvers cannot cope with this cooperative setting.

Sheriff [62] is a two-player, general-sum negotiation game. It consists of bargaining rounds between a smuggler, who is motivated to import contraband without getting caught, and a sheriff, who is motivated to find contraband or accept bribes. Figure 3.15c shows that JPSRO is capable of finding the optimal value.

3.2.2 Discussion

There has been significant recent interest in solving the equilibrium selection problem [138, 141]. This section provides a novel approach which is computationally tractable, supports general-support solutions, and has favourable scaling properties when the solution is full-support.

PSRO has proved to be a formidable learning algorithm in two-player, constant-sum games, and JPSRO, with (C)CE meta-solvers, is showing promising results on n-player, general-sum games. The secret to the success of these methods seems to lie in (C)CEs ability to compress the search space of opponent policies to an expressive and non-exploitable subset. For example, no dominated policies are part of CEs, and during execution there are no policies a player would rather deviate to. For (C)CE meta-solvers, if there is a value-improving BR it is guaranteed to be a novel policy.

There is a rich polytope of possible equilibria to choose from, however, a meta-solver must pick one at each time step. There are three competing properties which are important in this regard, exploitation, robustness, and exploration. For exploitation, maximum welfare equilibria appear

to be useful. However, to prevent JPSRO from stalling in a local equilibrium it is essential to randomize over multiple solutions satisfying the maximum welfare criterion. To produce robust BRs, entropy maximizing meta-solvers (such as MG(C)CE) have better empirical value and convergence than the uniform meta-solver. For exploration, we can randomly select a valid equilibrium at each iteration which outperforms random joint and random Dirichlet by a significant margin (similar to AlphaStar's "exploiter policies" [179]). Furthermore, one could also switch between meta-solvers at each iteration to achieve the best mix of exploitation and exploration.

Another strength of (C)CE meta-solvers is that they appear to perform well across many different games, with different numbers of players and payoff properties.

3.2.3 Conclusions

We have shown that JPSRO converges to an NF(C)CE over joint policies in extensive form and stochastic games. Furthermore, there is empirical evidence that some meta-solvers also result in high value equilibria over a variety of games. We argue that (C)CEs are an important concept in evaluating policies in n-player, general-sum games and thoroughly evaluate several meta-solvers. Finally, we believe that both MG(C)CE and JPSRO can scale to large problems, by using stochastic online meta-solvers for the former and exploiting function approximation and RL for the latter.

3.3 The Canonical PSRO Solver

In this thesis-exclusive section, we explore how to generalize the methods introduced in the above two developments, α-PSRO and JPSRO, to a general class of game-theoretic equilibrium. We first start by defining a general expression of game-theoretic equilibria, furnishing it with examples, and then show how this expression may be used to derive a PSRO algorithm converging towards the chosen equilibrium.

3.3.1 A General Equilibrium Framework: the SMD Decomposition

In this section, we will write $\Pi_i^R \subseteq \Pi_i$ an arbitrary, **R**estricted subset of Π_i, for all i; $\mathcal{J} = \{J : \Pi_1, ..., \Pi_N \to \mathbb{R}^N\}$ the set of payoff functions,.

We also write $\Pi^N = \Pi_1 \times ... \times \Pi_N$, and $\Pi^{N,R} = \Pi_1^R \times ... \times \Pi_N^R$

We define an SMD (Sigma-Metric-Deviation) Decomposition as a triplet (σ, m, $\mathcal{D}$), where

- $\sigma : \Pi^{N,R}, \mathcal{J} \to \Delta(\Pi^{N,R})$ is the equilibrium distribution function,
- $m : \Delta(\Pi^{N,R}), \mathcal{J}, \Pi^N \to \mathbb{R}$ is the equilibrium metric function,
- $\mathcal{D} : \Delta(\Pi^{N,R}), \mathcal{J}, \Pi^N \to 2^{\Pi_1}, ..., 2^{\Pi_N}$ is the equilibrium deviation function.

Note that the deviation function may return several policies for each player, hence its return domain being not in Π_i, but in 2^{Π_i}. To simplify our developments, we introduce the following notations:

For given Π_1, ..., Π_N, Π_1^R, ..., Π_N^R, $J \in \mathcal{J}$, we write

- $\sigma^R = \sigma(\Pi^{N,R}, J)$,
- $m^R = m(\sigma^R, J, \Pi^N)$,
- $D^R = \mathcal{D}(\sigma^R, J, \Pi^N)$, and $\mathcal{D}_k^R$ is the k-th player's deviation output.

For an SMD Decomposition to be an Equilibrium Decomposition, several conditions must be met:

Definition 13 (SMD Equilibrium Decomposition)**.** An SMD Decomposition is said to be an SMD Equilibrium Decomposition if

1. *(Metric calibration)* $m(\sigma^R,\, J,\, \Pi_1^R,\, ...,\, \Pi_N^R) \leq 0 \quad \forall\, \Pi_i^R \subseteq \Pi_i,\; i \in \mathcal{N}$,
2. *(Deviation well-definedness)* $\forall k \in \mathcal{N},\; D_k^R \neq \varnothing$,
3. *(Metric-Deviation calibration)* $m^R > 0 \Longrightarrow \exists k \in \mathcal{N},\; \pi_k \in D_k^R,\; \pi_k \notin \Pi_k^R$,
4. *(Deviation-Metric calibration)* $\forall k \in \mathcal{N},\; \forall \pi_k \in D_k^R,\; \pi_k \in \Pi_k^R \Longrightarrow m^R \leq 0$.

An SMD Equilibrium Decomposition characterizes an equilibrium via its m function:

Definition 14 (SMD-Decomposed Equilibrium)**.** The game-theoretic equilibrium characterized by SMD Equilibrium Decomposition $(\sigma, m, \mathcal{D})$ is a distribution σ^R over Π^N such that $m(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N) \leq 0$. $\sigma^R \in \Delta(\Pi^N)$ does not need to be an output of the function σ.

In the rest of this section, we will use the following objects: we define $J_{M,k}^R$ to be the tensor with entries $J_{M,k}^R[i_1,\, ...,\, i_N] = J_k(\pi_{i_1},\, ...,\, \pi_{i_N})$, where $\pi_{i_k} \in \Pi_k^R$ is the i_k-th policy of Π_k^R for a given indexing. We will also reuse the above definition of σ^R. We will also dispense with proving the correctness of the SMD decompositions to minimize tediousness.

We now show that several classical game-theoretical equilibria satisfy the above representation:

Nash equilibrium: For Nash equilibria, we have

- $\sigma_{Nash}(\Pi^{N,R}, J) = \text{Nash}\left((J_{M,k}^R)_{k=1..N}\right)$, a Nash equilibrium of the normal form game induced by $(J_{M,k}^R)_{k=1..N}$,
- $m_{Nash}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N) = \sum\limits_{i \in \mathcal{N}} \max\left(0, \max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi \in \Pi_1 \times ... \times \Pi_N} \sigma^R(\pi)\left(J_i(\pi_i', \pi_{-i}) - J_i(\pi_i, \pi_{-i})\right)\right)$, the exploitability function,
- $\mathcal{D}_{Nash}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N)_i = \arg\max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi \in \Pi_1 \times ... \times \Pi_N} \sigma^R(\pi) J_i(\pi_i', \pi_{-i})$, the best-response function.

Coarse-correlated equilibrium: The objects characterizing coarse-correlated equilibria are similar to those characterizing Nash equilibria:

- $\sigma_{CCE}(\Pi^{N,R}, J) = \text{CCE}\left((J_{M,k}^R)_{k=1..N}\right)$, a coarse-correlated equilibrium of the normal form game induced by $(J_{M,k}^R)_{k=1..N}$,
- $m_{CCE}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N) = \sum\limits_{i \in \mathcal{N}} \max\left(0, \max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi \in \Pi_1 \times ... \times \Pi_N} \sigma^R(\pi)\left(J_i(\pi_i', \pi_{-i}) - J(\pi_i, \pi_{-i})\right)\right)$, the CCE-gap function,
- $\mathcal{D}_{CCE}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N)_i = \arg\max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi \in \Pi_1 \times ... \times \Pi_N} \sigma^R(\pi) J_i(\pi_i', \pi_{-i})$, the best-response function.

Correlated equilibrium: The SMD decomposition of correlated equilibria starts to differ quite significantly from Nash equilibria's:

- $\sigma_{CE}(\Pi^{N,R}, J) = \text{CE}\left((J_{M,k}^R)_{k=1..N}\right)$, a correlated equilibrium of the normal form game induced by $(J_{M,k}^R)_{k=1..N}$,
- $m_{CE}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N) = \sum\limits_{i \in \mathcal{N}} \sum\limits_{\pi_i \in \Pi_i^R} \max\left(0, \max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi_{-i} \in \Pi_{-i}^R} \sigma^R(\pi_i, \pi_{-i})\left(J_i(\pi_i', \pi_{-i}) - J_i(\pi_i, \pi_{-i})\right)\right)$,
- $\mathcal{D}_{CE}(\sigma^R,\, J,\, \Pi_1,\, ...,\, \Pi_N)_i = \{\arg\max\limits_{\pi_i' \in \Pi_i} \sum\limits_{\pi_{-i} \in \Pi_{-i}} \sigma^R(\pi_i, \pi_{-i}) J_i(\pi_i', \pi_{-i}) \mid \pi_i \in \Pi_i^R,\; \sum\limits_{\pi_{-i} \in \Pi_{-i}} \sigma^R(\pi_i, \pi_{-i}) > 0\}$.

α-Rank: The SMD decomposition of α-Rank is indirectly explicitized in Section 3.1:

- $\sigma_{\alpha\text{-Rank}}(\Pi^{N,R}, J) = \alpha\text{-Rank}\left((J^R_{M,k})_{k=1..N}\right)$, the α-Rank equilibrium of the normal form game induced by $(J^R_{M,k})_{k=1..N}$,
- $m_{\alpha\text{-Rank}}(\sigma^R, J, \Pi_1, ..., \Pi_N) = \sum_{i\in\mathcal{N}} \max_{\pi'_i\in\Pi_i\setminus\Pi^R_i} \sum_{\pi\in\Pi_1\times...\times\Pi_N} \sigma^R(\pi)\mathbb{1}_{J_i(\pi'_i,\pi_{-i})>J_i(\pi_i,\pi_{-i})}$ the PBR-gap function with novelty-boundedness enforcement, under the convention that the max over an empty set is 0,
- $\mathcal{D}_{\alpha\text{-Rank}}(\sigma^R, J, \Pi_1, ..., \Pi_N)_i = \begin{cases} \arg\max_{\pi'_i\in\Pi_i\setminus\Pi^R_i} \sum_{\pi\in\Pi_1\times...\times\Pi_N} \sigma^R(\pi)\mathbb{1}_{J_i(\pi'_i,\pi_{-i})>J_i(\pi_i,\pi_{-i})} & \text{if } \Pi_i\setminus\Pi^R_i \neq \varnothing \\ \Pi^R_i, & \text{otherwise} \end{cases}$, the novelty-bound PBR oracle.

All the above SMD decompositions can be justified by the different PSRO variants' proofs of convergence.

Pareto front distribution: We call a Pareto front distribution a distribution over non-Pareto-dominated strategies.

- $\sigma_{\text{Pareto}}(\Pi^{N,R}, J) = \text{Pareto}\left((J^R_{M,k})_{k=1..N}\right)$, a Pareto distribution over the normal form game induced by $(J^R_{M,k})_{k=1..N}$,
- $m_{\text{Pareto}}(\sigma^R, J, \Pi_1, ..., \Pi_N) = \sum_{\pi\in\Pi^R_1\times...\times\Pi^R_N} \sigma^R(\pi) \max_{\pi'\in\Pi_1\times...\times\Pi_N} \left(\prod_{i\in\mathcal{N}} \mathbb{1}_{J_i(\pi')\geq J_i(\pi)}\right)\left(\mathbb{1}_{\sum_{i\in\mathcal{N}} J_i(\pi')>\sum_{i\in\mathcal{N}} J_i(\pi)}\right)$ the Pareto-gap function, where the product verifies that no player loses value, while the sum checks that at least one player increases value.
- $\mathcal{D}_{\text{Pareto}}(\sigma^R, J, \Pi_1, ..., \Pi_N)_i = \left\{\pi'_i \mid \pi' \in \arg\max_{\pi'\in\Pi_1\times...\times\Pi_N} \left(\prod_{i\in\mathcal{N}} \mathbb{1}_{J_i(\pi')\geq J_i(\pi)}\right)\left(\mathbb{1}_{\sum_{i\in\mathcal{N}} J_i(\pi')>\sum_{i\in\mathcal{N}} J_i(\pi)}\right), \pi \in \Pi^R_1\times...\times\Pi^R_N, \sigma^R(\pi) > 0\right\}$ the Pareto-oracle.

We notice that in all the above SMD decompositions, the Deviation function is obtained via value-maximization; however, this need not be the case, and *e.g.* softer deviation functions could also be used. For example, in the case of Nash equilibria, any function which increases (But not necessarily maximizes) value when possible also works as an oracle.

3.3.2 A General PSRO Framework: SMDRO

Our problem of interest is finding an equilibrium with SMD decomposition $(\sigma, m, \mathcal{D})$ in a given game. The Sigma-Metric-Deviation Response Oracle (SMDRO), described in Algorithm 16, exploits the SMD decomposition to produce an algorithm which provably converges towards the SMD-decomposed equilibrium, as we prove below:

Algorithm 16 SMDRO(σ, m, $\mathcal{D}$)

Initialize the players' policy sets $(\Pi^R_k)_{k=1..N}$ via random policies
Compute $\sigma^R = \sigma(\Pi^R_1, ..., \Pi^R_N, J^R_M)$
while $m(\sigma^R, J, \Pi_1, ..., \Pi_N) > 0$ **do**
    Compute $D^R = \mathcal{D}(\sigma^R, J, \Pi_1, ..., \Pi_N)$
    Append new policies to players' policy pools: $\Pi^R_k = \Pi^R_k \cup D^R_k \quad \forall k \in \mathcal{N}$
    Compute $\sigma^R = \sigma(\Pi^R_1, ..., \Pi^R_N, J^R_M)$
end while

Theorem 21. *Given (σ, m, $\mathcal{D}$) and SMD Equilibrium Decomposition, SMDRO(σ, m, $\mathcal{D}$) always converges to the SMD's decomposed equilibrium.*

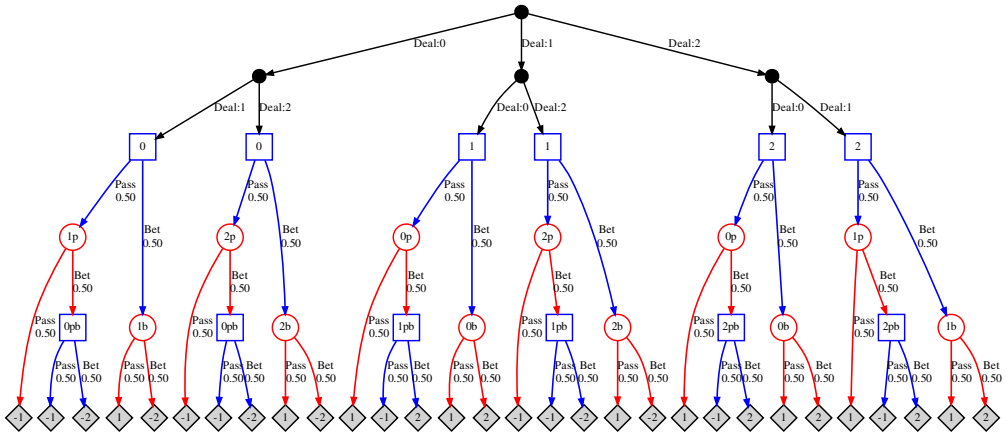
Proof. By Definition 14, we are at equilibrium when $m \leq 0$.

By property (4) of Definition 13, this happens when the deviation function only returns known policies.

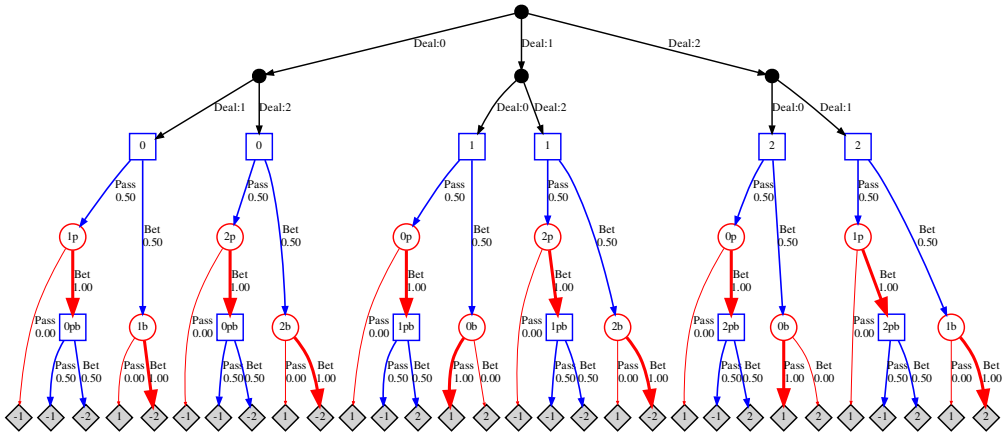
Since there are only a finite number of policies in the game, PSRO will necessarily reach a point where the deviation function only returns known policies, and thus a point where $m \leq 0$, which concludes the proof. □

3.4 Limitations and Future Work

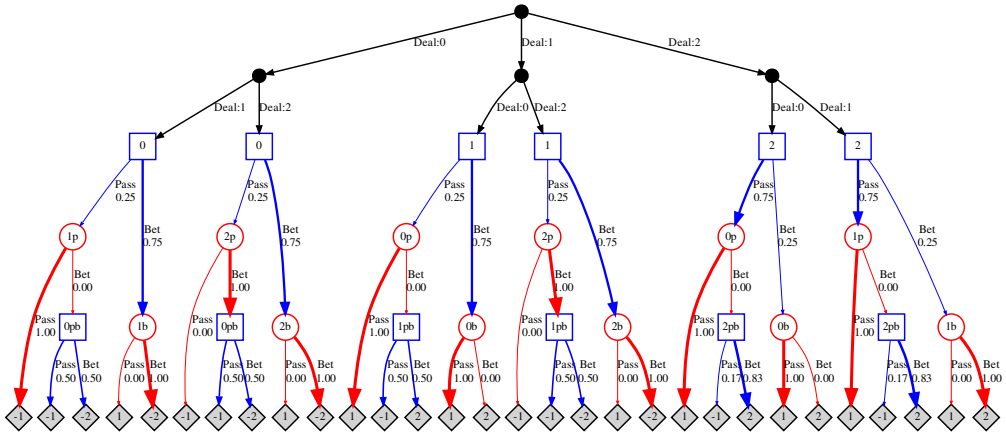
We have shown that PSRO-derived approaches are able to converge to any equilibrium which can be expressed following our framework. However, PSRO techniques potentially require as many iterations as there are deterministic policies in the game. This seriously limits their applications in *e.g.* high-N N-player games, as game-solving complexity typically scales exponentially with the number of players. Two options are available to us then: either find ways to make the algorithm faster as a function of total game size, or find a way to simplify the games which will lower their size. In this dissertation, we have investigated the second way, through the following question: when, and how, can we simplify large-N N-player games such that finding their equilibria is much easier ? We answer these questions in the next chapters, via the use of Mean-Field games.



(a) Initial (uniform) policies.

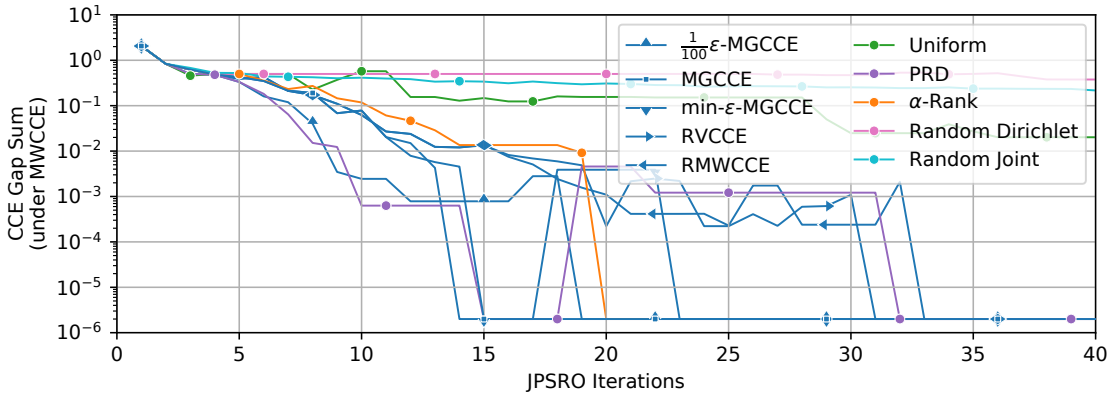


(b) Player 2's first best response indicated in red, and the policy it best-responded against in blue.

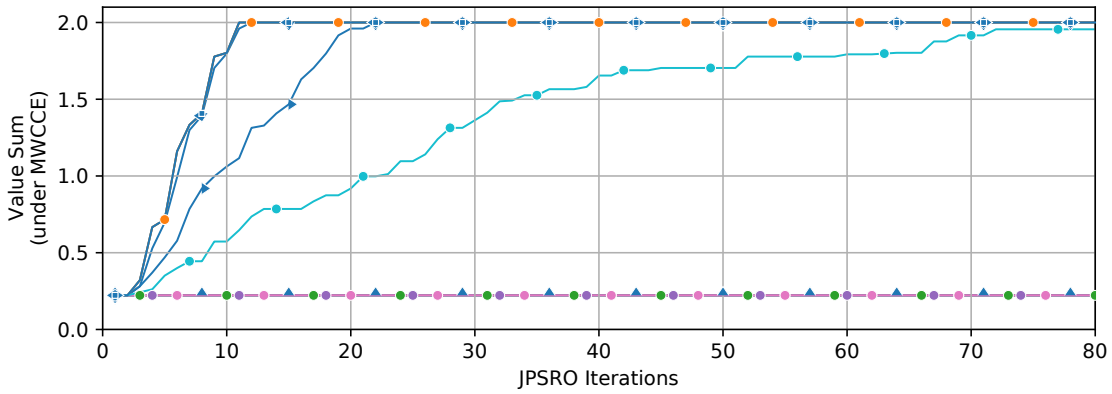


(c) Player 2's second best response indicated in red, and the policy it best-responded against in blue.

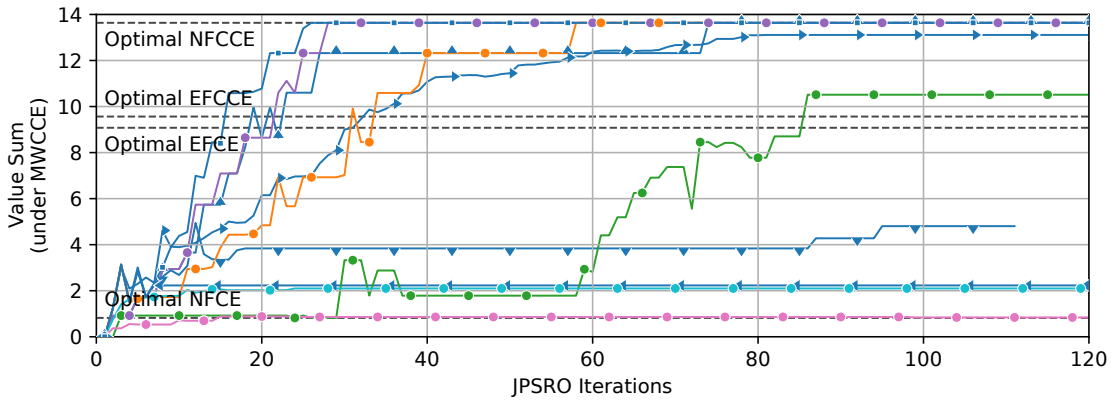
Figure 3.14: Game tree with both players' policies visualized. Player 1 decision nodes and action probabilities indicated, respectively, by the blue square nodes and blue arrows. Player 2's are likewise shown via the red counterparts.



(a) CCE Gap on three-player Kuhn Poker. Several meta-solver converge to within numerical accuracy (data is clipped) of a CCE.



(b) Value sum on three-item Trade Comm. The approximate CCE meta-solver was not sufficient to converge in this game, however all valid CCE meta-solvers were able to converge to the optimal value sum.



(c) Value sum on Sheriff. The optimal maximum welfare of other solution concepts are included to highlight the appeal of using NFCCE.

Figure 3.15: JPSRO(CCE) on various games. MGCCE is consistently a good choice of meta-solver over the games tested.

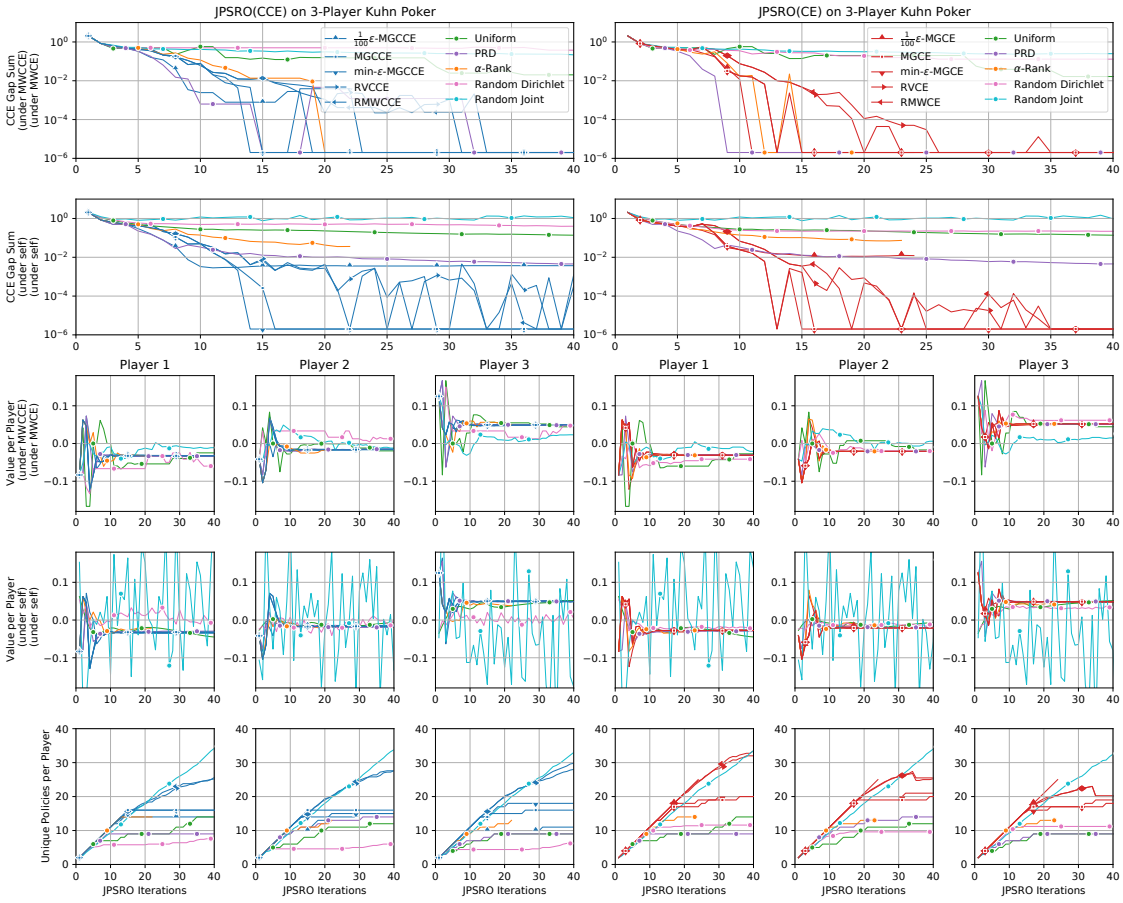


Figure 3.16: JPSRO(CCE) and JPSRO(CE) on three-player Kuhn Poker. All (C)CE meta-solvers, PRD and α-Rank find joint policies capable of supporting equilibrium (although α-Rank was slow and was terminated after 6 hours). This is some evidence that classic meta-solvers designed for the two-player, zero-sum setting can generalize well to the three-player, zero-sum.

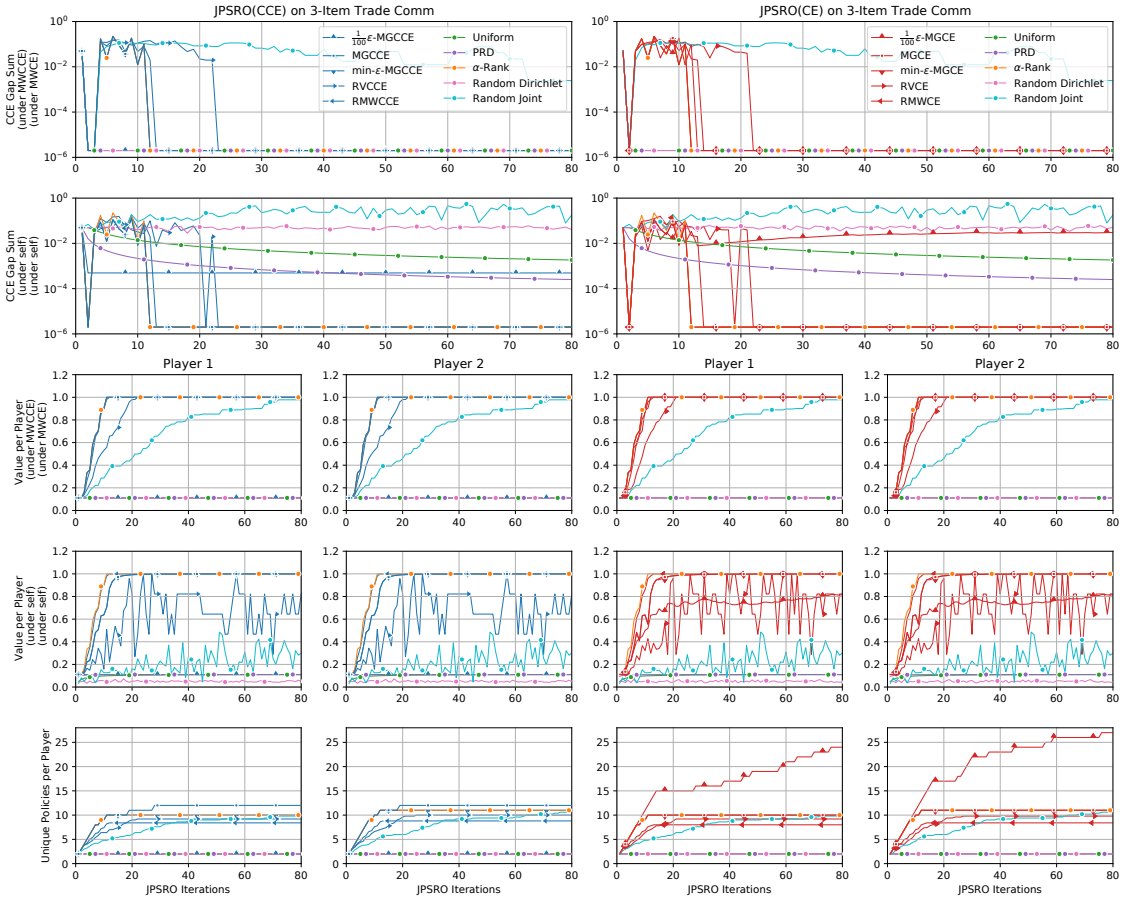


Figure 3.17: JPSRO(CCE) and JPSRO(CE) on three-item Trade Comm. In JPSRO(CCE), $\frac{1}{100}$ min-MGCCE fails to find the maximum welfare equilibrium, however, all other (C)CE meta-solvers find the maximum welfare equilibrium. Unexpectedly, α-Rank performs well on this game, while all other classic meta-solvers fail to make progress on this purely cooperative game. Performing well on this game requires exploration, so the random joint meta-solver is able to make progress, albeit naively and slowly.

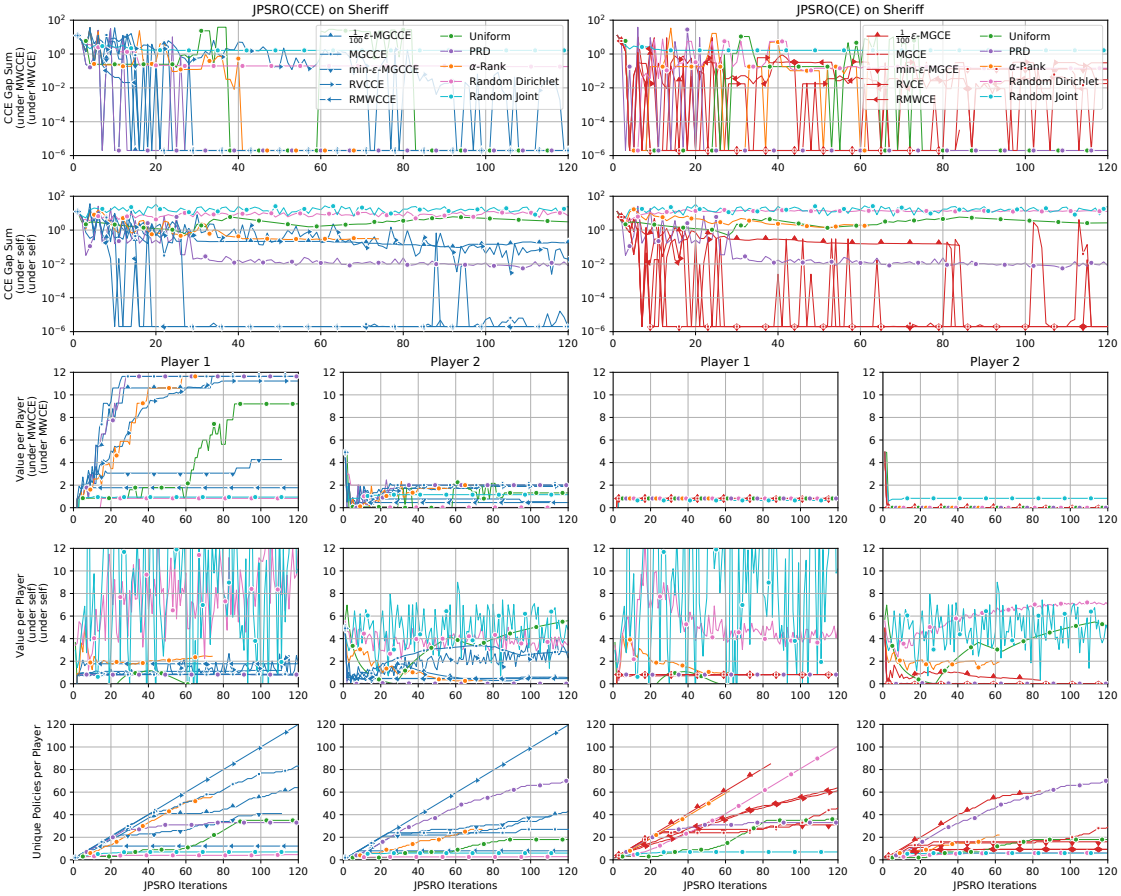


Figure 3.18: JPSRO(CCE) and JPSRO(CE) on Sheriff. This game is interesting because it is general-sum and different solution concepts have different optimal maximum welfare values. The maximum welfare NFCCE is 13.64 for the smuggler and 2.0 for the sheriff which JPSRO(CCE) successfully finds, while the maximum welfare NFCE is 0.82 for the smuggler and 0.0 for the sheriff which JPSRO(CE) successfully finds. This demonstrates the appeal of using NFCCE as a target equilibrium. Interestingly, for this game, $\frac{1}{100}\epsilon$-MG(C)CE was able to produce BRs of high enough quality to converge which is evidence that scaled methods that only approximate (C)CEs may be enough in some settings. RMWCCE converged to an equilibrium, but not the welfare maximizing one, providing evidence that greedy meta-solvers are not always suitable. In a similar argument, min-ϵ-MGCCE did not reach the maximum welfare solution within the allocated number of iterations. RV(C)CE is efficient at finding novel policies but ones of limited utility. PRD and α-Rank perform well and find the maximum welfare (C)CE equilibria.

Chapter 4

The mass is the sea: Scaling Equilibria Beyond Finite Players Through Mean-Field Games

The complexity of describing and computing equilibria in games with a finite number of players grows exponentially as the size of the population increases[1]. Such computations are however extremely useful in many different fields: traffic routing [59, 130, 186], energy management [6, 63, 109, 135, 173, 188, 190], mechanism-design [17, 131], among many others. Computation is hampered by, among others, the need to consider every individual player's states and actions. The joint player state space complexity thus grows combinatorially, the difficulty being akin to producing an exact simulation of an N particle system - easy for low N, impossible for high N. In such context, taking insight from statistical physics, we focus directly on the distribution of the population of particles instead of simulating every one of them, and thus consider *Mean-Field games*.

Mean-Field games (MFGs) have been introduced to simplify the analysis of Nash equilibria in games with a very large number of identical players interacting in a symmetric fashion (*i.e.*, through the distribution of all the players). The key idea is to solely focus on the interactions between a representative infinitesimal player and a (so-called Mean-Field) term capturing the effect of the population of players. Understanding the behavior of one typical player is enough, as the behavior of the whole population can be deduced from it, since all players are assumed to be identical. This approach circumvents the difficulties induced by representing an extremely large population of agents. Since their introduction by Lasry and Lions [102], and Caines, Huang and Malhamé [86], MFGs have been extensively studied both from a theoretical and a numerical viewpoint [2, 16, 35, 39, 40]. Applications in various fields such as energy management [56, 114, 128], financial markets [38, 42, 72], macroeconomics [5, 68, 72], vehicle routing [56, 77, 175], mechanism design [57, 87, 93] or epidemics dynamics [12, 26, 104, 151] have already been considered. Most of the literature focuses on stochastic differential games and characterize their solution via the consideration of partial differential or stochastic differential equations [16, 35, 39, 40]. A forward equation captures the full population dynamics, while a backward one represents the evolution of the value function for a representative agent. With few exceptions, such as in [96] which considers a class of closed-loop controls with a common signal or in [34, 52] which considers correlated equilibria as we explain below, only pure or mixed Nash equilibria have been considered so far. This is in stark contrast with the panoply of alternative notions of equilibria considered for games with a finite number of players [3, 11, 22, 61, 120, 121, 138, 152, 184, 185]. In the context of MFGs,

[1] To see this, picture a M-action N-player game. The payoff tensor of a such game is of size $N\,M^N$, a quantity exponential in N. Assuming that this game is such that its Nash equilibrium is fully mixed, thus computing the Nash equilibrium will require going through every payoff tensor cell at least once, hence leading to an at least exponential relationship between equilibrium computation time and number of players.

mixed Nash equilibria with relaxed controls have been studied in [41, 95]. For example, mixed controls arise naturally in the context of MFG with optimal stopping where players should avoid simultaneous actions, as studied by Bertuci [18] or Bouveret et al. [24]. Moreover, mixed policies are commonly considered in the setting of reinforcement learning for MFG, see for example [4, 73, 149]. More generally the question of learning equilibria in MFGs has gained momentum in the past few years [2, 36, 74, 75, 147, 149].

Studying and understanding learning behaviors in games has been a problem of fundamental importance within traditional game theory. Shortly after Von Neumann's seminal work on the existence and effective uniqueness of equilibria in zero-sum games via his minimax theorem [184, 185], Brown and Robinson [28, 155] developed the first learning procedures that converge successfully to equilibrium in zero-sum games in a time-average sense. Unfortunately, this initial glimmer of hope of general positive results connecting Nash equilibria and learning took a step backwards when Shapley [164] established that, even in the case of simple non-zero-sum games learning dynamics, one does not have to converge to Nash equilibria (even in a time-average sense). This result was a strong precursor of the evolution of the field with many, increasingly strong, negative results establishing the lack of any meaningful correlation between Nash equilibria and learning dynamics [48, 65, 90, 92, 160].

In the face of these persistent failures, a natural follow-up direction has been to pursue connections between the time-average of learning dynamics and other weaker game theoretic solutions concepts. The most well known approach of this type has focused on the tightly coupled notions of correlated equilibria (CE) [10] and coarse correlated equilibria (CCE) [123]. These solutions concepts are inspired by the possibility for a mediator to provide correlated advice to each player in regards to which action to pick from a joint distribution that is common knowledge to all players. Extending such concepts to MFGs somehow reduces the gap between Mean-Field Control, where a central coordinator imposes their will on decentralized controllers with no agency, and Mean-Field Games, where decentralized agents traditionally manifest their own will with no coordination mechanism. This bridge also entails the possibility of circumventing Price of Anarchy and Stability issues [136], i.e., achieving performance guarantees better those possibly by Nash equilibria, which is an known issue in Mean Field Games [37], by introducing a way for agents to coordinate their actions. Besides, unlike Nash equilibria, these solution concepts enjoy an inextricable connection to a wide class of learning procedures known as no-regret or regret-minimizing dynamics [80, 82, 158]. Specifically, all regret-minimizing dynamics converge in a time-average sense to coarse correlated equilibria and, vice versa, for any coarse correlated equilibrium in any game, there exists a tuple of regret-minimizing dynamics that converge to it [118]. Such notions of equilibria and related learning mechanisms have surprisingly been so far neglected in the context of Mean-Field games. Only Campi and Fischer [34] as well as DeglInnocenti [52] considered the notion of Mean Field correlated equilibria in both static and dynamic settings. They prove in particular, under suitable conditions and in the fully discrete (State, Action and Time) setting, that N-player CEs converge to Mean-Field CEs as N tends to infinity.

In contrast, this section presents another vision of Mean-Field correlated equilibria (and introduces coarse correlated ones), which we argue is closer to the one considered in the traditional game theory literature [11, 21, 69], as well as more intuitive and easier to manipulate. Yet, we are able to provide equivalence results between our definition and the one in [34] and focus our attention on relevant properties of these equilibria. In particular, we draw connections with no-regret learning in a Mean-Field setting and show that using a Mean Field Correlated Equilibrium policy in an N-player game generates a $O(1/\sqrt{N})$ approximate Correlated Equilibrium under suitable conditions. We study Correlated and Coarse Correlated Equilibria for a large class of Mean Field Games, both in the static and the evolutive settings. Importantly, this more flexible notion of equilibrium allows to capture the efficiency of learning mechanisms in Mean Field Games with several Nash equilibria. Building on the connection with no regret learning, we establish the convergence of classical learning algorithms for Mean Field Games to Coarse Correlated Equilibria in settings where no condition ensuring uniqueness of Nash (monotonicity, contraction property) is available. The three algorithms that we consider are Online Mirror Descent [147], a variant of Fictitious Play [36, 149], and Policy Space Response Oracle (PSRO) [98] (already introduced in [126] and reported

here for the sake of completeness). We summarize the main contributions of this section here:

- We provide the first formulation of coarse correlated equilibrium for Mean Field Games together with a more convenient one for correlated equilibria in this setting. Equivalence between our new formulation and the existing literature [34] is provided.
- We explore properties of our new equilibrium notions and in particular demonstrate that using a Mean-Field (coarse) correlated equilibrium in N-player games provides an $O(1/\sqrt{N})$ approximate Nash equilibrium.

This section is organized as follows. Section 4.1 revisits the notion of correlation device for Symmetric anonymous N-player games and paves the way to the intuitive notion of Mean Field (Coarse) Correlated Equilibrium presented in Section 4.2. Section 4.3 links this more intuitive notion to the existing literature [34] and derives some of its relevant theoretical properties, such as existence conditions and special cases characterization. Finally, Section 4.4 deals with the relationship between N-player games and Mean-Field games. It first establishes how to use a Mean-Field correlated equilibrium in an N-player game and then proves that sequences of N-player (coarse) correlated equilibria converge towards Mean-Field correlated equilibria with N, a property adapted from Campi and Fischer [34]. Moreover, it provides optimality bounds for using Mean-Field (coarse) correlated equilibria in N-player games.

Notations

We introduce here the main notations of this section.
Setting. Given a finite set $\mathcal{Y}$, we denote by $\Delta(\mathcal{Y})$ the set of distributions over $\mathcal{Y}$. To emphasize the difference between the finite and non-finite cases, if $\mathcal{Y}$ is not finite, we write $\mathcal{P}(\mathcal{Y})$ the set of distributions over $\mathcal{Y}$. A game - be it Mean-Field or N-player symmetric-anonymous - is a set $(\mathcal{X}, \mathcal{A}, r, P, \mu_0)$ where $\mathcal{X}$ is the finite set of states, $\mathcal{A}$ is the finite set of actions, $r : \mathcal{X} \times \mathcal{A} \times \Delta(\mathcal{X}) \to \mathbb{R}$ is a reward function, $p : \mathcal{X} \times \mathcal{A} \times \Delta(\mathcal{X}) \to \Delta(\mathcal{X})$ is a state transition function and $\mu_0 \in \Delta(\mathcal{X})$ is an initial state occupancy measure. The dependence of r and P on an element of $\Delta(\mathcal{X})$ captures the interaction between the players. It measures the influence of the full distribution of players over states on the reward and dynamics of each identical player. This assumption considers that all players are *anonymous*, *i.e.* only their state distribution affects others while their identity is irrelevant; and that the game is *symmetric*, since all players share the same reward and dynamic functions. In an N-player game, we denote by $\mathcal{N}$ the set of players. Note that we only consider symmetric-anonymous N-player games, hence we deviate from the traditional vision of considering all players' individual states, to consider that the other players affect one another only through their empirical distribution. Finally, since we consider finite games, we name $\mathcal{T}$ the discrete set of times.

Policy. A policy is a mapping $\bar{\pi} : \mathcal{X} \to \Delta(\mathcal{A})$, where $\bar{\pi}(x, a)$ represents the probability of playing action a while at state x. The set of such policies is denoted $\bar{\Pi}$. We also consider Π the set of deterministic policies, *i.e.* of the form: $\forall x \in \mathcal{X}$, $\exists a \in \mathcal{A}$, $\pi(x) = \delta_a$, or $\pi(x, a') = \mathbf{1}_{\{a'=a\}}$. The set Π of deterministic policies is *finite* and its convex hull is the set $\bar{\Pi}$ of all (stochastic) policies. In an N-player game, we write Π_i the set of policies of player i. If the game is symmetric, we write $\Pi = \Pi_1 = ... = \Pi_N$ the common set of policies available to all players.

Payoff. We name J_i the expected payoff function for player i : $J(\pi_i, \pi_{-i})$ is the expected return for player i when they play policy π_i while the population of all other players play the joint policy π_{-i}. If $\mu_{i,t}$ is the distribution of player i over states at time $t \in \mathcal{T}$, and $r^{\pi}_{i,t}$ is their expected reward vector (One component per state and per time, averaged over actions given their probability of occurrence following π) - both given the actions of other players, then

$$J(\pi_i, \pi_{-i}) \;=\; \sum_{t \in \mathcal{T}} \langle r^{\pi}_{i,t},\, \mu_{i,t} \rangle \,.$$

where the scalar product between two vectors $x, y \in \mathbb{R}^N$ is defined as $\langle x, y \rangle = \sum_{i=1}^{N} x_i y_i$.

Policy swap. Finally, we define the set of policy swap functions

$$\mathcal{U}_{CE} := \{u : \Pi \to \Pi\} \ , \quad \text{and} \quad \mathcal{U}_{CCE} := \{u : \Pi \to \Pi \mid u \text{ constant } \} \tag{4.1}$$

the set of unilateral deviation functions, i.e. the restriction of $\mathcal{U}_{CE}$ to constant functions. Intuitively, policy swaps which are defined over deterministic policies, are functions that shift the probability mass assigned on one policy to another - thereby swapping policies around in a distribution of play. Note that swaps do not need to be bijective, and can for example always return the same policy.

4.1 A Small Detour Through N-Player Games

By construction, Mean-Field Games identify to the limit of symmetric-anonymous N-player games, when N tends to infinity. Correlated and coarse correlated equilibria have been extensively studied in games with finite number of players [3, 11, 21, 22, 61, 69, 111]. Hence, we first ground our intuition and formalism by focusing on the particular case of symmetric-anonymous games with a finite number of players. We derive new expressions for correlated and coarse correlated equilibria for these games, paving the way to their straightforward extension in the Mean-Field setting in Section 4.2.

Considering (coarse) correlated equilibria removes this issue by letting the correlation device choose which joint policy to recommend. Note that a correlation device which only recommends one Nash equilibrium is a correlated equilibrium. thus correlated equilibria may thus be used to solve the equilibrium selection problem.

4.1.1 Notions and Intuitions of Equilibria in N-Player Games

For sake of completeness, we first briefly recall the classical notions of Nash [134], correlated and coarse correlated [10] equilibria in N-player games.

Definition 15 (N-Player Nash Equilibrium)**.** Given $\epsilon > 0$, we define an ϵ-Nash Equilibrium $(\pi_1, \dots, \pi_n) \in \bar{\Pi}_1 \times \cdots \times \bar{\Pi}_N$ as an n-tuple of strategies such that

$$\max_{\pi_i' \in \Pi_i} J_i(\pi_i', \pi_{-i}) - J_i(\pi_i, \pi_{-i}) \leq \epsilon \, , \qquad \forall i \in \mathcal{N}.$$

We will call a Nash *pure* if ever it is deterministic. Otherwise, we will call it *mixed*.

Contrarily to Nash Equilibria, where players choose separately which policy to follow, correlated and coarse correlated equilibria must be implemented with an additional entity atop the game whose only purpose is to coordinate agents' behaviors. It does so by selecting a joint strategy for the full population of players, and then recommends each player their policy within the joint strategy. Each player is aware of their own recommended policy together with the joint distribution over the population, but does not know the recommendation given to every other player: from a joint policy π, a player i only sees π_i.

The goal of the additional entity, termed a *correlation device*, is to render their recommendations stable in the presence of a payoff maximizing player. That is, given a policy recommendation, and given knowledge of the probability distribution over the joint policies recommended by the correlation device, does the player have an incentive to deviate and play something else? If the answer is negative, the correlation device is a correlated equilibrium:

Definition 16 (N-player ϵ-Correlated Equilibrium)**.** Given $\epsilon > 0$, we define an ϵ-Correlated Equilibrium $\rho \in \Delta(\Pi_1 \times \cdots \times \Pi_N)$ as a distribution over joint strategies such that

$$\mathbb{E}_{\pi \sim \rho} \left[J_i(u(\pi_i), \pi_{-i}) - J_i(\pi_i, \pi_{-i}) \right] \leq \epsilon \, , \qquad \forall u \in \mathcal{U}_{CE}, \, i \in \mathcal{N}.$$

Another question, different from and less restrictive than the correlated equilibria's, concerns the player's ability to *a priori* find a fixed deviating policy, independent of their received advice, so that they can improve their payoff without even taking their own recommendation into account. If this question's answer is negative, the correlation device is a coarse correlated equilibrium:

Definition 17 (N-player ϵ-Coarse Correlated Equilibrium)**.** Given $\epsilon > 0$, we define an ϵ-Coarse Correlated Equilibrium $\rho \in \Delta(\Pi_1 \times \cdots \times \Pi_N)$ as a distribution over joint strategies such that

$$\mathbb{E}_{\pi\sim\rho}\left[J_i(u(\pi_i), \pi_{-i}) - J_i(\pi_i, \pi_{-i})\right] \leq \epsilon\,, \qquad \forall u \in \mathcal{U}_{CCE},\, i \in \mathcal{N}.$$

We see here that correlated and coarse correlated equilibria are very similar, only differing by the collection of admissible deviation types $\mathcal{U}_{CE}$ and $\mathcal{U}_{CCE}$ defined in equation 4.1. In particular, any correlated equilibrium is obviously a coarse correlated equilibrium.

We note that the presence of the correlation device helps solve one issue which plagues Nash equilibria in N-player games: the equilibrium selection problem. Indeed, as mentioned above, Nash equilibria are characterized by all players acting in a payoff maximizing manner but without coordination. When several Nash equilibria exist in a game, players must all somehow choose the same Nash equilibrium to receive any individual optimality guarantee.

This formulation is however too general to provide straightforward definitions of these equilibria for Mean-Field games: there is no direct, general way to define a joint strategy over an infinity of unique players, hence neither is there one for distributions over this space. However, Mean-Field games are a particular class of infinite player games, i.e. infinite-player symmetric-anonymous games. In the following section, we provide an equivalent writing of (coarse) correlated equilibria in this setting which naturally scales to the Mean-Field limit.

4.1.2 The Special Case of Symmetric-Anonymous N-Player Games

We start this section with a remark: we will use interchangeably the terms *symmetric-anonymous* and *symmetric*, because all symmetric games are anonymous: indeed, take a symmetric game, a given player i, and the set of permutations σ_i composed of all permutations which do not permutate i. Since the game is symmetric, i's payoff remains identical whatever the permutation, hence i's payoff is only affected by the number of players playing a given strategy, but not by their identity. Since this is true for all players, the game is anonymous. This result is also derived in [76], along with many other properties of symmetric and anonymous games.

In symmetric-anonymous games, on top of all individual policy sets Π_i being identical and equal to Π, the payoff functions must not be impacted by player identities. Namely, all payoff functions J_i are such that, for any permutation $\tau : [1, N] \to [1, N]$, we have

$$J_i(\pi_1, \ldots, \pi_N) = J_{\tau^{-1}(i)}(\pi_{\tau(1)}, \ldots, \pi_{\tau(N)})\,, \qquad \pi = (\pi_1, \ldots, \pi_N) \in \Pi^N.$$

In other words, the reward for a given player i only depends on player i's own policy together with the distribution of policies over the population of all the other players, without any impact from each player identity. This rewrites analogously as follows: the payoff that player i receives when playing π_i only depends on the proportion of other players playing every policy in Π.

We therefore introduce the following concept:

Definition 18 (N-player Population Distribution)**.** The *Population Distribution* of N players playing policies in Π is defined as $\nu_N = \frac{1}{N}\sum_\pi n_\pi \delta_\pi$, where n_π is the number of players playing π, and δ_π is a dirac centered on π. The set of N-player population distributions is written $\Delta_N(\Pi)$.

We will analogously denote $\nu_{-i} \in \Delta_{N-1}(\Pi)$ the distribution over policies in the population of all players except player i. By construction, in symmetric-anonymous N-player games, we can express J_i as a function that is independent of the specific identity of the current player i, of i's policy and other players' policy distribution following

$$J_i(\pi_i, \pi_{-i}) = \mathcal{J}(\pi_i, \nu_{-i}). \tag{4.2}$$

When N players sample their policies from $\Delta_N(\Pi)$, *i.e.* they sample from $\nu_N \in \Delta(\Pi) \cap \Delta_N(\Pi)$ as a distribution, the policy distribution obtained as an outcome of this sample may not match ν anymore. To guarantee that this remains the case, and that no asymmetry exists between players when sampling from members of $\Delta_N(\Pi)$, we define a new notion of sampling:

Definition 19 (Symmetric sampling from $\Delta_N(\Pi)$)**.** When N players sample from $\nu_N \in \Delta_N(\Pi)$, they are *symmetrically* assigned a policy from ν_N such that their population distribution is equal to ν_N. The symmetrical assignment is such that the sampling distribution is invariant to player permutation.

We remark that sampling from $\Delta_N(\Pi)$ is akin to an assignment. This new sampling definition will guarantee that our new correlated equilibrium concept is symmetric.

Finally, we need to define the concept of population recommenders, which recommend different population distributions to the players:

Definition 20 (Population Recommenders)**.** A population recommender ρ is a distribution over population distributions, *i.e.* $\rho \in \Delta(\Delta_N(\Pi))$. A population distribution sampled by a population recommender is also called a *population recommendation*.

With these definitions introduced, we are in a position to rewrite both (C)CE definitions 16 and 17 for a representative player i.

Definition 21 (N-player Symmetric-Anonymous ϵ-(Coarse)-Correlated Equilibrium)**.** We define a symmetric-anonymous ϵ-(coarse)-correlated equilibrium ρ as a distribution in $\Delta(\Delta_N(\Pi))$ such that $\forall i, \forall u \in \mathcal{U}_{(C)CE}$,

$$\mathbb{E}_{\nu\sim\rho,\, \pi_i\sim\nu}\left[\mathcal{J}(u(\pi_i), \nu_{-i}) - \mathcal{J}(\pi_i, \nu_{-i})\right] \leq \epsilon \ .$$

By construction, we observe that the correlating device ρ defined above only samples population distributions $\nu \in \Delta(\Pi)$. Individual players then receive player-symmetric policy recommendations such that their marginal policy distribution is equal to ν in a permutation-invariant way. Hereby, all such correlated equilibria are symmetric and anonymous, hence their names: symmetric-anonymous equilibria. Note here that ν_{-i} is computed independently of the players' policy assignments, it is the result of removing from ν the policy assigned to player i.

We also see that being recommended a given policy does not necessarily imply knowing which ν_{-i} was sampled by ρ: knowledge of ρ only allows one to make estimates about others' expected behavior.

We see below that symmetric-anonymous equilibria are in fact equivalent to standard equilibria (as in Def. 16 or Def. 17) that are symmetric, *i.e.* that are in $\Delta_{sym}(\Pi^N) = \{\nu \in \Delta(\Pi^N) \mid \forall \tau \text{ permutation }, \nu \circ \tau = \nu\}$ the set of distributions over Π^N that are invariant to player permutations.

Theorem 22 (Equilibrium Equivalence)**.** *In symmetric-anonymous N-player games, there is one to one correspondence between symmetric-anonymous ϵ-(C)CE and ϵ-(C)CE with symmetric correlating device, i.e. such that $\rho \in \Delta_{sym}(\Pi^N)$.*

Proof of Theorem 22. Let $\Pi = \Pi_1 = \cdots = \Pi_N$. For $\pi \in \Pi^N$, and $\bar{\rho} \in \Delta(\Pi^N)$ a classical and *symmetric* (coarse) correlated equilibrium, let $\bar{\rho}_{\pi_i}$ be the conditional distribution on Π^{N-1} given player i is recommended policy π_i, and let $\nu_\pi^N = \frac{1}{N}\sum_{j=1}^N \delta_{\pi_j}$ be the empirical population distribution of $\pi \in \Pi^N$. From Equation 4.2, we have that

$$\mathcal{J}(\pi_i, \nu_{\pi_{-i}}^{N-1}) := J_i(\pi_i, \pi_{-i}).$$

and

$$\mathcal{J}(\pi_i, \nu_{\pi_{-i}}^{N-1}) = \mathcal{J}\left(\pi_i, \frac{N}{N-1}(\nu_\pi^N - \frac{1}{N}\delta_{\pi_i})\right),$$

so this quantity only depends on ν_π^N and π_i. Moreover, it will be useful to consider empirical distributions containing a given policy. More precisely, note that, for $\nu^N \in \Delta_N(\Pi)$,

$$\exists \tilde{\pi} \in \Pi^N \text{ s.t. } \tilde{\pi}_i = \pi_i \text{ and } \nu_{\tilde{\pi}}^N = \nu^N \Leftrightarrow \underbrace{\frac{N}{N-1}(\nu^N - \frac{1}{N}\delta_{\pi_i})}_{=:\nu^N_{-\pi_i}} \in \Delta_{N-1}(\Pi), \tag{4.3}$$

meaning that π_i is a point in the support of the empirical distribution ν^N if and only if $\frac{N}{N-1}(\nu^N - \frac{1}{N}\delta_{\pi_i}) \in \Delta_{N-1}(\Pi)$ is an empirical distribution with $N-1$ points. Let us denote by $\Delta_N(\Pi, \pi_i)$ the set of empirical distributions ν^N satisfying the above condition, *i.e.*,

$$\Delta_N(\Pi, \pi_i) = \Big\{\nu^N \in \Delta_N(\Pi) : \frac{N}{N-1}(\nu^N - \frac{1}{N}\rho_{\pi_i}) \in \Delta_{N-1}(\Pi)\Big\}.$$

For simplicity, we denote:

$$\nu^N_{-\pi_i} = \frac{N}{N-1}(\nu^N - \frac{1}{N}\rho_{\pi_i}), \qquad \nu^{N-1}_{+\pi_i} = \frac{N-1}{N}(\nu^{N-1} + \frac{1}{N-1}\rho_{\pi_i}).$$

For $\bar{\rho} \in \Delta(\Pi^N)$, let $\rho \in \Delta(\Delta_N(\Pi))$ be the distribution over empirical distributions induced by $\bar{\rho}$, i.e., for every $\nu^N \in \Delta_N(\Pi)$

$$\rho(\nu^N) = \bar{\rho}\left(\{\tilde{\pi} \in \Pi^N : \nu_{\tilde{\pi}}^N = \nu^N\}\right).$$

Since we assume that $\bar{\rho}$ is symmetric, we can say that for all π,

$$\sum_\pi \bar{\rho}(\pi)\mathcal{J}(u(\pi_i), \nu^{N-1}_{\pi_{-i}}) = \sum_\pi \sum_{\pi_i} \bar{\rho}(\pi)\frac{N_{\pi_i \in \pi}}{N}\mathcal{J}(u(\pi_i), \nu^{N-1}_{\pi_{-i}})$$

where $N_{\pi_i \in \pi}$ is the number of players playing π_i when the joint policy is π. What this means is that if $\bar{\rho}$ recommends π, then it will also recommend all possible permutations thereof, hence the expected payoff for player i when a given policy π is recommended is the same as the expected payoff averaged over all players when π is recommended.

$$\begin{aligned}
\mathbb{E}_{\pi\sim\bar{\rho}}[J_i(u(\pi_i), \pi_{-i})] &= \sum_\pi \bar{\rho}(\pi) J_i(u(\pi_i), \pi_{-i}) \\
&= \sum_\pi \bar{\rho}(\pi)\mathcal{J}(u(\pi_i), \nu^{N-1}_{\pi_{-i}}) \\
&= \sum_\pi \sum_{\pi_i} \bar{\rho}(\pi)\frac{N_{\pi_i}}{N}\mathcal{J}(u(\pi_i), \nu^{N-1}_{\pi_{-i}}) \\
&= \sum_{\nu^N} \sum_{\pi|\nu_\pi^N = \nu^N} \sum_{\pi_i} \bar{\rho}(\pi)\nu^N(\pi_i)\mathcal{J}(u(\pi_i), \nu^N_{-\pi_i}) \\
&= \sum_{\nu^N} \underbrace{\sum_{\pi|\nu_\pi^N = \nu^N} \bar{\rho}(\pi)}_{=\rho(\nu^N)} \sum_{\pi_i} \nu^N(\pi_i)\mathcal{J}(u(\pi_i), \nu^N_{-\pi_i}) \\
&= \sum_{\nu^N} \sum_{\pi_i} \nu^N(\pi_i)\rho(\nu^N)\mathcal{J}(u(\pi_i), \nu^N_{-\pi_i}) \\
&= \mathbb{E}_{\nu^N\sim\rho,\, \pi_i\sim\nu^N}[\mathcal{J}(u(\pi_i), \nu^N_{-i})],
\end{aligned}$$

which concludes the proof.

□

We have introduced the concept of Population Policy Distribution, and we observed that Correlation Devices can be distributions over Population Policy Distributions. Intuitively, the first concept can easily scale to the Mean-Field limit by taking N to infinity in $\Delta_N(\Pi)$, thus becoming $\Delta(\Pi)$: intuitively, the "granularity" of $\Delta_N(\Pi)$ is $\frac{1}{N}$; as N tends to infinity, this "granularity" tends to 0 and $\Delta_N(\Pi)$ is able to represent an increasing amount of members of $\Delta(\Pi)$ - when N is infinite, both sets coincide. The second concept can be transferred from $\Delta(\Delta_N(\Pi))$ to $\mathcal{P}(\Delta(\Pi))$. In the next section, after initially defining the Mean-Field setting of interest and recalling what Mean-Field Nash equilibria are, we define Mean-Field correlated and coarse correlated equilibria in the same spirit. Section 4.3 provides an analogue of Theorem 22 by proving that our new notion of correlated equilibrium is equivalent to the pre-existing ones established by Campi and Fischer [34].

4.2 Notions of Mean Field Equilibrium

We now describe a general setting, which is able to encompass the consideration of both static and dynamic Mean-Field games. The state space is denoted by $\mathcal{X}$. We denote by $\mathcal{T}$ the finite set of times within the game, so that $\mathcal{T}$ simply reduces to a singleton for static games. The set of distribution flows $\Delta(\mathcal{X})^{\mathcal{T}}$ on the state space $\mathcal{X}$ over times in $\mathcal{T}$ is denoted by $\mathcal{M}$.
Whenever every player in the population follows the policy $\pi \in \Pi$, the game generates a Mean-Field flow over $\mathcal{X}$ denoted by $\mu^\pi \in \mathcal{M}$. Formally, μ^π is defined by

$$\mu^\pi_{t+1}(x) = \sum_{x_t \in \mathcal{X}} \sum_{a \in \mathcal{A}} p(x \mid x_t, a, \mu^\pi_t)\, \pi(x_t, a) \mu^\pi_t(x_t) \qquad \forall t \in \mathcal{T}, x \in \mathcal{X}\,,$$

with $\mu^\pi_0 = \mu_0$ a predefined initial state distribution of the population.

Given a Mean-Field flow $\mu \in \mathcal{M}$ of the population, the expected reward of a representative player playing policy $\pi \in \Pi$ is given by

$$J(\pi, \mu) = \sum_{t \in \mathcal{T}} \sum_{x,a} r(x, a, \mu_t) \mu^\pi_t(x) \pi(x, a) \;=\; \sum_{t \in \mathcal{T}} \langle r^\pi(\cdot, \mu^\pi_t), \mu^\pi_t \rangle\,,$$

where μ^π the expected state distribution of policy π when the population follows the Mean-Field flow μ, and $r^\pi(x, \mu) = \sum_a \pi(x, a) r(x, a, \mu)$.

Given a fixed Mean-Field flow $\mu \in \mathcal{M}$, an individual player can maximise their expected return by solving the following Markov Decision Process (MDP) policy optimisation problem

$$\sup_{\pi \in \Pi} J(\pi, \mu)\,. \tag{4.4}$$

Whenever the population of players plays a distribution of strategies $\nu \in \Delta(\Pi)$, the induced Mean-Field flow over the state space $\mathcal{X}$ is denoted by

$$\mu(\nu) \in \mathcal{M}.$$

In the case when the dynamics depend on μ, it is difficult to express $\mu(\nu)$ in closed form, since policies' state distributions will interfere with one another's state distributions, leading to some potentially very strong non-linearities. However, in the μ-independent-dynamics case, $\mu(\nu)$ can be expressed in closed form:

Lemma 23 (Closed-form $\mu(\nu)$)**.** *In the μ-independent-dynamics case,*

$$\mu(\nu) = \sum_{\pi \in \Pi} \nu(\pi) \mu^\pi.$$

By extension, for $\nu \in \Delta(\Pi)$, we write

$$\pi(\nu)$$

the stochastic policy defined by sampling, at every initial state of the game, a policy $\pi \in \Pi$ with probability $\nu(\pi)$, and playing it until the end of the game. This definition ensures that $\mu^{\pi(\nu)} = \mu(\nu)$

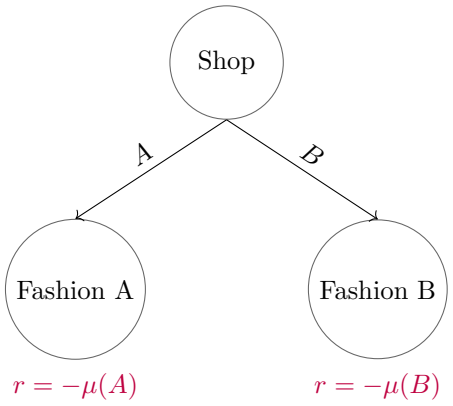


Figure 4.1: Two-Actions Hipster game.

by definition; however, we note that the set $\{\pi \mid \mu^\pi = \mu(\nu)\}$ may have more than one element; in degenerate cases where p does not depend on actions, for example, this set is equal to Π, as all policies have the same state distributions. This definition of $\pi(\nu)$ yields a *unique* policy.

Conversely, given a policy $\bar{\pi} \in \bar{\Pi}$, we write

$$\nu_{\bar{\pi}} \in \Delta(\Pi)$$

for the distribution such that

$$\pi(\nu_{\bar{\pi}}) = \bar{\pi}.$$

In the rest of this section, we will examine different types of game theoretic equilibria. These incorporate a notion of deviation: an equilibrium is only stable if no player has an incentive to deviate from its recommendations. These deviations are considered from the point of view of *all* players. However, since all players are identical, it is enough to make sure that a given, randomly chosen player never has an incentive to deviate. If that player has no incentive to change behavior, then neither does the population. We will refer to this player as the *representative player*.

4.2.1 Mean Field Nash Equilibrium

The literature on Mean Field Games mostly (and almost only) focused so far on the notion of Nash equilibrium between the infinite number of agents within the population. As a generalization of Definition 15, it is naturally defined as follows:

Definition 22 (Mean Field Nash Equilibrium, MFE)**.** Given $\epsilon > 0$, a policy $\bar{\pi} \in \bar{\Pi}$ is an **ϵ-Mean Field Nash Equilibrium** whenever

$$\sup_{\pi' \in \bar{\Pi}} J(\pi', \mu^{\bar{\pi}}) \quad \leq \quad J(\bar{\pi}, \mu^{\bar{\pi}}) + \epsilon.$$

It is a *Mean Field Nash Equilibrium* whenever the previous relation holds for $\epsilon = 0$. A Nash equilibrium is said to be *pure* if it is deterministic.

Example 6 (Two-Actions Hipster game)**.** *We give in Figure 4.1 an example of reward function in the Two-Actions Hipster Game: the goal for each player is to stand out from their peers by choosing the clothing item which is least frequent within the population. At a shop, agents choose either item A or item B, and are penalized for the non-uniqueness of their choice: if all agents choose Fashion A, Fashion B will grant the highest reward, and conversely. In this simplistic game, there is no pure Nash equilibrium, but only one mixed Nash (Agents choose Fashion A or B with probability* $\frac{1}{2}$*).*

One of the most prominent properties of Mean Field Nash equilibria relies on their strong connection with equilibria in N player games. We will later, in Section 4.4, explain in detail how one can use Mean-Field equilibria in N-player games. This process is at the core of the usefulness of Mean-Field games, since plugging a Mean Field Nash equilibrium in an N-player game yields an $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-approximate Nash equilibrium, which is known in the continuous time and continuous space setting in e.g. [35, 86], and which we prove in this chapter for discrete games.

Nevertheless, while the existence of Nash equilibria is a very straightforward property for MFGs to have, with clear and arguably non-restrictive conditions, deriving the uniqueness of such equilibria in general is a difficult and tedious task. One possible approach relies on additional strong Lipschitz conditions leading to a contracting mapping operator [86]. Alternatively, the so-called monotonicity condition introduced in [102] intuitively provides players the incentive to behave differently than the full population and ensures uniqueness of the Nash Equilibrium. Whenever this well established condition is not satisfied, uniqueness of Mean Field Nash equilibrium can be hard to enforce. A natural example for this is the converse of the Hipster game (presented in Figure 4.1) as described below.

Example 7 (Suits Game)**.** *In the Suit game, rewards are inverted compared to the Hipster game (players are incentivized to act similarly to others). This game does not satisfy the monotonicity condition [102] and has 3 Nash equilibria: all-in on Fashion A, B; or 50% on each.*

When the MFE is not unique, one possible option is to help the players synchronize using an extraneous noise or signal. Restoring uniqueness of MFE via the addition of vanishing common noise has been observed in [54]. Alternatively, the addition of a common signal sent to the full population naturally calls for notions of correlated or coarse correlated equilibria [10, 11]. With the exception of [34, 52], Nash equilibria are surprisingly the only type of equilibrium considered in the MFG literature. This is in stark contrast with the literature on N-player games, where weaker notions of equilibria are well established and understood [3, 11, 21, 22, 61, 69, 120, 121]. Specifically, it is understood that there exists a tight correspondence between no-regret dynamics and coarse correlated equilibria [118]. Moreover, worst case analysis for Nash equilibria can sometimes automatically be extended without any further degradation of performance to worst case (coarse) correlated equilibria via what is known as robust Price of Anarchy analysis [157].

4.2.2 Intuition on Correlation Device and Correlated Equilibria

We are now in position to generalize the concepts of correlation device and (coarse) correlated equilibrium to the Mean-Field setting by building on new formulations derived in Section 4.1. Before doing so, let first provide relevant intuitions for these new concepts and facilitate their interpretation.

Correlation Device. A correlation device makes a single policy recommendation to each player in the game. It coordinates the population's actions. In the well-known traffic lights example[2], the correlation device sets the lights' colors, and lets agents (cars) decide whether to follow or not the lights' signals.

Coarse-Correlated Equilibrium - Agent's perspective. From the perspective of the agent, we can imagine that the correlation device is a mediator who is partially aligned to the agent's interests and has a bird's eye view of what the population is doing. In a coarse correlated equilibrium, the agent has two choices: either delegate all decisions to the mediator - despite the partial misalignment -, or take its own decisions, without the mediator's knowledge of what the rest of the population will be doing. If the agent has a larger incentive to use the services of the mediator on average, then the mediator's recommendations may be said to be a coarse correlated equilibrium.

[2]In a hypothetical intersection where traffic laws would not hold.

Correlated Equilibrium - Agent's perspective. Keeping the mediator's analogy, in a correlated equilibrium situation, the agent has two choices: accept the mediator's suggested course of action, or refuse it and choose their own course. This case differs from the coarse correlated case by the fact that here, the agent sees which course of action the mediator has prepared, and, from it, can estimate what the other agents may be recommended by the mediator. Having more information - but not as much information as the mediator -, the agent may therefore take better-informed decisions. However, if despite this, the agent prefers to follow the mediator's suggestion, then the mediator's recommendations may be said to be a correlated equilibrium[3].

Whenever correlation devices are discrete probability distributions, a visualization of how correlation devices operate, for the homogeneous (only one recommended policy to the population) and non-homogeneous (heterogeneous, several deterministic policies may be recommended at once) cases, are respectively available in Figures 4.3 and 4.2.

4.2.3 Mean-Field Correlation Device

Whenever several Nash equilibria exist, an equilibrium selection problem arises: the population needs more guidance in order to be able to coordinate and synchronize. As noted before, in N-player games, the notions of correlated and coarse correlated equilibria bypass this issue through the use of a correlation device, which provides a signal allowing the population to synchronize; and so do they in Mean-Field games.

Definition 23 (Population distribution/recommendation)**.** We introduce the following.

- A **population distribution, or population recommendation** $\nu \in \Delta(\Pi)$ is a distribution over the set of policies Π;
- Given a population distribution $\nu \in \Delta(\Pi)$, each player receives an **individual recommendation** $\pi \in \Pi$ uniformly sampled from ν, so that the distribution of all individual recommendations over the population is ν.

As detailed in Section 4.2.4 below, correlated equilibria encompass an information asymmetry component: while the recommender knows the full population recommendation, the players - the recommendees - only have access to their own recommendation, which can allow for complex cooperative behavior. Nevertheless, all players are also aware of the possible population distributions, together with their probability of occurrences. This information is contained into what we call correlation devices, whose definition in the Mean-Field setting is as follows.

Definition 24 (Correlation device)**.** A **correlation device** is a distribution ρ over $\Delta(\Pi)$. It encapsulates the possible population recommendations given to the population - we denote $\mathcal{P}(\Delta(\Pi))$ the set of correlation devices.

A Mean-Field correlation device is a distribution over population recommendations that synchronizes all individuals in the population. Its structure is presented in Figure 4.2. The exogenous recommender picks a realization of a random variable with distribution $\rho \in \mathcal{P}(\Delta(\Pi))$ and gives each player its own individual recommendation $\pi \in \Pi$ as a signal. All players know ρ together with their own individual recommendation $\pi \in \Pi$, but do not have access to the population recommendation $\nu \in \Delta(\Pi)$ sampled by the recommender. Whenever a player receives $\pi \in \Pi$ as recommendation, their belief about the possible population distributions shifts to $\rho(\cdot \mid \pi)$ defined by: for $\nu \in \Delta(\Pi)$,

$$d\rho(\nu \mid \pi) \;:=\; \frac{\nu(\pi) d\rho(\nu)}{\int_{\nu' \in \Delta(\Pi)} \nu'(\pi) d\rho(\nu')}. \tag{4.5}$$

[3]On a philosophical note, a striking relationship between the concept of Correlated Equilibrium and that of Manager Efficiency developed by MacIntyre [108]. We hope to see such philosophical links developed more thoroughly in the future.

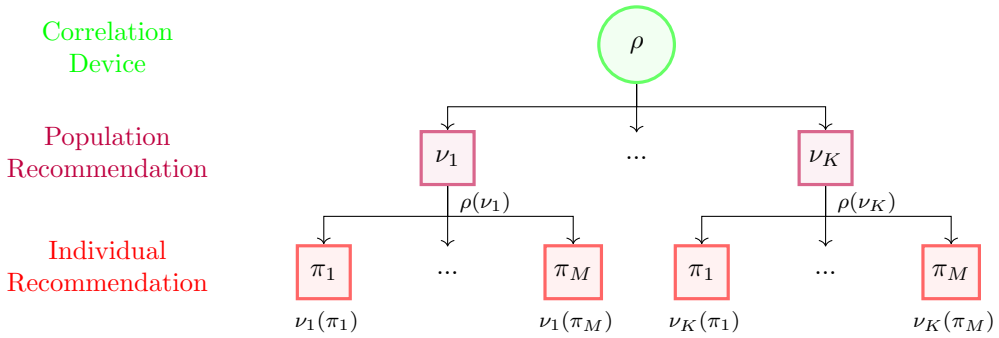


Figure 4.2: Structure of a discrete Mean-Field correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$.

This conditional distribution goes in pair with the distribution ρ_Π over Π induced by the correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$ which is defined by

$$\rho_\Pi(\pi) := \int_{\nu \in \Delta(\Pi)} \nu(\pi) d\rho(\nu) \ , \tag{4.6}$$

so that

$$d\rho(\nu) = \sum_{\pi \in \Pi} \rho_\Pi(\pi) d\rho(\nu \mid \pi).$$

By our definition, agents never observe ν : the whole stochasticity of the process resides in the centralized instance, which samples both ν and a policy from ν for each agent. However, we could also imagine that ρ would send ν to each agent, and lets agents sample their policy from ν for the duration of an episode. In this case, agents all play the same policy $\pi(\nu) \in \bar{\Pi}$, and all know what the other agents are playing. We call such ρ, which communicate ν to the players, ***homogeneous*** *correlation devices*. We note that ρ samples ν and transfers it to players, which then play $\pi(\nu)$. We can therefore view ρ as a distribution over the possible values of $\pi(\nu)$, *i.e.* over $\bar{\Pi}$. We formalize this notion:

Definition 25 (Homogeneous correlation device)**.** A **homogeneous correlation device** $\rho_h \in \mathcal{P}(\bar{\Pi})$ is a special type of correlation device that samples stochastic policies, and only recommends one stochastic policy to all players in the population.

Here is an example of a homogeneous correlation device.

Example 8. *Let us consider again the Suits Game, defined in Example 7, in which each player is incentivized to pick a fashion well represented in the population. A correlation device alternatively recommending all players to choose Fashion A and Fashion B (*i.e. *50% of the time, it recommends Fashion A to all players; 50% of the time, Fashion B) is a homogeneous correlation device, that happens to generate a Mean-Field correlated equilibrium, as discussed in the next section.*

Intuitively, since all players know what other players are playing, some homogeneous equilibria should find themselves very restricted. We show that this is indeed the case in Section 4.3.5.

4.2.4 Mean-Field Correlated Equilibrium

We now turn to the definition of correlated equilibrium for Mean-Field games, which is built as a natural extension to the one considered in N-player games. We define Mean-Field correlated equilibria similarly to their anonymous N-player version derived in Definition 21 above.

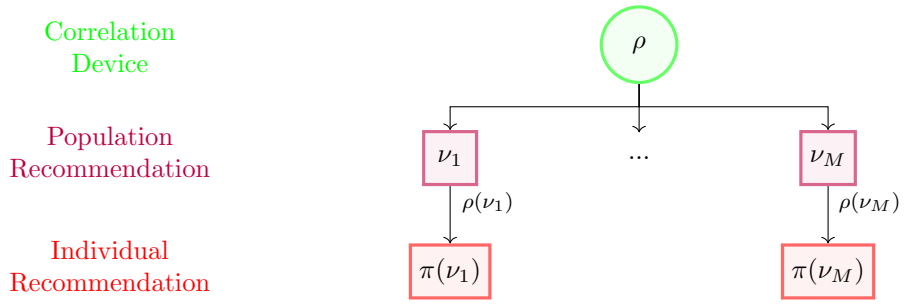


Figure 4.3: Structure of a discrete homogeneous Mean-Field correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$.

Definition 26 (Mean Field Correlated Equilibrium, MFCE)**.** Given $\epsilon > 0$, a correlation device ρ is an ϵ**- Mean Field Correlated Equilibrium** if, $\forall u \in \mathcal{U}_{CE}$

$$\mathbb{E}_{\nu\sim\rho,\,\pi\sim\nu}\left[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))\right] \leq \epsilon\,. \tag{4.7}$$

It is called **Mean Field Correlated Equilibrium** whenever the previous relation holds for $\epsilon = 0$.

This definition of Mean Field Correlated equilibrium aligns naturally with the one developed in the Game theory literature [11]. Besides, we will verify in Section 4.3.4 below that it also connects in an elegant fashion to the one introduced recently in [34].

The next result provides a geometric property of the set of Mean-Field correlated equilibria.

Proposition 24. *For all $\epsilon \geq 0$, the set of ϵ-MFCEs is convex.*

Proof. Let $\epsilon \geq 0$, ρ_0, ρ_1 be two ϵ-MFCE. Let $0 \leq \alpha \leq 1$ and let ρ_α be the barycentric correlation device $\alpha\rho_0 + (1-\alpha)\rho_1 \in \mathcal{P}(\Delta(\Pi))$.

Let $u \in \mathcal{U}_{CE}$.

$$\begin{aligned}
&\mathbb{E}_{\nu\sim\rho_\alpha,\,\pi\sim\nu}[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))]\\
=&\alpha\mathbb{E}_{\nu\sim\rho_0,\,\pi\sim\nu}[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))] + (1-\alpha)\mathbb{E}_{\nu\sim\rho_0,\,\pi\sim\nu}[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))]\\
\leq&\epsilon
\end{aligned}$$

□

The set of correlated equilibria is behaving as we expect. We now turn towards the set of *homogeneous correlated equilibria*. There is a significant information difference between correlated equilibria and homogeneous correlated equilibria: while the former's agents only observe their own recommendation, the latter's observe the full population recommendation. This means that the deviations they consider will have more granularity than $\mathcal{U}_{CE}$: each population recommendation will correspond to one specific deviation, *i.e.* homogeneous correlated equilibria's deviation functions are $\mathcal{U} = \{u \mid u : \bar{\Pi} \to \bar{\Pi}\}$. This concept can be linked with the notion of Φ-regret introduced in Piliouras et al. [152]. We formally define homogeneous correlated equilibria, given their deviation set $\mathcal{U}^h_{CE} = \{u \mid u : \bar{\Pi} \to \bar{\Pi}\}$,

Definition 27 (Homogeneous Mean Field Correlated Equilibrium, MFCE)**.** Given $\epsilon > 0$, a homogeneous correlation device ρ is an ϵ**- Homogeneous Mean Field Correlated Equilibrium** if,

$$\mathbb{E}_{\nu\sim\rho}\left[J(u(\pi(\nu)),\mu(\nu)) - J(\pi(\nu),\mu(\nu))\right] \leq \epsilon \qquad \forall u \in \mathcal{U}^h_{CE}. \tag{4.8}$$

It is called **Homogeneous Mean Field Correlated Equilibrium** whenever the previous relation holds for $\epsilon = 0$.

4.2.5 Mean-Field Coarse Correlated Equilibrium

In N-player games, computing Correlated Equilibria can be very expensive [121]. Hereby, another set of equilibria, wider and easier to compute, was introduced in this setting: coarse correlated equilibria. Up to our knowledge, such a notion has never been studied in the framework of Mean-Field Games. A coarse correlated equilibrium is a weaker notion of equilibrium, where each player may only choose to deviate from their recommendation before having observed it - though players are still assumed to have knowledge of the correlation device's behavior $\rho \in \mathcal{P}(\Delta(\Pi))$. This larger class of equilibria contains correlated equilibria and is more easily reachable by classical learning algorithms, as will be discussed in Section 5.2.

Definition 28 (Mean Field Coarse Correlated equilibrium, MFCCE)**.** Given $\epsilon > 0$, a correlation device ρ is an **ϵ-Mean-Field Coarse Correlated Equilibrium** if

$$\mathbb{E}_{\pi\sim\nu,\nu\sim\rho}\left[J\left(u(\pi),\mu(\nu)\right) - J\left(\pi,\mu(\nu)\right)\right] \quad \leq \quad \epsilon\,, \quad \forall u \in \mathcal{U}_{CCE}\,. \tag{4.9}$$

It is a *Mean-Field Coarse Correlated Equilibrium* whenever the previous equation holds for $\epsilon = 0$.

Recall that $\mathcal{U}_{CCE}$ denotes the set constant deviations over Π, i.e. the mappings from Π to Π which a fixed constant policy $\pi \in \Pi$. MFCCEs can also be defined in an alternative way.

Proposition 25 (MFCCE characterization using best-responses)**.** *A correlation device ρ is an ϵ-MFCCE if and only if,*

$$\sup_{\pi'\in\Pi} \mathbb{E}_{\nu\sim\rho}[J(\pi',\mu(\nu))] \quad \leq \quad \mathbb{E}_{\pi\sim\nu,\nu\sim\rho}\left[J(\pi,\mu(\nu))\right] + \epsilon$$

Proof. The proof follows from identifying Π with $\{u(\pi),\, u \in \mathcal{U}_{CCE} \text{ and } \pi \in \Pi\}$. □

Proposition 26 (MFCEs are MFCCEs)**.** *The set of ϵ-MFCE is included in the set of ϵ-MFCCE.*

Proof. This property is a direct implication from the definition of MFCEs and Proposition 25, when it is noted that $\mathcal{U}_{CCE} \subseteq \mathcal{U}_{CE}$. □

Inclusions between the sets of Nash, correlated and coarse correlated equilibria are represented in Figure 4.4. Besides, MFCCEs being much less restrictive than MFCEs, both sets rarely coincide. However, they can consistently coincide in very small games.

Proposition 27. *In two-action one-state Mean-Field games, the set of MFCEs and MFCCEs are equal.*

Proof. We already know that the set of MFCEs is included in the set of MFCCEs. The reverse inclusion is proven by observing that in this particular setting, unilateral deviation to either action is equivalent to deviating when being recommended the other action - thus being optimal for unilateral deviations is equivalent to being optimal for per-action deviations.

Note that this does not imply that $\mathcal{U}_{CE} = \mathcal{U}_{CCE}$ - indeed, members of $\mathcal{U}_{CE}$ which switch both policies at the same time can not be members of $\mathcal{U}_{CCE}$. □

We note that this does not mean that $U_{CE} = U_{CCE}$ in these settings. Indeed, a deviation function which switches both actions is a member of U_{CE} but not of U_{CCE}. However, if a payoff stands to be gained by deviating from one action to another, then it means that the other action is more profitable in general, and thus that only unilateral deviations towards it matter.

Just like MFCEs, the set of MFCCEs is also convex:

Proposition 28. *For all $\epsilon \geq 0$, the set of ϵ-MFCCE is convex.*

Proof. Similar to the one of Proposition 24. □

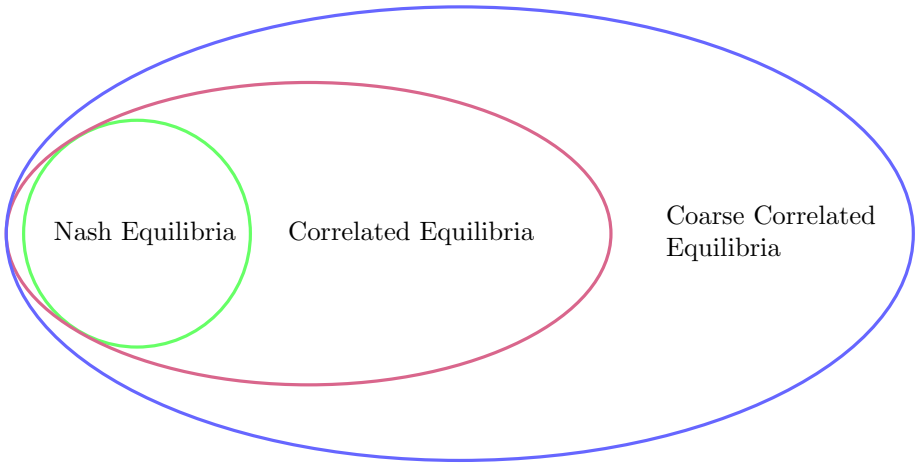


Figure 4.4: Visualization of the typical inclusion relationships between equilibrium sets.

The definition of a homogeneous coarse correlated equilibrium is similar to that of a correlated equilibrium: indeed, coarse correlated equilibria deviate before receiving any play information. More formally, with $\mathcal{U}^h_{CCE} = \{u \mid u : \bar{\Pi} \to \bar{\Pi},\ u \text{ constant}\}$ their deviation set,

Definition 29 (Mean Field Coarse Correlated equilibrium, MFCCE)**.** Given $\epsilon > 0$, a homogeneous correlation device ρ is an **ϵ-Homogeneous Mean-Field Coarse Correlated Equilibrium** if

$$\mathbb{E}_{\pi\sim\nu,\nu\sim\rho}\left[J\left(u(\pi),\mu(\nu)\right) - J\left(\pi,\mu(\nu)\right)\right] \quad \leq \quad \epsilon\,, \quad \forall u \in \mathcal{U}^h_{CCE}\,. \tag{4.10}$$

It is a *Homogeneous Mean-Field Coarse Correlated Equilibrium* whenever the previous equation holds for $\epsilon = 0$.

4.2.6 Equilibrium Sets Visualization in a Toy Example

This section aims to highlight how vast the set of correlated equilibria can be in comparison to the set of Nash equilibria, and more strikingly how vast the set of coarse correlated equilibria is compared to the set of correlated equilibria. In a word, we illustrate the assertion depicted in Figure 4.4:

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

and evaluate the size of these sets in a simple game.

Example 9. *Let consider the following 3-actions (A, B and C) static Mean-Field Dominated-Action game:*

$$r(A,\mu) = \mu(A) + \mu(C)\,, \qquad r(B,\mu) = \mu(B)\,, \qquad r(C,\mu) = \mu(A) + \mu(C) - 0.05\,,$$

where $\mu(X)$ abusively denotes the proportion of players picking action X in the population (i.e. the state of a player reduces to their action). A visualization of its Mean-Field Nash, correlated and coarse correlated equilibria is provided in Figure 4.5.

In general, visualizing the sets of correlated equilibria is difficult. Indeed, each correlated equilibrium is a distribution over distribution of policies. Therefore, a correlated equilibrium is in general composed of several different mixed policies at once. It is easy to see how to visualize one such equilibrium, but less obvious how to visualize their set, especially when the number of such mixed policies may be infinite. However, in our example, one of the three available actions is dominated: whenever an agent is recommended to play C, they know that they should play A

instead! Correlated equilibria are therefore restricted to recommending either A or B. We know that any mixture between homogeneously recommending A and B to the population yields a CE, so that the set of CEs is the straight line between A and B in Figure 4.5.

Visualizing the set of coarse correlated equilibria is much harder, even more so in this simple game. Indeed, one can recommend action C homogeneously and still get many coarse correlated equilibria, so we can not use the simplifying assumption used for CEs. We choose to restrict to the set of homogeneous CCEs, more precisely, we represent (α, β, γ) such that $\rho = \alpha\delta_A + \beta\delta_B + \gamma\delta_C$ is a CCE. We observe in Figure 4.5 that the set of CCEs is represented by a very large triangle in the simplex, so that the correlation device can recommend the dominated action C. More strikingly, the set of CCEs reveals to be significantly larger than the set of CEs and Nash equilibria. Keeping this in mind is important for understanding their existence relationships (there exists games where CCEs exist, but CEs do not, for example), but also that the Price of Stability of CCEs is greater than or equal to CEs', and their Price of Anarchy is lower than or equal to CE's, sometimes by a great factor.

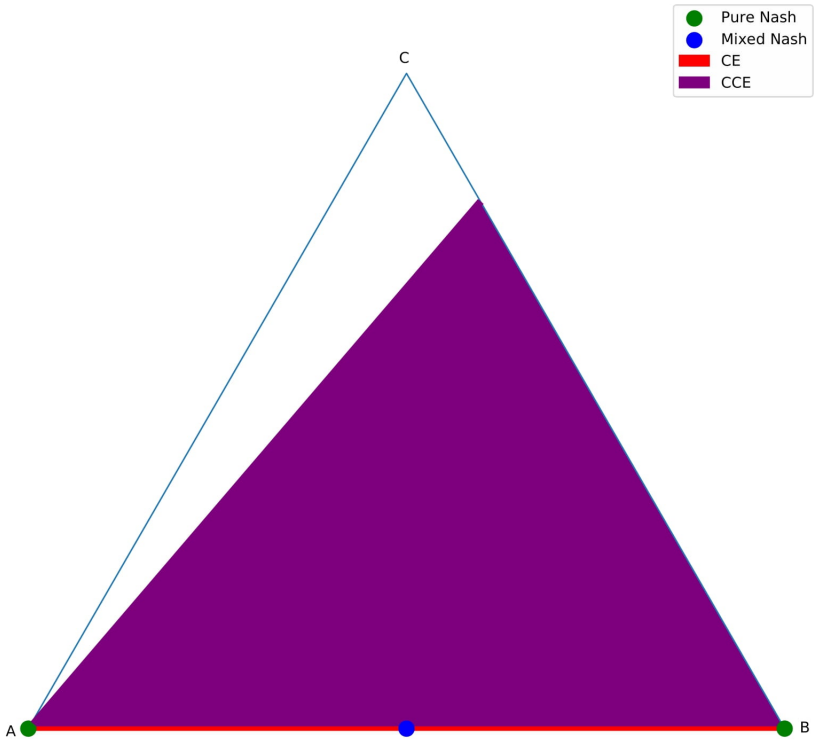


Figure 4.5: Visualization of Mean-Field Equilibria on the Dominated-Action game.

4.3 Properties of Mean Field (Coarse) Correlated Equilibria

In this section, we investigate several properties of our (coarse) correlated equilibrium framework. First, in Section 4.3.1, we detail relationships between Nash equilibria and (coarse) correlated equilibria. Then, in Section 4.3.2, we detail existence conditions for Mean-Field (coarse) correlated equilibria, and find surprising situations where *no (coarse) correlated equilibrium exists*. This is mitigated by the existence, for all $\epsilon > 0$, of ϵ-(coarse) correlated equilibria. Then, in Section 4.3.4,

we establish equivalence between our notion of correlated equilibrium and the one presented by Campi and Fischer [34], thereby inheriting all their asymptotic properties. Finally, in Section 4.3.5, we characterize special properties of homogeneous Mean-Field correlated equilibria.

4.3.1 Relationship Between (Coarse) Correlated Equilibria and Nash Equilibria

In 2-player zero-sum games, correlated equilibria are strongly linked to Nash equilibria: their marginalizations are Nash equilibria; a correlation device recommending a Nash equilibrium is also a correlated equilibrium, and a (coarse) correlated equilibrium which only recommends one (possibly stochastic) joint policy actually recommends a Nash equilibrium !

Mirroring these statements, we first show how any ϵ-Nash equilibrium can be transformed into an ϵ-correlated equilibrium; then that, given any ϵ-correlated equilibrium recommending only one ν, $\pi(\nu)$ is an ϵ-Nash equilibrium. Finally, we analyze the question of (coarse) correlated equilibrium marginalizations, defining what they exactly are, when they exist, and conditions for them to be Nash equilibria.

From Nash Equilibria to Correlated Equilibria

We first start by showing how one can derive Correlated equilibria from Nash equilibria.

Proposition 29 (Nash-derived Correlated Equilibrium)**.** *Every ϵ-Nash equilibrium can be transformed into a Correlated Equilibrium.*

Proof. Let $\pi^* \in \bar{\Pi}$ be a Nash equilibrium. We write $\nu^* = \nu_{\pi^*}$ for conciseness, and take $\rho = \delta_{\nu^*}$.

ρ is an ϵ-correlated equilibrium: if there existed $u \in \mathcal{U}_{CE}$ such that

$$J(u(\pi(\nu^*)), \mu(\nu^*)) - J(\pi(\nu^*), \mu(\nu^*)) > \epsilon$$

then, since $\pi(\nu^*) = \pi^*$ and $\mu(\nu^*) = \mu^{\pi^*}$, this would imply that $u(\pi^*)$ is a policy which has higher value against the Nash than the Nash policy π^* plus ϵ, which is strictly impossible by definition.

Therefore every ϵ-Nash equilibrium can be transformed into an ϵ-correlated equilibrium. □

From Coarse Correlated Equilibria to Nash Equilibria

We now examine the converse of the above property - when can we extract an ϵ-Nash equilibrium from an ϵ-correlated equilibrium ? We show that this is at least possible when the correlated equilibrium is a single Dirac:

Proposition 30 (Coarse correlated equilibrium-derived Nash equilibrium)**.** *Assume $\rho = \delta_\nu$, with $\nu \in \Delta(\Pi)$, is an ϵ-coarse correlated equilibrium. Then $\pi(\nu)$ is an ϵ-Nash equilibrium.*

Proof. We write the optimality condition of ρ for all $u \in \mathcal{U}_{CCE}$:

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}[J(u(\pi), \mu(\nu)) - J(\pi, \mu(\nu))] \leq \epsilon,$$

$$\mathbb{E}_{\pi\sim\nu}[J(u(\pi), \mu(\nu)) - J(\pi, \mu(\nu))] \leq \epsilon,$$

i.e., $\forall \pi' \in \Pi$,

$$J(\pi', \mu(\nu)) - \mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(\nu))] \leq \epsilon.$$

Finally, we note that $\mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(\nu))] = J(\pi(\nu), \mu(\nu))$, which concludes the proof:

$$J(\pi', \mu(\nu)) - J(\pi(\nu), \mu(\nu)) \leq \epsilon,$$

i.e. $\pi(\nu)$ is an ϵ-Nash equilibrium. □

We also show that, in certain classes of games, the marginalization - defined in Definition 30 - of an ϵ-Mean-Field coarse-correlated equilibrium yields an ϵ-Nash equilibrium

We first define properly what the marginalization of a correlation device is:

Definition 30 (Correlation Device Marginalization)**.** The marginalization $\hat{\pi}$ of a correlation device ρ is defined as the policy whose distribution is equal to $\int_\nu \mu(\nu) d\rho(\nu)$.

Note that it always exists when the dynamics do not depend on the distribution:

Proposition 31 (Existence of the marginalization)**.** *In games where the dynamics do not depend on the mean field flow, the marginalization of a correlation device always exists, and is equal to*

$$\hat{\pi}_t(s,a) = \sum_{\pi \in \Pi} \frac{\int_\nu \nu(\pi) \mu_t^\pi(s) d\rho(\nu)}{\sum_{\pi' \in \Pi} \int_{\nu'} \nu'(\pi') \mu_t^{\pi'}(s) d\rho(\nu')} \pi(s,a).$$

Proof. Let first write the distribution evolution equation for $\hat{\pi}$:

$$\mu_{t+1}^{\hat{\pi}}(x) = \sum_{x_t, a} p(x \mid x_t, a) \hat{\pi}(x_t, a) \mu_t^{\hat{\pi}}(x_t)\,.$$

We prove by induction that $\mu_t^{\hat{\pi}} = \int_\nu \mu_t(\nu) d\rho(\nu)$ for all t.

The result holds for $t = 0$ since μ_0 is fixed. If this is true for t, then

$$\begin{aligned}
\mu_{t+1}^{\hat{\pi}}(x) &= \sum_{x_t, a} p(x \mid x_t, a) \hat{\pi}(x_t, a) \mu_t^{\hat{\pi}}(x_t) \\
&= \sum_{x_t, a} p(x \mid x_t, a) \int_{\nu'} \int_\nu \sum_\pi \frac{\nu(\pi) \mu_t^\pi(x_t) d\rho(\nu)}{\int_{\nu'} \mu_t(\nu') d\rho(\nu')} \mu_t(\nu') \pi(x_t, a) d\rho(\nu') \\
&= \int_\nu \sum_\pi \nu(\pi) \underbrace{\sum_{x_t, a} p(x \mid x_t, a) \mu_t^\pi(x_t) \pi(x_t, a)}_{= \mu_{t+1}^\pi(x)} d\rho(\nu) \\
&= \int_\nu \mu_{t+1}(\nu)(x) d\rho(\nu)\,,
\end{aligned}$$

which concludes the induction argument. □

Finally, we will need to define what monotonicity, introduced by Lasry and Lions [102] is:

Definition 31 (Monotonicity)**.** A mean field game is said to be monotonic if

$$\langle \mu - \mu', r(\cdot, \mu) - r(\cdot, \mu') \rangle \leq 0\,, \forall \mu, \mu' \in \mathcal{M}\,.$$

We can now present cases where we can link the marginalization of a coarse correlated equilibrium with its optimality as a Nash equilibrium:

Proposition 32. *In monotonic games where the reward function is affine with respect to μ, the marginalization of an ϵ-Mean-Field-coarse correlated equilibrium, if it exists, is a 2ϵ-Mean-Field-Nash-equilibrium.*

Proof. Let ρ be an ϵ-MFCCE, and $\hat{\pi}$ its marginalization. Let first observe that the monotonicity property implies:

$$\langle \mu - \mu', r(\cdot, \mu) \rangle \leq \langle \mu - \mu', r(\cdot, \mu') \rangle\,, \forall \mu, \mu' \in \mathcal{M}\,. \tag{4.11}$$

From there, we compute

$$\begin{aligned}
J(\pi, \mu^{\hat{\pi}}) - J(\hat{\pi}, \mu^{\hat{\pi}}) &= \langle \mu^{\pi} - \mu^{\hat{\pi}}, r(\cdot, \mu^{\hat{\pi}}) \rangle \\
&= \sum_{\nu} \sum_{\nu'} \rho(\nu)\rho(\nu') \langle \mu^{\pi} - \mu(\nu), r\left(\cdot, \mu(\nu')\right) \rangle \\
&= \sum_{\nu} \sum_{\nu'} \rho(\nu)\rho(\nu') \left(\langle \mu^{\pi} - \mu(\nu'), r\left(\cdot, \mu(\nu')\right) \rangle + \langle \mu(\nu') - \mu(\nu), r\left(\cdot, \mu(\nu')\right) \rangle \right) \\
&\leq \sum_{\nu} \sum_{\nu'} \rho(\nu)\rho(\nu') \left(\epsilon + \langle \mu(\nu') - \mu(\nu), r\left(\cdot, \mu(\nu')\right) \rangle \right) \\
&\leq \sum_{\nu} \sum_{\nu'} \rho(\nu)\rho(\nu') \left(\epsilon + \langle \mu(\nu') - \mu(\nu), r\left(\cdot, \mu(\nu)\right) \rangle \right) \\
&\leq 2\epsilon
\end{aligned}$$

where the second line comes from the affine character of r with respect to μ, and $\hat{\pi}$ being the marginalization of ρ; the third and fifth lines come from ρ being ϵ-optimal, and the fourth line comes from Equation 4.11. □

Remark 3 (Translation-invariance)**.** *We note that the above property also holds if a state-independent dependency on μ is added to the reward function.*

Remark 4 (Extension to ϵ-monotonicity)**.** *If the game is ϵ'-quasi-monotonic, i.e.*

$$\langle \mu - \mu', r(\cdot, \mu) - r(\cdot, \mu') \rangle \leq \epsilon' \, , \forall \mu, \mu' \in \mathcal{M} \, ,$$

then the marginalization of an ϵ-MFCCE, if it exists, is a $(2\epsilon + \epsilon')$-MFE.

Remark 5 (On the non existence of marginalization in distribution-dependent settings)**.** *Consider the hole-trap game depicted in Figure 4.6. In this game, one initially chooses between going left or right. Once in the Left or Right node, the next state does not depend on the players' actions anymore: if every player is in the current node, then it transitions to its + version (Left+ or Right+), otherwise all players are sent to the hole.*

Taking a reward structure which makes Left+ and Right+ equivalent, and the Hole node very penalizing, we can take a Mean-Field Coarse Correlated Equilibrium which alternatively selects between Left and Right 50% of the time.

Its marginalized policy is a policy for which 50% of players end up in Left+ and 50% of players end up in Right+. However, this is strictly impossible, as this requires that 50% of players be on the Left and Right nodes, which would automatically send all players to the hole, and none to Right+ and Left+. The marginalization of this correlated equilibrium is therefore impossible.

4.3.2 Existence of (Coarse) Correlated Equilibria

We have not yet established conditions for correlated equilibria to exist. A set of conditions can be derived immediately from the fact that Nash equilibria can be used as correlated equilibria, as we proved in Proposition 29. Existence conditions for Nash equilibria, namely, continuity of the reward and dynamics functions with respect to μ, hence also imply existence of correlated equilibria. Perhaps surprisingly, we find that the famous result derived by Hart and Schmeidler [81] that correlated equilibria (and therefore coarse correlated equilibria) exist in *all finite N-player games* (*i.e.* N-player games with finite $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{T}$ but not necessarily with continuous reward and or dynamic functions) does not hold in Mean-Field games: Example 10 shows a game where no exact correlated equilibrium exists. We summarize the existence relationships between different Mean-Field equilibria in Figure 4.7, and visually represent them in Figure 4.4.

Remark 6. *Note that deriving a Mean-Field version of Hart and Schmeidler [81]'s proof of existence in the case of infinite players remains an open problem, due in part to the Mean-Field*

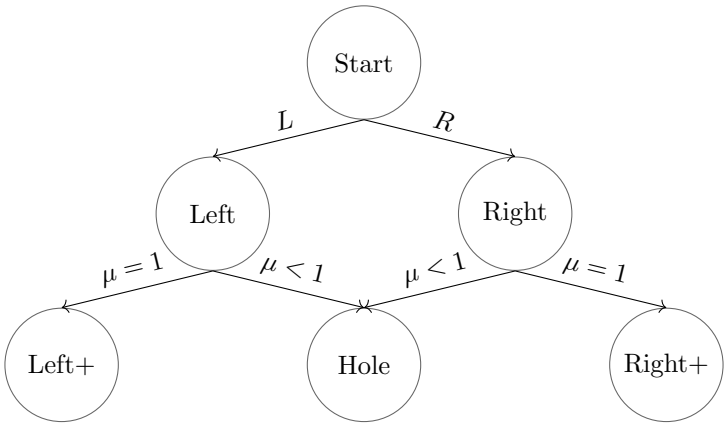


Figure 4.6: Hole-trap game

$*$: Proposition 29. $**$: Proposition 26. † : Example 11. †† : Example 12

Figure 4.7: Existence relationship between equilibrium concepts. $A \implies B$ means that the existence of A implies the existence of B; whereas $A \not\Longleftarrow B$ means that the existence of B does not imply the existence of A.

assumption that any finite set of players changing their policies would have no impact at all on the Mean-Field reward function - but Hart and Schmeidler [81]'s proof relies precisely on the fact that this isn't the case in their framework.

We begin by the following proposition, which will be the core argument for the existence proof.

With Proposition 29 proven, we know that if the game admits a Nash equilibrium, then it admits a correlated equilibrium. Therefore, for correlated equilibria to exist, it suffices that Nash equilibria exist. A sufficient condition for their existence is the continuity of r with respect to μ. This has been proven in a very similar setting by [126], for what they call a *restricted game*. We straightforwardly adapt their argument to our setting to prove Theorem 33.

Theorem 33 ((Coarse) Correlated equilibrium existence)**.** *If the reward function r and the dynamics kernel function p are continuous with respect to μ, then the game admits at least one (coarse) correlated equilibrium.*

Proof. We begin by recalling Proposition 29: if the game admits a Nash equilibrium, then it admits a correlated equilibrium.

We now prove that, under the condition that r is continuous with respect to μ, the game admits at least one Nash equilibrium.

Let $\phi : \Delta(\Pi) \to 2^{\Delta(\Pi)}$ be the best-response map,

$$\forall \nu \in \Delta(\Pi), \quad \phi(\nu) := \underset{\nu' \in \Delta(\Pi)}{\arg\max}\, J(\pi(\nu'), \mu(\nu)).$$

$\Delta(\Pi)$ is non-empty and convex; it is besides closed and bounded in a finite-dimensional space, and therefore compact.

Non-emptiness and convexity of ϕ:

For all $\nu \in \Delta(\Pi)$,

$$\arg\max_{\nu' \in \Delta(\Pi)} J(\pi(\nu'), \mu(\nu)) \subseteq \Delta(\Pi),$$

because $\Delta(\Pi)$ is compact. Therefore $\phi(\nu)$ is non-empty: the argmax exists. We now prove that $\phi(\nu)$ is convex.

Let $\nu_1,\ \nu_2 \in \phi(\nu)$, $t \in [0,1]$. Then

$$J(\pi(t\nu_1 + (1-t)\nu_2), \mu(\nu)) = tJ(\pi(\nu_1), \mu(\nu)) + (1-t)J(\pi(\nu_2), \mu(\nu)),$$

by linearity of J with respect to its first argument. This proves us that $t\nu_1 + (1-t)\nu_2 \in \phi(\nu)$, and thus that $\phi(\nu)$ is convex.

Graph(ϕ) closedness:

Graph$(\phi) = \{(\nu, \nu') \in \Delta(\Pi) \times \Delta(\Pi) \mid \nu' \in \phi(\nu)\}$. Let $(\nu_k^1, \nu_k^2)_k$ be a sequence of elements of Graph(ϕ) which converges towards $(\nu_*^1, \nu_*^2) \in \Delta(\Pi) \times \Delta(\Pi)$.

r and p are continuous in μ, therefore J is also continuous in μ. Since $J : (\nu_1, \nu_2) \to J(\pi(\nu_1), \mu(\nu_2))$ is linear in ν_1 because $J\left(\pi(\nu_1), \mu(\nu_2)\right) = \sum_i \nu_1^i J\left(\pi_i, \mu(\nu_2)\right)$, it is also continuous in both variables at the same time.

Since J is continuous in both variables at (ν_*^1, ν_*^2), let $\epsilon > 0$ and $\alpha > 0$ be such that $\forall (\nu_1, \nu_2) \in \Delta(\Pi) \times \Delta(\Pi)$ such that $d\left((\nu_1, \nu_2), (\nu_*^1, \nu_*^2)\right) \leq \alpha$,

$$\left|J\left(\pi(\nu_1), \mu(\nu_2)\right) - J(\pi(\nu_*^1), \mu(\nu_*^2))\right| \leq \epsilon$$

with d a metric over $\Delta(\Pi) \times \Delta(\Pi)$ under which J is continuous. Let $N_0 > 0$ be such that $\forall n \geq N_0,\ d\left((\nu_k^1, \nu_k^2), (\nu_*^1, \nu_*^2)\right) \leq \alpha$, and let $n \geq N_0$.

By uniform continuity (J is continuous over a compact) and triangle inequality, taking n large enough, for all $\nu \in \Delta(\Pi)$,

$$J\big(\pi(\nu), \mu(\nu_*^2)\big) \leq \epsilon + J\big(\pi(\nu), \mu(\nu_n^2)\big)$$

where the first line is obtained by uniform continuity of J.

$$-J\big(\pi(\nu_*^1), \mu(\nu_*^2)\big) \leq \epsilon - J\big(\pi(\nu_n^1), \mu(\nu_n^2)\big),$$

and by optimality of ν_n^1 against $\mu(\nu_n^2)$, $\forall \nu \in \Delta(\Pi_n)$,

$$J\big(\pi(\nu), \mu(\nu_n^2)\big) - J\big(\pi(\nu_n^1), \mu(\nu_n^2)\big) \leq 0$$

We then have, $\forall \nu \in \Delta(\Pi_n)$,

$$\begin{aligned} J(\pi(\nu), \mu(\nu_*^2)) - J(\pi(\nu_*^1), \mu(\nu_*^2)) &\leq 2\epsilon + J(\pi(\nu), \mu(\nu_n^2)) - J(\pi(\nu_n^1), \mu(\nu_n^2)) \\ &\leq 2\epsilon \end{aligned}$$

This is true for all ν, so also for their sup:

$$0 \leq \sup_{\nu} J(\pi(\nu), \mu(\nu_*^2)) - J(\pi(\nu_*^1), \mu(\nu_*^2)) \leq 2\epsilon,$$

where the first inequality comes from the sup.

Finally, this is true for all $\epsilon > 0$. Taking ϵ to 0, we have that $J(\pi(\nu_*^1), \mu(\nu_*^2)) = \sup_\nu J(\pi(\nu), \mu(\nu_*^2))$, and thus $(\nu_*^1, \nu_*^2) \in$ Graph(ϕ). Therefore Graph(ϕ) is closed.

We have all the hypotheses required to apply Kakutani's fixed point theorem [91]: there thus exists $\nu^* \in \Delta(\Pi)$ such that $\nu^* \in \phi(\nu^*)$, ie. $\nu^* = \arg\max_{\nu'} J(\pi(\nu'), \mu(\nu^*))$, which means that $\forall \nu' \in \Delta(\Pi), J(\pi(\nu'), \mu(\nu^*)) \leq J(\pi(\nu^*), \mu(\nu^*))$, in other words: ν^* is a Nash equilibrium of the game, and therefore, by Proposition 29, there exists a correlated equilibrium in the game. □

Finally, we address the question of whether (coarse) correlated equilibria are always guaranteed to exist for Mean-Field games with finite state and action spaces. Theorem 33 has already established the existence of such equilibria when the reward function r is continuous in the population distribution μ. The following example illustrates that equilibria do not necessarily exist when this continuity assumption does not hold, by highlighting a game where neither correlated nor coarse correlated equilibria exist !

Example 10 (Reward for the few)**.** *We consider a stateless Mean-Field game with two actions, a and b. The reward function is set up so as to reward the players who select the least popular action. More precisely, letting $\mu \in \mathscr{P}(\{a,b\})$ denote the population distribution over actions, we define*

$$r(a,\mu) = \begin{cases} 1 & \text{if } \mu(a) < 1/2 \\ 0 & \text{if } \mu(a) = 1/2 \\ 0 & \text{if } \mu(a) > 1/2 \end{cases}, \quad r(b,\mu) = \begin{cases} 0 & \text{if } \mu(a) < 1/2 \\ 1 & \text{if } \mu(a) = 1/2 \\ 1 & \text{if } \mu(a) > 1/2 \end{cases},$$

noting that in the case where the population is evenly split between actions a and b, the players taking action b are the one who are rewarded. Note that this payoff function is not continuous at $\mu = 1/2\delta_a + 1/2\delta_b$. Now, suppose ρ is the correlation device of a coarse correlated equilibrium. The expected return of a representative player accepting the recommendation generated by this correlation device is

$$\int \left(\nu(a) \mathbb{1}_{\{\nu(a)<1/2\}} + \nu(b) \mathbb{1}_{\{\nu(a)>1/2\}} + \frac{1}{2} \mathbb{1}_{\{\nu(a)=1/2\}} \right) \rho(\mathrm{d}\nu) = \int \min(\nu(a), \nu(b)) \rho(\mathrm{d}\nu) \,.$$

Now, the expected reward of a player that decides to deviate to action a before seeing the recommendation generated by the correlation device is $\int \mathbb{1}\{\nu(a) < 1/2\}\rho(\mathrm{d}\nu)$, and similarly the expected reward for deviating to b is $\int \mathbb{1}\{\nu(a) \geq 1/2\}\rho(\mathrm{d}\nu)$.

In order for ρ to encode a coarse correlated equilibrium, it must be the case that these expected rewards under deviation from the recommended play are no greater than the expected reward when following the recommendation:

$$\int \mathbb{1}\{\nu(a) < 1/2\}\rho(\mathrm{d}\nu) \,, \int \mathbb{1}\{\nu(a) \geq 1/2\}\rho(\mathrm{d}\nu) \leq \int \min(\nu(a), \nu(b))\rho(\mathrm{d}\nu) \,.$$

However, adding these two inequalities yields

$$1 \leq \int 2\min(\nu(a), \nu(b))\rho(\mathrm{d}\nu) \,.$$

Since $2\min(\nu(a), \nu(b)) \leq 1$, this inequality can only hold if $\nu(a) = \nu(b)$ ρ-almost surely, meaning that $\rho(1/2\delta_a + 1/2\delta_b) = 1$. However, this is clearly not a coarse correlated equilibrium, since an individual player benefits from deviating to b in this case.

We conclude no coarse correlated equilibrium (and hence no correlated equilibrium nor Nash equilibrium) exist for this Mean-Field game.

However, the following example below mitigates the previous one, by showing a game where, despite the lack of existence of Nash equilibria, correlated and coarse correlated equilibria do exist.

Example 11 (Existence of Mean-Field games with a CE and a CCE but no Nash equilibrium.)**.** *Consider a Mean-Field variant of rock-paper-scissors. If there are at least two distinct actions in the population distribution, then rock wins, and scissor loses most. If there is only a single action taken in the population, then the payoffs to each individual player are as in the standard game. More precisely, when $\mu \in \mathscr{P}(\{R,P,S\})$ is not a Dirac, we have*

$$r(R,\mu) = 1 \,, r(P,\mu) = -1, \, r(S,\mu) = -3 \,.$$

Moreover, when μ is a Dirac, say δ_R, we have the usual payoffs presented to the individual agent:

$$r(R,\delta_R) = 0 \,, r(S,\delta_R) = -1 \,, r(P,\delta_R) = 1 \,.$$

Note that this reward function is not continuous at μ when μ is a Dirac. There is no Nash equilibrium in this game: a mixed policy π cannot be a Nash equilibrium, since there is benefit in deviating to Rock, and a Dirac π cannot be a Nash equilibrium, since there is benefit to an individual agent in deviating to the superior action.

Now, we argue that the correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$ given informally by first selecting one of rock, paper, scissors uniformly at random, and then recommending this action to all players, is a coarse correlated equilibrium; mathematically, this is given by

$$\rho = 1/3\delta_{\delta_R} + 1/3\delta_{\delta_P} + 1/3\delta_{\delta_S}\,.$$

The payoff when accepting this recommendation is 0*. The average payoff when deviating to a fixed action prior to seeing the recommendation is also* 0*, hence we have a CCE. Note this is not a CE, since one can clearly deviate to a better action after seeing the recommendation.*

However, the correlating device which alternates between everyone playing paper, and half the population playing paper while the other half plays rock is a mean field correlated equilibrium. More formally,

$$\rho = 1/2\delta_{1/2\delta_P + 1/2\delta_R} + 1/2\delta_{\delta_P}$$

is a Mean Field CE. To see this, let us consider each action's deviation incentive. When players are recommended to play rock, they always have an incentive to follow this recommendation. Players are never recommended to play scissors. Therefore, we must only examine the deviation payoffs from paper to rock on the one hand, and from paper to scissors on the other hand.

$$Payoff(S \mid P) = 1\,\mathbb{P}(\nu = \delta_P \mid P) - 3\,\mathbb{P}(\nu = 1/2\delta_P + 1/2\delta_R \mid P) = 1\,\frac{2}{3} - 3\,\frac{1}{3} = -\frac{1}{3}$$

Similarly, we find that the expected deviation payoff when switching from paper to rock is $-\frac{1}{3}$. Finally, we see that the expected payoff when being recommended paper is $-\frac{1}{3}$. Players therefore never have an incentive to deviate from paper, and ρ is thus a correlated equilibrium.

We have thereby provided an instance of a game where correlated and coarse correlated equilibria exist, but Nash equilibria do not. Hence, the set of correlated equilibria of all games is strictly larger than the set of Nash equilibria.

We also need to nuance the non-existence result: as we will see in Section 5.1, although (coarse) correlated equilibria do not always exist as we have just shown, we can always find ϵ-(coarse) correlated equilibria, with $\epsilon > 0$ as small as we like. We provide here a theorem stating this property, though its proof will be the entirety of Section 5.1.

Theorem 34 (Existence of $\epsilon > 0$-(coarse) correlated equilibria)**.** *For all $\epsilon > 0$ small enough, there exists ϵ-(coarse) correlated equilibria in* all games*.*

Proof. All algorithms of Section 5.1 provably converge towards $\epsilon > 0$ (coarse) correlated equilibria, with $\epsilon \to 0$. □

To illustrate Theorem 34, we remark that in Example 10, although no exact equilibrium exists, one can easily design an ϵ-Nash equilibrium for all $1/2 > \epsilon > 0$. Indeed, taking $\nu_a = (1/2+\epsilon)\delta_a + (1/2-\epsilon)\delta_b$ and $\nu_b = (1/2-\epsilon)\delta_a + (1/2+\epsilon)\delta_b$, $\rho = 1/2\delta_{\nu_a} + 1/2\delta_{\nu_b}$ is a 4ϵ-Nash equilibrium. However, a single policy will always be $> 1/2$-exploitable, thereby showing that ϵ-Nash equilibria do not always exist for ϵ small enough.

At last, we exhibit a game where the existence of Mean Field CCE does not imply the existence of Mean Field CE.

Example 12. *Let consider the following (stateless) Mean-Field variant of rock-paper-scissors. Each member of the population selects an action from $\{R, P, S\}$, and the payoff structure is specified as:*

- *If $\mu(P) > 0$ (that is, a non-zero proportion of the population play paper), then $r(S, \mu) = 1$, $r(P, \mu) = 0$, $r(R, \mu) = -1$.*
- *If $\mu(P) = 0$ but $\mu(S) > 0$ (that is, almost no one plays paper, but a non-zero proportion play scissors), then $r(R, \mu) = 1$, $r(S, \mu) = 0$, $r(P, \mu) = -1$.*
- *Finally, if $\mu = \delta_R$, then $r(P, \mu) = 1$, $r(R, \mu) = 0$, $r(S, \mu) = -1$.*

Is there a correlated equilibrium in this game? No: if a player is ever recommended P, they realise that the sampled recommendation distribution puts mass on P, so they would benefit from deviating to S. So no MFCE can ever recommend P. But now similarly, any player recommended S could similarly benefit from deviating to R, so S cannot be recommended in a MFCE. This leaves only one possibility: that the MFCE always recommends R, but this is clearly also not an MFCE.

We now claim that $\rho = 1/3\delta_{\delta_S} + 1/3\delta_{\delta_P} + 1/3\delta_{\delta_R}$ is an MFCCE for this game. Following the recommendation leads to a payoff of 0. However, playing a fixed action also clearly leads to an expected payoff of 0, hence we have an MFCCE.

4.3.3 Uniqueness of (Coarse) Correlated Equilibria

The uniqueness of correlated and coarse correlated equilibria is less crucial than it is for Nash equilibria: indeed, when a game has a unique Nash, there can be no equilibrium selection problem, which is why Nash unicity is of interest for us. In contrast, correlated equilibria do not suffer from equilibrium selection problems due to the correlation device's role in coordinating agents. However, we identified an important situation where correlated and coarse correlated equilibria are unique: the presence of a dominant strategy, which we define as follows:

Definition 32 (Strictly-dominant strategy)**.** A strategy $\pi^* \in \bar{\Pi}$ is said to be strictly dominant if

$$J(\pi^*, \mu) > J(\pi, \mu)\,, \qquad \forall \pi \in \Pi,\, \mu \in \Delta(\mathcal{X})^{\mathcal{T}}\,.$$

Indeed, if a correlated or coarse-correlated equilibrium were to recommend any other action than the dominant one, the players would all have an incentive to play that dominant strategy instead, as we show here:

Proposition 35 ((Coarse) Correlated equilibria uniqueness)**.** *If there exists a strictly dominant strategy in the game, then the game only admits a unique coarse correlated equilibrium, and therefore a unique correlated equilibrium, which only recommends $\nu^* \in \Delta(\Pi)$ such that $\pi(\nu^*) = \pi^*$.*

Proof. Let ρ be a coarse correlated equilibrium of a game with strictly dominant strategy π^*.

Then $\forall \nu \in \Delta(\Pi)$ such that $\pi(\nu) \neq \pi^*$,

$$J(\pi^*, \mu(\nu)) > J(\pi(\nu), \mu(\nu))$$

since π^* is a strictly dominant strategy.

Therefore, unless ρ only recommends $\nu \in \Delta(\Pi)$ such that $\pi(\nu) = \pi^*$, π^* is always a strictly-value-increasing deviation. For ρ to be a coarse correlated equilibrium, it must therefore only recommend $\nu^* \in \Delta(\Pi)$ such that $\pi(\nu^*) = \pi^*$. Since there only exists one such ν^* (Otherwise two different deterministic policies would have equal state-action distribution, which is impossible), equilibrium uniqueness follows. Of course, equilibrium properties derive directly from the optimality of π^*. □

We provide an example of such a situation in the Mean-Field Prisoner's Dilemma:

Example 13 (Mean-Field Prisoner's Dilemma)**.** *Consider the two-action normal-form Mean-Field game with actions C(ooperate) and D(efect), and reward function*

$$r(C, \mu) = 3\mu(C) - \mu(D)\,,$$
$$r(D, \mu) = 4\mu(C) - 0\mu(D)$$

This game has a strictly dominating action, D.

4.3.4 Connection to the Notion of Correlated Equilibrium Derived by Campi and Fischer

A notion of correlated solution in Mean-Field games has already been introduced by Campi and Fischer [34]. The main difference between their framework and ours is that they chose to work with (policy, distribution flow) pairs (π, μ) instead of population recommendations, which led to difficulties in adapting their equilibrium concept from Mean-Field settings to N-player settings. In contrast, the concept of population distribution adapted seamlessly to N-player games and allows us to provide deeper theoretical properties such as optimality bounds in the next sections of this work.

We investigate how our definition of MFCE coincides with their notion of correlated solution. Following our notations, $\mathcal{T} := \{0, \dots, T\}$ with $T \in \mathbb{N}$ in their framework, while the state space $\mathcal{X}$ and the action space $\mathcal{A}$ are finite. The set Π is the finite set of deterministic strategies, that is

$$\Pi = \{\pi : \{0, \dots, T-1\} \times \mathcal{X} \to \mathcal{A}\} .$$

In our approach, the correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$ is introduced in order to generate different distributions of policies over the full population and synchronise hereby the players actions. In the approach detailed in [34], the synchronisation between the representative player and the population is viewed as a constraint to which their correlation device must conform. In more detail, their correlation device analogue recommends directly the representative individual policy $\pi \in \Pi$ together with the population Mean-Field flow $\mu \in \mathcal{M} = \Delta(\mathcal{X})^{\mathcal{T}}$. This gives rise to the notion of correlation flow $\bar{\rho}$, a distribution over $\Pi \times \mathcal{M}$. The main drawback of this approach is that, written as such, there is no guarantee that the policies generated by a correlating flow $\bar{\rho}$ induce a Mean-Field flow consistent with the one sampled by $\bar{\rho}$. This additional property in [34] corresponds to a consistency condition on the correlating flow $\bar{\rho}$, which can be adapted from the one described in Definition 4.1 in [34] and rewritten as follows.

Definition 33 (Consistent correlating flow [34])**.** A **consistent correlating flow** is a distribution $\bar{\rho}$ over $\Pi \times \mathcal{M}$ that satisfies the following consistency condition:

$$\mu \left(\frac{\bar{\rho}(\cdot, \mu)}{\sum\limits_{\pi \in \Pi} \bar{\rho}(\pi, \mu)} \right) = \mu , \qquad \text{for any } \mu \text{ in the support of } \bar{\rho} . \tag{4.12}$$

The consistency condition indicates that, for a potentially recommended Mean-Field flow $\mu \in \mathcal{M}$, the population recommendation induced by the correlation flow $\bar{\rho}$ conditioned by μ, that is

$$\bar{\rho}_p(\cdot \mid \mu) := \frac{\bar{\rho}(\cdot, \mu)}{\sum_{\pi \in \Pi} \bar{\rho}(\pi, \mu)}, \tag{4.13}$$

generates its own Mean-Field flow $\mu(\bar{\rho}_p(\cdot \mid \mu))$ that coincides with μ. This condition is naturally inspired by the structure of Nash equilibria definition in MFGs and is required in order to properly define the notion of correlated solution of Mean Field Games in [34]. Nevertheless, directly providing recommendations to the population when manipulating (C)CEs allows to automatically satisfy this condition. This is the approach naturally followed by our notion of correlation device.

We are now in position to establish a one to one correspondence between consistent correlation flows $\bar{\rho}$ considered in [34] and correlation devices $\rho \in \mathcal{P}(\Delta(\Pi))$ as introduced in Definition 24.

Theorem 36. *For any consistent correlating flow $\bar{\rho}$ on $\Pi \times \mathcal{M}$, there exists a correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$ that generates the same distribution over $\Pi \times \mathcal{M}$. The opposite result holds similarly.*

The derivation of this property first requires the following result.

Lemma 37. *For any $\mu \in \mathcal{M}$, the set $\mathcal{D}_\mu = \{\nu \in \Delta(\Pi) \mid \mu(\nu) = \mu\}$ is convex.*

Proof. Define $\mu^\pi(\nu)$ the state distribution flow of agents playing π when the population distribution is $\nu \in \mathcal{V}$, and $p^\pi(x' \mid x, \mu) = \sum_{a \in \mathcal{A}} \pi(x', a) p(x' \mid a, x, \mu)$ the proportion of agents going from x' to x when playing π, under population distribution μ.

The state distribution of an agent playing π in a population playing $\pi(\nu)$, is by definition

$$\mu^\pi_{t+1}(\nu)(x) = \sum_{x'} \mu^\pi_t(\nu)(x') p^\pi(x' \mid x, \mu_t(\nu)).$$

Fix $\mu \in \mathcal{M}$, ν_1, $\nu_2 \in \mathcal{D}_\mu$ and define $\nu = \alpha\nu_1 + (1-\alpha)\nu_2$ with $\alpha \in [0,1]$.

We will prove by induction on $t \in \mathcal{T}$ that, for each $\pi \in \Pi$, $\mu^\pi_t(\nu) = \mu^\pi_t(\nu_1) = \mu^\pi_t(\nu_2)$ and $\mu_t(\nu) = \alpha\mu_t(\nu_1) + (1-\alpha)\mu_t(\nu_2)$ so that $\mu_t(\nu) = \mu_t$. We first observe that this is satisfied for $t = 0$, since the initial distribution is fixed.

Suppose that the result holds at time t. Then

$$\begin{aligned}
\mu^\pi_{t+1}(\nu)(x) &= \sum_{x'} \mu^\pi_t(\nu)(x') p^\pi(x' \mid x, \mu_t(\nu)) \\
&= \sum_{x'} \mu^\pi_t(\nu_1)(x') p^\pi(x' \mid x, \mu_t) \\
&= \sum_{x'} \mu^\pi_t(\nu_1)(x') p^\pi(x' \mid x, \mu_t(\nu_1)) \\
&= \mu^\pi_{t+1}(\nu_1)(x)
\end{aligned}$$

and similarly

$$\begin{aligned}
&= \sum_{x'} \mu^\pi_t(\nu_2)(x') p^\pi(x' \mid x, \mu_t(\nu_2)) \\
&= \mu^\pi_{t+1}(\nu_2)(x) \,.
\end{aligned}$$

Besides, we observe that

$$\begin{aligned}
\mu_{t+1}(\nu)(x) &= \sum_\pi \mu^\pi_{t+1}(\nu)(x)\nu(\pi) \\
&= \alpha \sum_\pi \mu^\pi_{t+1}(\nu)(x)\nu_1(\pi) + (1-\alpha) \sum_\pi \mu^\pi_{t+1}(\nu)(x)\nu_2(\pi) \\
&= \alpha \sum_\pi \mu^\pi_{t+1}(\nu_1)(x)\nu_1(\pi) + (1-\alpha) \sum_\pi \mu^\pi_{t+1}(\nu_2)(x)\nu_2(\pi) \\
&= \alpha\mu_{t+1}(\nu_1)(x) + (1-\alpha)\mu_{t+1}(\nu_2)(x) \\
&= \mu_{t+1}(x) \,.
\end{aligned}$$

The property is initialized and hereditary, which concludes the proof: $\mathcal{D}_\mu$ is convex. □

With Lemma 37 proven, we are now in position to prove Theorem 36, *i.e.* the equivalence between our correlated equilibrium representation and the one presented in [34].

Proof of Theorem 36. Take any consistent correlation flow $\bar{\rho}$ in the sense of Campi and Fischer [34]. It can be decomposed as a distribution $\bar{\rho}_m$ over $\mathcal{M}$ combined with a conditional distribution $\bar{\rho}_p$ over Π:

$$\bar{\rho}(\pi, \mu) = \bar{\rho}_p(\pi \mid \mu)\bar{\rho}_m(\mu) \,.$$

To any $\mu \in \mathcal{M}$, we associate the induced population distribution $\nu(\mu) := \bar{\rho}_p(. \mid \mu)$. Because the correlating flow is consistent, the Mean-Field flow induced by $\nu(\mu)$ coincides with μ - i.e. $\mu(\nu(\mu)) = \mu$. Therefore the distribution over $\Pi \times \mathcal{M}$ induced by $\bar{\rho}$ is similar to the one generated by the following correlation device

$$d\rho(\nu) \quad = \quad \int_{\mu \in \mathcal{M}} \mathbf{1}_{\bar{\rho}_p(.|\mu) = \nu} \, d\bar{\rho}_m(\mu) \,.$$

Indeed, the distribution over $\Pi \times \mathcal{M}$ generated by ρ is given for $(\pi, \mu) \in \Pi \times \mathcal{M}$ by

$$\begin{aligned}\int_{\nu\in\mathcal{V}} \mathbf{1}_{\mu(\nu)=\mu}\nu(\pi)d\rho(\nu) &= \int_{\nu\in\mathcal{V}}\int_{\mu'\in\mathcal{M}} \mathbf{1}_{\bar{\rho}_p(.|\mu')=\nu}\mathbf{1}_{\mu(\nu)=\mu}\nu(\pi)d\bar{\rho}_m(\mu') \\ &= \int_{\nu\in\mathcal{V}} \mathbf{1}_{\bar{\rho}_p(.|\mu)=\nu}\nu(\pi)d\bar{\rho}_m(\mu) \\ &= d\bar{\rho}_p(\pi \mid \mu)d\bar{\rho}_m(\mu) \\ &= d\bar{\rho}(\pi,\mu)\ ,\end{aligned}$$

where we used the consistency condition in the second equality.

On the other hand, take a correlation device $\rho \in \mathcal{P}(\Delta(\Pi))$. It induces on $\Pi \times \mathcal{M}$ the following correlation flow:

$$d\bar{\rho}(\pi,\mu) = \int_{\nu\in\mathcal{V}} \mathbf{1}_{\mu(\nu)=\mu}\ \nu(\pi)d\rho(\nu)\ .$$

It remains to verify that the induced correlation flow is indeed consistent. By construction, we have that $d\bar{\rho}(\pi,\mu) \neq 0 \iff \exists\nu,\ \mu(\nu)=\mu,\ \nu(\pi)\neq 0,\ d\rho(\nu)\neq 0$.

Whenever there exists a unique $\nu \in \mathcal{D}$ such that $\mu(\nu)=\mu$, $d\rho(\nu)\neq 0$, then directly $\bar{\rho}_p(.\mid\mu)=\nu$ and the consistency condition holds.

Otherwise, $\bar{\rho}_p(.\mid\mu)$ is a mixture of several population recommendations $\nu \in \mathcal{V}$ such that $\mu(\nu)=\mu$. But Lemma 37 ensures that the set $\mathcal{D}_\mu = \{\nu \in \Delta(\Pi) \mid \mu(\nu)=\mu\}$ is convex, so that $\mu(\bar{\rho}_p(.\mid\mu))=\mu$ and the correlation flow induced by ρ is also consistent.

□

As our notion of correlating device connects now naturally to the notion of correlation flow considered in Campi and Fischer [34], we are now in position to draw connections between our notion of Correlated equilibria and the notion of correlated solution described in [34]. Before doing so, let's turn to the definition of correlated solution introduced by Campi and Fischer [34] which requires the following notion of expected return when using a deviation mapping $u \in \mathcal{U}_{CE}$ in the presence of a correlating flow $\bar{\rho}$.

Definition 34 (Correlated solution, Definition 4.1 in [34])**.** A consistent correlation flow $\bar{\rho}$ is a correlated solution to the Mean Field Game whenever the following optimality condition holds:

$$\mathbb{E}_{(\pi,\mu)\sim\bar{\rho}}\left[J(u(\pi),\mu) - J(\pi,\mu)\right] \leq 0\ ,\ \text{for any } u \in \mathcal{U}_{CE}\ .$$

Proposition 38. *A correlating flow $\bar{\rho}$ is a correlated solution in the Campi-Fischer [34] sense if and only if a corresponding correlation device ρ - which exists by Proposition 36 - is a Mean Field Correlated Equilibrium according to our definition.*

Proof. Let $\bar{\rho}$ be a consistent correlation flow generating the same distribution over $\Pi \times \mathcal{M}$ than the correlation device $\rho \in \mathcal{C}$, see Proposition 36. The consistent correlation flow $\bar{\rho}$ is a correlated solution of the MFG if and only if

$$\mathbb{E}_{(\pi,\mu)\sim\bar{\rho}}\left[J(u(\pi),\mu) - J(\pi,\mu)\right] \leq 0\ , \qquad u \in \mathcal{U}_{CE}\ .$$

On the other hand, the correlation device is a correlated equilibrium if and only if

$$\mathbb{E}_{\pi\sim\nu,\nu\sim\rho}\left[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))\right] \leq 0\ , \qquad u \in \mathcal{U}_{CE}\ .$$

The proof is complete, recalling that $\bar{\rho}$ and ρ induce the same distribution on $\Pi \times \mathcal{M}$. □

4.3.5 Homogeneous Correlated Equilibrium Characterization

Homogeneous correlation devices presented in Definition 25 are such that any agent knows what any other agent is playing, since everyone is playing the same policy. Therefore, a homogeneous ϵ-correlated equilibrium should intuitively only recommend ϵ'-Nash equilibria, with some $\epsilon' \geq 0$ to be specified. In this section, we clarify the relationship between Nash equilibria and homogeneous correlated equilibria.

We first start with linking the components of a homogeneous ϵ-Mean-Field correlated equilibrium to ϵ-Nash-equilibria.

Proposition 39. *Let $\epsilon \geq 0$ and ρ be a homogeneous ϵ-MFCE. Then all $\pi \in \Pi$ atoms of ρ are $\frac{\epsilon}{\rho(\pi)}$-MFE.*

Proof. Let $\epsilon \geq 0$, ρ be a homogeneous ϵ-MFCE and $\pi^* \in \Pi$ such that $\rho(\pi^*) > 0$.

Since ρ is an ϵ-homogeneous Mean-Field correlated equilibrium, we have

$$\int_{\pi \in \bar{\Pi}} (J(u(\pi), \mu^\pi) - J(\pi, \mu^\pi)) \rho(d\pi) \leq \epsilon \qquad \forall \pi' \in \Pi,\, u : \bar{\Pi} \to \bar{\Pi}.$$

Given $\pi^* \in \bar{\Pi}$ an atom of ρ and $\pi' \in \bar{\Pi}$ any policy, we select u such that $\forall \pi \in \bar{\Pi}$, $\pi \neq \pi^*$, $u(\pi) = \pi$ and $u(\pi^*) = \pi'$. Plugging this in the above equation, we get

$$\begin{aligned} \rho(\pi^*)(J(\pi', \mu^{\pi^*}) - J(\pi^*, \mu^{\pi^*})) &\leq \epsilon \\ J(\pi', \mu^{\pi^*}) - J(\pi^*, \mu^{\pi^*}) &\leq \frac{\epsilon}{\rho(\pi^*)}\,. \end{aligned}$$

Which means that π^* is an $\frac{\epsilon}{\rho(\pi^*)}$-MFE. □

We now know that the components of homogeneous ϵ-correlated equilibria are necessarily ϵ'-Nash equilibria. This shows that, at least in Mean-Field games, only homogeneous correlation devices recommending solely Nash equilibria can have no Φ-regret.

Finally, we answer the converse question - if a homogeneous correlated equilibrium only recommends ϵ-Mean-Field Nash equilibria, is it an ϵ-Mean-Field correlated equilibrium?

Proposition 40. *Any homogeneous correlation device recommending only ϵ-Mean-Field Nash equilibria is an ϵ-MFCE.*

Proof. Let ρ be a homogeneous correlation device with support only over ϵ-Nash equilibria.

For all $u \in \mathcal{U}^h_{CE}$, we compute

$$\begin{aligned} \mathbb{E}_{\pi \sim \rho}[J(u(\pi), \mu^\pi) - J(\pi, \mu^\pi)] &= \int_{\pi \in \bar{\Pi}} \rho(d\pi)\,(J(u(\pi), \mu^\pi) - J(\pi, \mu^\pi)) \\ &= \int_{\pi \in \epsilon\text{-Nash}} \rho(d\pi)\, \underbrace{(J(u(\pi), \mu^\pi) - J(\pi, \mu^\pi))}_{\leq \epsilon} \\ &\leq \epsilon, \end{aligned}$$

hence ρ is an ϵ-MFCE. □

4.4 Connections Between N-Player and Mean-Field Equilibria

In this section, we explore the connections between N-player and Mean-Field equilibria. We first properly define how to use Mean-Field equilibria in N-player games in section 4.4.1. We then build

in section 4.4.2 on the correspondence between our approach and the one in Campi and Fischer [34] to investigate the behavior of N-player equilibria as N tends to infinity. We show that they converge towards Mean-Field equilibria. Finally, in Section 4.4.3, we derive a key practical property by computing optimality bounds whenever using a Mean-Field equilibrium in an N-player game.

4.4.1 Mean-Field Games to N-Player Games

Before we use Mean-Field correlation devices in N-player games, we must first define how we can do so.

The population recommendation framework is very straightforward to use in N-player games : just like in Mean-Field games, we first sample a population recommendation $\nu \sim \rho$, and then, for each player, sample a policy from ν. Since there are now only N players, sampling N policies from ν yields $\nu_N \in \Delta_N(\Pi)$, a random variable with a law determined by ν and N. This means that we can view ρ as a distribution over $\Delta_N(\Pi)$, *i.e.* $\rho \in \mathcal{P}(\Delta_N(\Pi))$: ρ is an N-player correlation device !

When sampling an N-player population recommendation ν_N from a Mean-Field population recommendation $\nu \in \Delta(\Pi)$, we will use the abusive notation $\nu_N \sim \nu$. The discussions above yield the following property:

Proposition 41 (Mean-Field to N-player equilibria)**.** *Taking ρ a Mean-Field correlation device, and ρ_N its N-player version, we have that*

$$\mathbb{E}_{\nu\sim\rho,\,\nu_N\sim\nu,\,\pi\sim\nu_N}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu_N)\right]\right] = \mathbb{E}_{\nu_N\sim\rho_N,\,\pi\sim\nu_N}\left[J(u(\pi),\mu_N)\right] \quad \forall u : \Pi \to \Pi.$$

However, a Mean-Field correlation device can only be used in an N-player game if it makes sense to do so, that is, if the N-player game corresponds to the Mean-Field game. We define this notion more precisely:

Definition 35 (Corresponding N-player game)**.** Given a Mean-Field game with payoff function J and deterministic policies Π, its corresponding N-player game is the N-player game where all N players play the Mean-Field game as independent agents, and the Mean-Field population distribution is replaced by the N-players' distribution.

In other words, taking μ_N the state distribution of all N players, replace $r(x, a, \mu)$ by $r(x, a, \mu_N)$ and $p(x' \mid x, a, \mu)$ by $p(x' \mid x, a, \mu_N)$.

To rephrase the definition, players play a modified version of the Mean-Field game where their distribution flow is considered to be the game's Mean-Field flow as far as rewards and dynamics are concerned.

4.4.2 N-Player to Mean-Field Equilibria

Given the equilibrium equivalence shown in section 4.3.4 between Campi and Fischer [34]'s concepts and ours, we inherit their convergence proofs going from N-player games to the Mean-Field case: any sequence of N-player (coarse) correlated equilibria converges towards a Mean-Field (coarse) correlated equilibrium as N increases, given some conditions.

Theorem 42 (N-player CEs to Mean-Field CEs)**.** *Let $(\rho_N)_N$ be a sequence of ϵ_N-correlated equilibria in the corresponding N-player game. If the reward function and state transition functions are continuous in μ, and if $\epsilon_N \to 0$, then the limit of the sequence $(\rho_N)_N$ is a Mean-Field correlated equilibrium.*

Proof. The result follows from a direct application of Theorem 6.1 in [34] and we only need to verify that the 5 required assumptions (A1)-(A5) identified by Campi and Fischer [34] are satisfied. Assumption (A1) holds since the state transition function is continuous in μ. Assumption (A2) follows from the continuity of the reward function with respect to μ. Assumption (A3) and (A4) are valid as $(\rho_N)_N$ is a sequence of ϵ_N-correlated equilibria and $\epsilon_N \to 0$. Finally, Assumption (A5) holds by virtue of $(\rho_N)_N$ being correlated equilibria of the corresponding N-player game, so that $\mu_0^N = \mu_0$ for all N. □

Theorem 43 (N-player CCEs to Mean-Field CCEs)**.** *A similar statement holds for coarse correlated equilibria.*

Proof. The proof follows the line of argument of the one of Theorem 6.1 in [34] and simply requires to restrict the set of deviations $\mathcal{U}_{CE}$ to the more restrictive $\mathcal{U}_{CCE}$. □

We now explore the converse of these properties: which population behavior is induced by plugging Mean-Field equilibria policy in N-player games?

4.4.3 Mean-Field Equilibria in N-Player Games

Spending resources computing Mean-Field equilibria can be reasonably justified whenever we can use these equilibria in real-world situations, where, typically, agents aren't infinite, but present in very large numbers. It is therefore useful to be sure that our Mean-Field-generated equilibria work reasonably well in the large-N N-player games of interest. The purpose of this section is to provide conditions for which using a Mean-Field ϵ-(coarse) correlated equilibrium in N-player games provides an N-player $\mathcal{O}\left(\epsilon + \frac{1}{\sqrt{N}}\right)$-(coarse) correlated equilibrium !

We first consider in Theorem 44 the simple situation, where transitions do not depend on μ, then ramp up to transition functions that are Lipschitz with respect to μ, first with ρ as sums of diracs in Theorem 45, then for all correlating devices in Theorem 46.

Theorem 44. *Let ρ be an $\epsilon \geq 0$-Mean-Field (coarse) correlated equilibrium. If*

- *the reward function is γ_r-Lipschitz in μ for the L_2 norm, and*
- *the transition function does not depend on μ,*

then ρ is an $\epsilon + \frac{2\gamma_r T\left(1+\sqrt{\frac{1}{2N}}\right)}{\sqrt{N}}$-(coarse) correlated equilibrium of the corresponding N-player game.

Proof. We consider correlated equilibria, but dealing with coarse correlated ones simply requires to replace the set of deviations $\mathcal{U}_{CE}$ by $\mathcal{U}_{CCE}$. An ϵ-Mean-Field correlated equilibrium ρ in the Mean-Field's corresponding N-player game is characterized by, according to Proposition 41,

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \epsilon\,, \quad \forall u \in \mathcal{U}_{CE}\,.$$

Fix $u \in \mathcal{U}_{CE}$. The outline of the proof is the following: We first control the difference between $J(u(\pi),\mu^N_{-\pi,u(\pi)})$ and $J(u(\pi),\mu^N)$, and then bound the difference between $J(u(\pi),\mu^N)$ and $J(u(\pi),\mu(\nu))$, both using the Lipschitz property of r, and therefore of J.

We write δ_μ the indicator function of a player's position and time: if a given player i is in state x at time t, then $\delta^i_\mu(x,t) = 1$, and it is 0 for all other states at time t. Directly, we have that $\mu^N = \frac{1}{N}\sum_i \delta^i_\mu$. We overload the notation to write δ^π_μ the indicator function of the location of a given player playing π. Observe that, since $\mu^N = \sum_i \delta^i_\mu$, we can separate this sum following

$$\mu^N = \frac{1}{N}\sum_{i\neq j}\delta^i_\mu + \frac{1}{N}\delta^j_\mu,$$

i.e.

$$\mu^N = \mu^N_{-j} + \frac{1}{N}\delta^j_\mu.$$

Since this is true for all j, we can exclude the player which deviated from playing π to $u(\pi)$ from the sum:

$$\mu^N_{-\pi,u(\pi)} = \frac{N-1}{N}\mu^{N-1}_{-\pi} + \frac{1}{N}\delta^{u(\pi)}_\mu \quad \text{and} \quad \mu^N = \frac{N-1}{N}\mu^{N-1}_{-\pi} + \frac{1}{N}\delta^{\pi}_\mu\,,$$

therefore

$$\mu^N_{-\pi,u(\pi)} - \mu^N = \frac{1}{N}\left(\delta^{u(\pi)}_\mu - \delta^\pi_\mu\right)\,.$$

We will now prove that J is $T\gamma_r$-Lipschitz w.r.t. μ. Take μ_1, $\mu_2 \in \mathcal{M}$ and $\pi \in \Pi$.

$$\begin{aligned}
J(\pi,\mu_1) - J(\pi,\mu_2) &= \sum_{t\in\mathcal{T}}\sum_{x\in\mathcal{X}} \mu_t^\pi(x)\left(r^\pi(x,\mu_{1,t}) - r^\pi(x,\mu_{2,t})\right) \\
&\leq \sum_{t\in\mathcal{T}}\sum_{x\in\mathcal{X}} \mu_t^\pi(x)\gamma_r\|\mu_{1,t}-\mu_{2,t}\|_2 \\
&\leq \gamma_r \sum_{t\in\mathcal{T}} \|\mu_{1,t}-\mu_{2,t}\|_2 \\
&\leq \gamma_r \sum_{t\in\mathcal{T}} 1\sqrt{\sum_x (\mu_{1,t}(x)-\mu_{2,t}(x))^2} \\
&\leq \gamma_r \sqrt{\sum_{t\in\mathcal{T}}\sum_x (\mu_{1,t}(x)-\mu_{2,t}(x))^2}\sqrt{\sum_{t\in\mathcal{T}} 1^2} \\
&\leq \sqrt{T}\gamma_r\|\mu_1-\mu_2\|_2
\end{aligned}$$

where the first line is true because μ^π does not depend on μ_1 or μ_2, since dynamics are independent of distribution.

Since J is $\sqrt{T}\gamma_r$-Lipschitz w.r.t. μ, we deduce

$$|J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(u(\pi),\mu^N)| \leq \frac{\sqrt{T}\gamma_r}{N}\|\delta_\mu^{u(\pi)} - \delta_\mu^\pi\|_2\,.$$

Because the number of states in the game is finite, $\|\delta_\mu^{u(\pi)} - \delta_\mu^\pi\|_2$ is bounded. The maximum value of this difference is reached in the hypothetical situation where π and $u(\pi)$ never reach the same state at the same time. Hence, we have

$$\|\delta_\mu^{u(\pi)} - \delta_\mu^\pi\|_2 = \sqrt{\sum_t \underbrace{\sum_s \left(\delta_\mu^{u(\pi)}(s,t) - \delta_\mu^\pi(s,t)\right)^2}_{\leq 2}} \;\leq\; \sqrt{2T}\,,$$

so that

$$|J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(u(\pi),\mu^N)| \leq \frac{T\gamma_r\sqrt{2}}{N}. \tag{4.14}$$

Note that the above is true for any realization of the random variables δ_μ^π.

We have bounded the difference between $J(u(\pi),\mu^N_{-\pi,u(\pi)})$ and $J(u(\pi),\mu^N)$; now let us bound the difference between $J(\pi,\mu^N)$ and $J(\pi,\mu(\nu))$ for all π, as we will use the following equality later on:

$$\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(\pi,\mu^N)\right] = \mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(\pi,\mu^N) - J(\pi,\mu(\nu))\right] + J(\pi,\mu(\nu)). \tag{4.15}$$

We start with the Lipschitz property of J:

$$|J(\pi,\mu^N) - J(\pi,\mu(\nu))| \leq T\gamma_r\|\mu^N - \mu(\nu)\|_2.$$

By the Jensen inequality, we have

$$\mathbb{E}[\|\mu^N - \mu(\nu)\|_2] = \mathbb{E}\left[\sqrt{\sum_{x,t}|\mu^N(x,t) - \mu(\nu)(x,t)|^2}\right] \leq \sqrt{\sum_{x,t}\mathbb{E}\left[|\mu^N(x,t) - \mu(\nu)(x,t)|^2\right]}\,.$$

Recall that μ^N is the Mean-Field flow resulting from N players independently sampling *and* playing their policies from ν. Since the policy sampling *and* the state sampling via policy

playing are independent of other players, the *expected* distribution of all players is the Mean-Field distribution of a population playing ν, *i.e.* $\mathbb{E}[\mu^N] = \mu(\nu)$ (Though their *actual* state distribution will of course typically differ from their expected state distribution). Therefore $\forall x \in \mathcal{X}$, $t \in \mathcal{T}$, $\mathbb{E}[\mu^N(x,t)] = \mu(\nu)(x,t)$ and therefore, $\forall x \in \mathcal{X}$, $t \in \mathcal{T}$,

$$\mathbb{E}\left[\left|\mu^N(x,t) - \mu(\nu)(x,t)\right|^2\right] = \mathrm{Var}(\mu^N(x,t)).$$

The term $\mu^N(x,t) = \frac{1}{N}\sum_{i=1}^{N} \delta_\mu^i(x,t)$ is the empirical mean of N independent Bernoulli random variables with mean $\mu(\nu)(x,t)$, and therefore has variance $\frac{1}{N}\mu(\nu)(x,t)(1-\mu(\nu)(x,t))$.

We notice that whatever the value of μ, $\mu(1-\mu) \leq \mu$ since $1-\mu \leq 1$. Therefore $\forall t \leq T$, $\forall \mu \in \mathcal{M}$,

$$\sum_x \mu(x,t)(1-\mu(x,t)) \leq \sum_x \mu(x,t) = 1,$$

which yields

$$\mathbb{E}\left[\|\mu^N - \mu(\nu)\|_2\right] \leq \sqrt{\frac{T}{N}},$$

and finally gives us

$$\mathbb{E}\left[|J(u(\pi), \mu^N) - J(u(\pi), \mu(\nu))|\right] \leq \frac{T\gamma_r}{\sqrt{N}}. \tag{4.16}$$

Plugging this property into Equation 4.15, we obtain

$$\begin{aligned}\mathbb{E}_{\mu^N \sim \mu(\nu)}\left[J(\pi, \mu^N)\right] &= \mathbb{E}_{\mu^N \sim \mu(\nu)}\left[J(\pi, \mu^N) - J(\pi, \mu(\nu))\right] + J(\pi, \mu(\nu)) \\ J(\pi, \mu(\nu)) - \frac{T\gamma_r}{\sqrt{N}} &\leq \mathbb{E}_{\mu^N \sim \mu(\nu)}\left[J(\pi, \mu^N)\right] \leq J(\pi, \mu(\nu)) + \frac{T\gamma_r}{\sqrt{N}},\end{aligned}$$

where the second line comes from Equation 4.16 and the fact that $-\mathbb{E}[|X|] \leq \mathbb{E}[X] \leq \mathbb{E}[|X|]$.

We recall Equation 4.14:

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi), \mu^N_{-\pi,u(\pi)})\right]\right] \leq \frac{T\gamma_r\sqrt{2}}{N} + \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi), \mu^N)\right]\right] \quad \forall u : \Pi \to \Pi.$$

Combining all these equations, we have

$$\begin{aligned}&\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\Big[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi), \mu^N_{-\pi,u(\pi)}) - J(\pi, \mu^N)\right]\Big] \\ &\qquad \leq \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\Big[\mathbb{E}_{\nu_N\sim\nu}\left[J(u(\pi), \mu(\nu_N)) - J(\pi, \mu(\nu_N))\right]\Big] + \frac{T\gamma_r\sqrt{2}}{N} \\ &\qquad \leq \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[J(u(\pi), \mu(\nu)) - J(\pi, \mu(\nu))\right] + \frac{T\gamma_r\sqrt{2}}{N} + 2\frac{\gamma_r T}{\sqrt{N}} \\ &\qquad \leq \epsilon + \frac{2\gamma_r T\left(1 + \sqrt{\frac{1}{2N}}\right)}{\sqrt{N}}.\end{aligned}$$

where the last inequality comes from the fact that ρ is an ϵ-Mean-Field (coarse) correlated equilibrium. □

Remark 7. *We see that in this case, equilibrium approximation accuracy decreases with the time horizon, however, it does so at speed $\mathcal{O}(T)$ - surprisingly, the inaccuracy is not just not exponential in the time horizon, but it is linear ! It also linearly depends on γ_r: the lower the Lipschitz coefficient, the more accurate the approximation. There is no dependency on state space size $|\mathcal{X}|$, however we infer that it is hidden within the L_2-Lipschitz condition.*

We now tackle the more complex case of μ-dependent transitions. In N-player games, sampling recommendations from a Mean-Field correlation device ρ induces a sampling noise: when ρ has sampled population distribution ν, although the N players sample their distributions from ν, their population distribution will not be equal to ν.

Moreover, the N-players' action choices, and the game's intrinsic stochasticity will also render players' trajectories different from their expected values.

This adds a third expectation in the computation of a (coarse) correlated equilibrium's payoff, which we abusively write $\mu^N \sim \mu(\nu)$ as the distribution of N players who sampled their policies from ν.

Finally, we write μ_π^N the distribution flow associated with all N_π players playing policy $\pi \in \Pi$. Similarly, we write $\mu_\pi(\nu)$ the Mean-Field flow associated with players playing policy π when the population distribution is ν.

We first tackle the distribution-dependent dynamics in the particular case where ρ is a finite sum of diracs.

Theorem 45. *Let ρ be an $\epsilon \geq 0$-Mean-Field (coarse) correlated equilibrium. If*

- *the reward and transition functions are Lipschitz in μ for the L_2 norm, and*
- *ρ is a finite sum of diracs,*

then ρ is an $\epsilon + O\left(\frac{1}{\sqrt{N}}\right)$ (coarse) correlated equilibrium of the corresponding N-player game.

Proof. We provide here a proof outline to introduce the reader to the main arguments in the full proof, which can be found in Section 4.4.3.

Using Lipschitz arguments, we bound the (coarse) correlated equilibrium equation in the N-player game by the same equation in the Mean-Field game, with the addition of a distance term between the Mean-Field distribution and the N-player distribution

$$\mathbb{E}_{\mu^N \sim \mu(\nu)}\left[J(u(\pi), \mu^N_{-\pi,u(\pi)}) - J(\pi, \mu^N)\right] \leq J(u(\pi), \mu(\nu)) - J(\pi, \mu(\nu)) \\ + \gamma_r \mathbb{E}_{\mu^N \sim \mu(\nu)}\left[\|\mu(\nu) - \mu^N\|_2 + \|\mu(\nu) - \mu^N_{-\pi,u(\pi)}\|_2\right].$$

The rest of the proof is focused on finding bounds for the $\mathbb{E}_{\mu^N \sim \mu(\nu)}\left[\|\mu(\nu) - \mu^N\|_2\right]$ term, which can be straightforwardly extended to the $\mathbb{E}_{\mu^N \sim \mu(\nu)}\left[\|\mu(\nu) - \mu^N_{-\pi,u(\pi)}\|_2\right]$ term.

The dependence of p on μ forces us to consider every sampled policy's state distributions separately, as they influence one another: it is difficult otherwise to know policy state distributions, and thus which mixed policy is being played at which state and time.

To bound the difference between μ^N and $\mu(\nu)$, we proceed by induction over game time using a lemma which reconciles per-policy correctness (Closeness to μ_π for every π) with global correctness (Closeness to μ for every μ^N).

Finally, we conclude the proof by summing over the finite number of atoms of ρ to recover the first expectation. □

We see that we still keep a bound in $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ for this more complex case, though deriving it was much more difficult. Unfortunately, allowing ρ to be any type of distribution degrades the bounds, given our proof technique, to $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$, the square root of the former one, as we see in the following theorem.

Theorem 46. *Let ρ be an $\epsilon \geq 0$-Mean-Field (coarse) correlated equilibrium. If*

- *the reward and transition functions are Lipschitz in μ for the L_2 norm*

then ρ is an $\epsilon + O\left(\frac{1}{N^{\frac{1}{3}}}\right)$ (coarse) correlated equilibrium of the corresponding N-player game.

Proof. The line of arguments is similar to that of Theorem 45, with a few alterations to the end, that are developed in Section 4.4.3.

Indeed, the end of the proof of Theorem 45 requires summing over a finite number of values of $\frac{\rho(\nu)}{\sqrt{\nu_m N}}$, which is always finite and is indeed $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. However, when ρ is not finite, it could well be that it puts mass on a sequence for which ν_m tends to 0, and this bound therefore diverges.

To counter this, we introduce a new scalar, α,that we use to filter out policies with selection probabilities $\leq \alpha$. We prove that, while $\alpha \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, the policies that weren't filtered out will still have their state distribution $\mu_\pi \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, and so will the global state distribution. Once this is proven, we search for the best value of α leading to the best bound on N. We find that $\alpha = \frac{1}{\sqrt{N}}$ yields the bound of $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$. □

Remark 8. *We note that these proofs are much more difficult than for Nash equilibria because of Mean-Field correlated equilibria's induced stochasticities: they provide deterministic policy recommendations, and in N-player games, the number of players playing a given policy* is a random variable. *What this means is that we cannot consider that the whole population plays a policy $\pi(\nu)$, which greatly complexifies the proof.*

It is unclear whether the bound $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$ is ever reached, or if non-discrete MF(C)CEs have tighter bounds; we leave this question for future work.

However, before closing this section, we would like to make the remark that, since Nash equilibria can be cast as correlated equilibria, the above bounds also apply to Nash equilibria. Surprisingly, this is the first result of the sort of which we are aware in the fully discrete setting:

Remark 9 (Mean-Field Nash Equilibrium N-player ϵ-optimality)**.** *This development, since it applies to coarse correlated and correlated equilibria, also straightforwardly applies to Nash equilibria by Proposition 29, which, given the conditions of the above theorem, are thus $\epsilon = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-Nash equilibria in their corresponding N-player games since Proposition 30 can be adapted to N-player games).*

To the best of our knowledge, this is the first time that optimality bounds have been provided for Mean-Field Nash equilibria's optimality in N-player games for the fully discrete setting.

Useful Lemmas

We define the following lemma, which we will use in the rest of this section. Its role will be to link per-policy optimality to population optimality.

Lemma 47 (Local to Global Flow Gap)**.** *If, for $t \in \mathcal{T}$, $\forall \pi \in \Pi$ such that $\nu(\pi) > 0$, $\mathbb{E}\left[\|\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)\|_2^2\right] = \mathcal{O}\left(\frac{1}{N\nu_m}\right)$, then*

$$\mathbb{E}\left[\|\mu_N(t) - \mu(\nu)(t)\|_2^2\right] = \mathcal{O}\left(\frac{1}{N\nu_m}\right).$$

where $\nu_m = \min_{\pi,\, \nu(\pi)>0} \nu(\pi)$.

Proof of Lemma 47. We write N_π the number of players playing policy π. As the result of N independent samples from a Bernoulli random variable with law $\nu(\pi)$, this is a binomial random variable with parameters $\nu(\pi)$ and N.

We develop the squared l2 distance expression:

$$\mathbb{E}\left[\|\mu_N(t) - \mu(\nu)(t)\|_2^2\right] = \mathbb{E}\left[\|\sum_\pi \frac{N_\pi}{N}\mu_{N,\pi}(t) - \nu(\pi)\mu_\pi(\nu)(t)\|_2^2\right]$$

$$\leq \mathbb{E}\left[\left(\sum_\pi \|\frac{N_\pi}{N}\mu_{N,\pi}(t) - \nu(\pi)\mu_\pi(\nu)(t)\|_2\right)^2\right]$$

$$\leq \sum_\pi \sum_{\pi'} \mathbb{E}\left[||\frac{N_\pi}{N}\mu_{N,\pi}(t) - \nu(\pi)\mu_\pi(\nu)(t)||_2 \, ||\frac{N_{\pi'}}{N}\mu_{N,\pi'}(t) - \nu(\pi')\mu_{\pi'}(\nu)(t)||_2\right]$$

$$\leq \sum_\pi \sum_{\pi'} \mathbb{E}[\left(\nu(\pi)||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2 + |\frac{N_\pi}{N} - \nu(\pi)| \underbrace{||\mu_{N,\pi}(t)||_2}_{\leq 1}\right)$$

$$\left(\nu(\pi')||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2 + |\frac{N_{\pi'}}{N} - \nu(\pi')| \underbrace{||\mu_{N,\pi'}(t)||_2}_{\leq 1}\right)]$$

$$\leq \sum_\pi \sum_{\pi'} \nu(\pi)\nu(\pi')\mathbb{E}[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2 \, ||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2] +$$

$$\nu(\pi)\mathbb{E}[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2 \, |\frac{N_{\pi'}}{N} - \nu(\pi')|] +$$

$$\nu(\pi')\mathbb{E}[||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2 \, |\frac{N_\pi}{N} - \nu(\pi)|] +$$

$$\mathbb{E}[|\frac{N_\pi}{N} - \nu(\pi)| \, |\frac{N_{\pi'}}{N} - \nu(\pi')|)].$$

We use the Cauchy-Schwarz inequality to separate-out terms in the expectations:

$$\mathbb{E}[||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2 \, |\frac{N_{\pi'}}{N} - \nu(\pi')|] \leq \sqrt{\mathbb{E}[||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2^2] \, \mathbb{E}[|\frac{N_{\pi'}}{N} - \nu(\pi')|^2]}$$

and similarly so for the other expressions; then bound each term.

By assumption, $\mathbb{E}[||\mu_{N,\pi'}(t) - \mu_{\pi'}(\nu)(t)||_2^2] = \mathcal{O}\left(\frac{1}{N\nu_m}\right)$.

N_π is the number of players who have sampled policy π. This is a binomial random variable with parameters $(\nu(\pi), N)$, and therefore $\mathbb{E}\left[\left(\nu(\pi) - \frac{N_\pi}{N}\right)^2\right] = \frac{1}{N}\nu(\pi)(1 - \nu(\pi))$.

Finally, on the interval $[\nu_m, 1]$, where $\nu_m = \min_{\pi,\, \nu(\pi)>0} \nu(\pi)$, $\sqrt{\nu(\pi)(1-\nu(\pi))} \leq \nu(\pi)\sqrt{\frac{1-\nu_m}{\nu_m}}$.

Plugging these back in the former expressions, we obtain

$$\mathbb{E}\left[\|\mu_N(t) - \mu(\nu)(t)\|_2^2\right] \leq \sum_\pi \sum_{\pi'} \nu(\pi)\nu(\pi')\mathcal{O}\left(\frac{1}{N\nu_m}\right) + 2\nu(\pi)\nu(\pi')\mathcal{O}\left(\frac{1}{N\nu_m}\right)$$

$$+\nu(\pi)\nu(\pi')\mathcal{O}\left(\frac{1}{N\nu_m}\right)$$

$$= \mathcal{O}\left(\frac{1}{N\nu_m}\right)$$

which concludes the proof. □

We will also implicitly use a lemma linking Lipschitzness in p and r to Lipschitzness in J.

Lemma 48 (Lipschitzness of J)**.** *Assume r and p are γ_r- and γ_p-Lipschitz in μ, respectively. Then J is $\left(\frac{|\mathcal{X}|}{|\mathcal{X}|-1}\gamma_p R_M\sqrt{T - 2\frac{1-|\mathcal{X}|^T}{1-|\mathcal{X}|} + \frac{1-|\mathcal{X}|^{2T}}{1-|\mathcal{X}|^2}} + \sqrt{T}\gamma_r\right)$-Lipschitz in μ where R_M is the highest absolute reward obtainable in the game.*

Proof. Take μ_1, $\mu_2 \in \mathcal{M}$. We start by proving that, given a policy π, its expected distribution $\mu^\pi_{\mu_1}$ under μ_1 and $\mu^\pi_{\mu_2}$ under μ_2 are such that $\forall t,\ x,\ \mu^\pi_{\mu_1,t}(x) - \mu^\pi_{\mu_2,t}(x) \leq \gamma_p \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_1 - \mu_2\|_2$.

We prove this by induction over game time.

At $t = 0$, $\mu^\pi_{\mu_1,0} = \mu^\pi_{\mu_2,0} = \mu_0$, hence the relationship is verified.

Assuming that $\mu^\pi_{\mu_1,t}(x) - \mu^\pi_{\mu_2,t}(x) \leq \gamma_p \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2$, then take $x \in \mathcal{X}$.

$$\mu^\pi_{\mu_1,t+1}(x) = \sum_{x_t} p^\pi(x \mid x_t, \mu_{1,t}) \mu^\pi_{\mu_1,t}(x_t),$$

where $p^\pi(\cdot \mid x, \mu) = \sum_a \pi(x,a) p(\cdot \mid x, a, \mu)$.

$$\begin{aligned}
\mu^\pi_{\mu_1,t+1}(x) &\leq \sum_{x_t} (p^\pi(x \mid x_t, \mu_{2,t}) + \gamma_p \|\mu_{1,t} - \mu_{2,t}\|_2) \mu^\pi_{\mu_1,t}(x_t) \\
\mu^\pi_{\mu_1,t+1}(x) &\leq \sum_{x_t} p^\pi(x \mid x_t, \mu_{2,t}) \mu^\pi_{\mu_1,t}(x_t) + \sum_{x_t} \gamma_p \|\mu_{1,t} - \mu_{2,t}\|_2 \mu^\pi_{\mu_1,t}(x_t) \\
\mu^\pi_{\mu_1,t+1}(x) &\leq \sum_{x_t} p^\pi(x \mid x_t, \mu_{2,t}) \mu^\pi_{\mu_1,t}(x_t) + \gamma_p \|\mu_{1,t} - \mu_{2,t}\|_2 \underbrace{\sum_{x_t} \mu^\pi_{\mu_1,t}(x_t)}_{=1} \\
\mu^\pi_{\mu_1,t+1}(x) &\leq \sum_{x_t} \underbrace{p^\pi(x \mid x_t, \mu_{2,t})}_{\leq 1} (\mu^\pi_{\mu_2,t}(x_t) + \gamma_p \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2) + \gamma_p \|\mu_{1,t} - \mu_{2,t}\|_2 \\
\mu^\pi_{\mu_1,t+1}(x) &\leq \mu^\pi_{\mu_2,t+1}(x) + |\mathcal{X}| \gamma_p \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2 + \gamma_p \|\mu_{1,t} - \mu_{2,t}\|_2 \\
\mu^\pi_{\mu_1,t+1}(x) - \mu^\pi_{\mu_2,t+1}(x) &\leq \gamma_p \frac{1-|\mathcal{X}|^{t+1}}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2.
\end{aligned}$$

The property is hereditary and initialized, hence it is true.

We now turn to the proof of J being Lipschitz. Take $\pi \in \Pi$, μ_1, $\mu_2 \in \mathcal{M}$. Then we have

$$
\begin{aligned}
|J(\pi,\mu_1) - J(\pi,\mu_2)| &= |\sum_x \sum_t \mu^\pi_{\mu_1,t}(x) r^\pi(x,\mu_{1,t}) - \mu^\pi_{\mu_{2,t},t}(x) r^\pi(x,\mu_{2,t})| \\
&= |\sum_x \sum_t \left(\mu^\pi_{\mu_1,t}(x) - \mu^\pi_{\mu_2,t}(x)\right) r^\pi(x,\mu_{2,t}) + \mu^\pi_{\mu_1,t}(x)\left(r^\pi(x,\mu_1) - r^\pi(x,\mu_{2,t})\right)| \\
&\leq \sum_x \sum_t |\mu^\pi_{\mu_1,t}(x) - \mu^\pi_{\mu_2,t}(x)||r^\pi(x,\mu_{2,t})| + \sum_x \sum_t \mu^\pi_{\mu_1,t}(x)|r^\pi(x,\mu_{1,t}) - r^\pi(x,\mu_{2,t})| \\
&\leq \sum_x \sum_t \gamma_p \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2 \underbrace{|r^\pi(x,\mu_{2,t})|}_{\leq R_M} + \sum_t \underbrace{\sum_x \mu^\pi_{\mu_1,t}(x)}_{=1} |r^\pi(x,\mu_{1,t}) - r^\pi(x,\mu_{2,t})| \\
&\leq |\mathcal{X}|\gamma_p R_M \sum_{t\geq 1} \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t-1} - \mu_{2,t-1}\|_2 + \gamma_r \sum_t \|\mu_{1,t} - \mu_{2,t}\|_2| \\
&\leq |\mathcal{X}|\gamma_p R_M \sum_t \frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|} \|\mu_{1,t} - \mu_{2,t}\|_2 + \gamma_r \sqrt{\sum_t \|\mu_{1,t} - \mu_{2,t}\|_2^2} \sqrt{\sum_t 1} \\
&\leq |\mathcal{X}|\gamma_p R_M \sqrt{\sum_t \left(\frac{1-|\mathcal{X}|^t}{1-|\mathcal{X}|}\right)^2} \sqrt{\sum_t \|\mu_{1,t} - \mu_{2,t}\|_2^2} + \sqrt{T}\gamma_r \|\mu_1 - \mu_2\|_2 \\
&\leq |\mathcal{X}|\gamma_p R_M \sqrt{\sum_t \frac{1 - 2|\mathcal{X}|^t + |\mathcal{X}|^{2t}}{\left(1-|\mathcal{X}|\right)^2}} \|\mu_1 - \mu_2\|_2 + \sqrt{T}\gamma_r \|\mu_1 - \mu_2\|_2 \\
&\leq |\mathcal{X}|\gamma_p R_M \sqrt{\frac{T - 2\frac{1-|\mathcal{X}|^T}{1-|\mathcal{X}|} + \frac{1-|\mathcal{X}|^{2T}}{1-|\mathcal{X}|^2}}{\left(1-|\mathcal{X}|\right)^2}} \|\mu_1 - \mu_2\|_2 + \sqrt{T}\gamma_r \|\mu_1 - \mu_2\|_2 \\
&\leq \frac{|\mathcal{X}|}{|\mathcal{X}|-1}\gamma_p R_M \sqrt{T - 2\frac{1-|\mathcal{X}|^T}{1-|\mathcal{X}|} + \frac{1-|\mathcal{X}|^{2T}}{1-|\mathcal{X}|^2}} \|\mu_1 - \mu_2\|_2 + \sqrt{T}\gamma_r \|\mu_1 - \mu_2\|_2
\end{aligned}
$$

which concludes the proof. Of course, the case $|\mathcal{X}| = 1$ is trivially solved: if there is only one state, then all distributions are equal. □

Note that if $\gamma_p = 0$, *i.e.* the transition function does not depend on μ, the above Lipschitz constant becomes the same as in the transition-independent case in Theorem 44's proof.

Proof of Theorem 45

We recall Theorem 45:

Theorem. *Let ρ be an $\epsilon \geq 0$-Mean-Field (coarse) correlated equilibrium. Then, if*

- *The reward and transition functions are Lipschitz in μ for the L_2 norm, and*
- *ρ is a finite sum of diracs,*

then ρ is an $\epsilon + O\left(\frac{1}{\sqrt{N}}\right)$ (coarse) correlated equilibrium of the corresponding N-player game.

Proof. Let $u \in \mathcal{U}_{\{CE,CCE\}}$. An ϵ-(coarse) correlated equilibrium ρ in the Mean-Field's corresponding N-player game satisfies:

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \epsilon.$$

The outline of the proof is the following: We proceed first by bounding the difference between μ^N and $\mu(\nu)$, which we do by induction over timesteps and by separating players by the policy

they sampled. Once this is done, we bound the difference between μ^N and $\mu^N_{u(\pi),-\pi}$, and finally use a Lipschitz argument to relate $\mathbb{E}\left[|J(u(\pi),\mu(\nu)) - J(u(\pi),\mu^N)|\right]$ to $\mathbb{E}\left[\|\mu(\nu)-\mu^N\|_2\right]$, which we have just bounded: indeed, Lemma 48 shows that if p and r are μ-Lipschitz, then so is J.

Indeed, assuming $\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[\|\mu^N-\mu(\nu))\|_2\right] = \mathcal{O}\left(\frac{1}{\sqrt{\alpha N}}\right)$ with any $\alpha > 0$, we have

$$\begin{aligned}
\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(\pi,\mu^N)\right] &= J(\pi,\mu(\nu)) + \mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(\pi,\mu^N) - J(\pi,\mu(\nu))\right], \\
&\leq J(\pi,\mu(\nu)) + \mathbb{E}_{\mu^N\sim\mu(\nu)}\left[|J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)|\right], \\
&\leq J(\pi,\mu(\nu)) + \gamma_r\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[\|\mu(\nu)-\mu^N\|_2\right], \\
&\leq J(\pi,\mu(\nu)) + \mathcal{O}\left(\frac{1}{\sqrt{\alpha N}}\right).
\end{aligned}$$

Once we have reached this point, we do the same operation and get the expected result for $J(u(\pi),\mu^N_{-\pi,u(\pi)})$. Unfortunately, the term α in the $\mathcal{O}$ depends on ν, hence we will need to be careful when taking the expectation with respect to ν. This yields to two different cases: In the case when ρ is discrete, which is the case which typically interests us, we keep the $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ bound; but in the case when ρ is continuous, we are left with a less strong bound of $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$.

Let us first prove that, for all ν, there exists some $\alpha > 0$ which we will show is equal to $\min_{\pi|\nu(\pi)>0}\nu(\pi)$, such that $\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[\|\mu^N-\mu(\nu))\|_2\right] = \mathcal{O}\left(\frac{1}{\sqrt{\alpha N}}\right)$. We start by noting that $\mu_N(t) = \sum_\pi \frac{N_\pi}{N}\mu_{N,\pi}(t)$ and $\mu(\nu)(t) = \sum_\pi \nu(\pi)\mu_\pi(\nu)(t)$.

The above development shows that the proof requires us to bound $\mathbb{E}\left[\|\mu_N-\mu(\nu)\|_2\right]$. We proceed to do precisely this by induction on time, bounding each term $\mathbb{E}\left[\|\mu_N(t)-\mu(\nu)(t)\|_2^2\right]$ for any $\nu\in\Delta(\Pi)$.

Indeed,

$$\begin{aligned}
\mathbb{E}\left[\|\mu_N-\mu(\nu)\|_2\right] &= \mathbb{E}\left[\sqrt{\sum_t\sum_x\left(\mu_N(t)(x)-\mu(\nu)(t)(x)\right)^2}\right] \\
&\leq \sqrt{\sum_t\mathbb{E}\left[\sum_x\left(\mu_N(t)(x)-\mu(\nu)(t)(x)\right)^2\right]} \\
&\leq \sqrt{\sum_t\mathbb{E}\left[\|\mu_N(t)-\mu(\nu)(t)\|_2^2\right]}
\end{aligned}$$

Our induction hypothesis is, $\forall\nu\in\Delta(\Pi)$, $\forall t\in\mathcal{T}$, $\forall\pi\in\Pi$ such that $\nu(\pi)>0$,

$$\mathbb{E}\left[\|\mu_{N,\pi}(t)-\mu_\pi(\nu)(t)\|_2^2\right] = \mathcal{O}\left(\frac{1}{N\nu_m}\right)$$

and

$$\mathbb{E}\left[\|\mu_N(t)-\mu(\nu)(t)\|_2^2\right] = \mathcal{O}\left(\frac{1}{N\nu_m}\right).$$

Induction initialization: We initialize the induction with $t=0$, and consider any π such that $\nu(\pi)>0$.

$$\mathbb{E}\left[\|\mu_{N,\pi}(0)-\mu_\pi(\nu)(0)\|_2^2\right] = \sum_{n=0}^N\mathbb{P}(N_\pi=n)\mathbb{E}\left[\|\mu_{N,\pi}(0)-\mu_\pi(\nu)(0)\|_2^2\,\middle|\,N_\pi=n\right].$$

When $N_\pi=0$, then $\mu_{N,\pi}=0$ everywhere, as there are no agents playing π. In this case, we have that

$$\mathbb{E}\left[\|\mu_{N,\pi}(0)-\mu_\pi(\nu)(0)\|_2^2 \mid N_\pi=0\right] = \mathbb{E}\left[\|\mu_\pi(\nu)(0)\|_2^2\right] \leq 1.$$

We have that $\forall x \in \mathcal{X}$, $\mu_{N,\pi}(0)(x)$ is the empirical mean of N_π i.i.d. variables $\delta_{X_0=x} \sim \mathcal{B}(\mu_0(x))$, since at time 0, all N players are independently distributed according to $\mu(\nu)(0) = \mu_0$.

Therefore $\mathbb{E}\left[(\mu_{N,\pi}(0)(x) - \mu_\pi(\nu)(0)(x))^2 \,\middle|\, N_\pi\right] = \frac{1}{N_\pi}\mu(\nu)(0)(x)(1-\mu(\nu)(0)(x))$ and thus, since $\mu(\nu)(0)(x)(1-\mu(\nu)(0)(x)) \leq \frac{1}{2}$,

$$\mathbb{E}\left[\|\mu_N(0) - \mu(\nu)(0)\|_2^2 \,\middle|\, N_\pi\right] \leq \frac{1}{2N_\pi}.$$

Taking the expectation over N_π yields, since N_π is a binomial random variable with parameters $(\nu(\pi), N)$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mu_{N,\pi}(0) - \mu_\pi(\nu)(0)\|_2^2\right] \\
&\leq \binom{N}{0}(1-\nu(\pi))^N \mathbb{E}\left[\|\mu_{N,\pi}(0) - \mu_\pi(\nu)(0)\|_2^2 \mid N_\pi = 0\right] + \sum_{n=1}^{N}\binom{N}{n}\nu(\pi)^n(1-\nu(\pi))^{N-n}\frac{1}{2n} \\
&\leq (1-\nu(\pi))^N + \frac{1}{\nu(\pi)(N+1)}\underbrace{\sum_{n=1}^{N}\binom{N+1}{n+1}\nu(\pi)^{n+1}(1-\nu(\pi))^{(N+1)-(n+1)}}_{\leq 1}\underbrace{\frac{n+1}{2n}}_{\leq 1} \\
&\leq (1-\nu(\pi))^N + \frac{1}{\nu(\pi)(N+1)} \\
&= \mathcal{O}\left(\frac{1}{\nu_m N}\right). \qquad (4.17)
\end{aligned}$$

Applying Lemma 47 concludes the initialization step.

Induction step: Let $t \geq 0$, $x \in \mathcal{X}$, and assume that $\mathbb{E}\left[\|\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)\|_2^2\right] = \mathcal{O}\left(\frac{1}{\nu_m N}\right)$ for all $\pi \in \Pi$ such that $\nu(\pi) > 0$. We also write $p_x = \sum_{x_t} p^\pi(x \mid x_t, \mu_N(t))\mu_{N,\pi}(t)(x_t)$ the expected state density at state x.

$$\begin{aligned}
&\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x) - \mu_\pi(\nu)(t+1)(x))^2\right] \\
&= \mathbb{E}\left[((\mu_{N,\pi}(t+1)(x) - p_x) + (p_x - \mu_\pi(\nu)(t+1)(x)))^2\right] \\
&= \mathbb{E}\left[(\mu_{N,\pi}(t+1)(x) - p_x)^2\right] + 2\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x) - p_x)(p_x - \mu_\pi(\nu)(t+1)(x))\right] \\
&\qquad + \mathbb{E}\left[(p_x - \mu_\pi(\nu)(t+1)(x))^2\right] \qquad (4.18)
\end{aligned}$$

We will bound each term in Equation 4.18 separately. We start with its first term.
The evolution equation for the subpopulation playing π is

$$\mathbb{E}\left[\mu_{N,\pi}(t+1)(x)\right] = \mathbb{E}\left[\sum_{x_t} p^\pi(x \mid x_t, \mu_N(t))\mu_{N,\pi}(t)(x_t)\right]$$

We note that for all $x \in \mathcal{X}$, we can write $\mu_{N,\pi}(t+1)(x) = \sum_{x_t}\mu_{N,\pi}(t+1)(x|x_t)\mu_{N,\pi}(t)(x_t)$, where $\mu_{N,\pi}(t+1)(x|x_t)$ is the proportion of particles at state x_t at time t which went to state x at time $t+1$.

We observe that $(N_\pi\mu_{N,\pi}(t)(x_t))\mu_{N,\pi}(t+1)(x|x_t)$ is the number of players playing π present at state x_t at time t who moved to state x at time $t+1$, which is a binomial random variable of parameters $p^\pi(x \mid x_t, \mu_N(t))$, the probability of moving to x from x_t when playing π, and $N_\pi\mu_{N,\pi}(t)(x_t)$, the number of players playing π at x_t at time t.

Hence, if we write $\Delta(x, x_t) = \mu_{N,\pi}(t+1)(x|x_t) - p^\pi(x \mid x_t, \mu_N(t))$ and recall that $p_x = \sum_{x_t} p^\pi(x \mid x_t, \mu_N(t))\mu_{N,\pi}(t)(x_t)$

$$
\begin{aligned}
&\mathbb{E}\left[(\mu_{N,\pi}(t+1)(s)-p_x)^2 \mid N_\pi, p_x\right] \\
=&\mathbb{E}\left[(\sum_{x_t}(\mu_{N,\pi}(t+1)(x|x_t)-p^\pi(x \mid x_t,\mu_N(t)))\mu_{N,\pi}(t)(x_t))^2 \mid N_\pi, p_x\right] \\
=&\mathbb{E}\left[\sum_{x_t}\sum_{x_t'}\Delta(x,x_t)\Delta(x,x_t')\mu_{N,\pi}(t)(x_t)\mu_{N,\pi}(t)(x_t') \mid N_\pi, p_x\right] \\
\leq&\sum_{x_t}\sum_{x_t'}\sqrt{\mathbb{E}\left[\Delta(x,x_t)^2\mu^2_{N,\pi}(t)(x_t) \mid N_\pi, p_x\right]}\sqrt{\mathbb{E}\left[\Delta(x,x_t')^2\mu^2_{N,\pi}(t)(x_t') \mid N_\pi, p_x\right]}
\end{aligned}
$$

By virtue of $\mu^2_{N,\pi}(t)(x_t)$ being $\mu_{N,\pi}(t)(x_t)$-measurable, we have

$$
\mathbb{E}\left[\Delta(x,x_t)^2\mu^2_{N,\pi}(t)(x_t) \mid N_\pi, p_x\right] = \mathbb{E}\left[\mathbb{E}\left[\Delta(x,x_t)^2 \mid \mu_{N,\pi}(t)(x_t)\right]\mu^2_{N,\pi}(t)(x_t) \mid N_\pi, p_x\right]
$$

Given that $(N_\pi\mu_{N,\pi}(t)(x_t))\mu_{N,\pi}(t+1)(x|x_t)$ is a binomial random variable with parameters $p^\pi(x \mid x_t,\mu_N(t))$ and $N_\pi\mu_{N,\pi}(t)(x_t)$,

$$
\begin{aligned}
\mathbb{E}\left[\Delta(x,x_t)^2 \mid \mu_{N,\pi}(t)(x_t), N_\pi, p_x\right] &= \frac{1}{N_\pi^2\mu^2_{N,\pi}(t)(x_t)}\mathbb{E}\left[N_\pi^2\mu^2_{N,\pi}(t)(x_t)\Delta(x,x_t)^2 \mid \mu_{N,\pi}(t)(x_t), N_\pi, p_x\right] \\
&= \frac{1}{N_\pi^2\mu^2_{N,\pi}(t)(x_t)}p^\pi(x \mid x_t,\mu_N(t))(1-p^\pi(x \mid x_t,\mu_N(t)))N_\pi\mu_{N,\pi}(t)(x_t) \\
&\leq \frac{1}{N_\pi\mu_{N,\pi}(t)(x_t)}
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathbb{E}\left[\Delta(x,x_t)^2\mu^2_{N,\pi}(t)(x_t) \mid N_\pi, p_x\right] &\leq \mathbb{E}\left[\frac{1}{N_\pi}\mu_{N,\pi}(t)(x_t) \mid N_\pi, p_x\right] \\
&\leq \frac{1}{N_\pi}.
\end{aligned}
$$

Plugging this back into the former equation, this yields

$$
\begin{aligned}
\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x)-p_x)^2 \mid N_\pi, p_x\right] &\leq \sum_{x_t}\sum_{x_t'}\frac{1}{N_\pi} \\
&\leq \frac{1}{N_\pi}|\mathcal{X}|^2.
\end{aligned}
$$

Taking the expectation over N_π and following the same steps as Equation 4.17, we have $\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x)-p_x)^2\right] \leq \left((1-\nu(\pi))^N + \frac{2}{\nu(\pi)(N+1)}\right)|\mathcal{X}|^2 = \mathcal{O}\left(\frac{1}{\nu_m N}\right)$.

Bounding the second and third terms in Equation 4.18:

The middle-term is simplified using the Cauchy-Schwarz inequality:

$$
\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x)-p_x)(p_x-\mu_\pi(\nu)(t+1)(x))\right] \leq \sqrt{\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x)-p_x)^2\right]\mathbb{E}\left[(p_x-\mu_\pi(\nu)(t+1)(x))^2\right]}
$$

We have an upper bound for the first term, let us now bound the second term.

$$
\begin{aligned}
&\mathbb{E}[(p_x - \mu_\pi(\nu)(t+1)(x))^2] \\
&= \mathbb{E}\left[\left(\sum_{x_t}\sum_{a} \pi(x_t,a)\left(p(x \mid x_t, a, \mu_N(t))\mu_{N,\pi}(t)(x_t) - p(x \mid x_t, a, \mu(\nu)(t))\mu_\pi(\nu)(t)(x_t)\right)\right)^2\right] \\
&\leq \mathbb{E}[(\underbrace{\sum_{x_t} p^\pi(x \mid x_t, \mu(\nu)(t))\left(\mu_{N,\pi}(t)(x_t) - \mu_\pi(\nu)(t)(x_t)\right)}_{\leq \sqrt{\sum_{x_t} p^\pi(x|x_t,\mu(\nu)(t))^2}\sqrt{\sum_{x_t}(\mu_{N,\pi}(t)(x_t)-\mu_\pi(\nu)(t)(x_t))^2}} + \\
&\qquad \underbrace{\sum_{x_t}\mu_{N,\pi}(t)(x_t)}_{=1}\underbrace{\sum_a \pi(x_t,a)}_{=1}\gamma_p\|\mu_N(t) - \mu(\nu)(t)\|_2)^2] \\
&\leq \mathbb{E}[(\underbrace{\sqrt{\sum_{x_t} p^\pi(x \mid x_t, \mu(\nu)(t))^2}}_{\leq\sqrt{|\mathcal{X}|}}\|\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)\|_2 + \gamma_p\|\mu_N(t) - \mu(\nu)(t)\|_2)^2] \\
&\leq \mathbb{E}\left[\left(\sqrt{|\mathcal{X}|}\|\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)\|_2 + \gamma_p\|\mu_N(t) - \mu(\nu)(t)\|_2\right)^2\right] \\
&\leq |\mathcal{X}|\underbrace{\mathbb{E}\left[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2^2\right]}_{=\mathcal{O}(\frac{1}{\nu_m N})} + 2\gamma_p\sqrt{|\mathcal{X}|}\mathbb{E}\left[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2\,||\mu_N(t) - \mu(\nu)(t)||_2\right] + \gamma_p\underbrace{\mathbb{E}\left[||\mu_N(t) - \mu(\nu)(t)||_2^2\right]}_{=\mathcal{O}(\frac{1}{\nu_m N})},
\end{aligned}
$$

where the third line inequality comes from p being γ_p-Lipschitz in μ, and the last equalities in underbraces come from the induction assumption and Lemma 47. We apply the Cauchy-Schwarz inequality to the middle term:

$$
\begin{aligned}
&\mathbb{E}[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2\,||\mu_N(t) - \mu(\nu)(t)||_2] \\
&\quad\leq \sqrt{\mathbb{E}\left[||\mu_{N,\pi}(t) - \mu_\pi(\nu)(t)||_2^2\right]\mathbb{E}\left[||\mu_N(t) - \mu(\nu)(t)||_2^2\right]} \\
&\quad= \mathcal{O}\left(\frac{1}{\nu_m N}\right)
\end{aligned}
$$

Therefore, for all $t \geq 0$, $\pi \in \Pi$, $\nu(\pi) > 0$, $\mathbb{E}\left[(\mu_{N,\pi}(t+1)(x) - \mu_\pi(\nu)(t+1)(x))^2\right] = \mathcal{O}\left(\frac{1}{\nu_m N}\right)$, and thus $\mathbb{E}\left[\|\mu_N - \mu(\nu)\|_2\right] = \mathcal{O}\left(\frac{1}{\sqrt{\nu_m N}}\right)$.

Neglecting the deviation term: We now consider the case when one player deviates from policy π to policy $u(\pi)$. The effect of this defection is an impurity of the policy distribution with, as a result, an increase of $N_{u(\pi)}$ and a decrease of N_π by 1 each. We briefly describe how this change can be neglected.

If the deviated-to policy is in the support of ν: We see that the result of Lemma 47 remains unchanged, as the additional $\frac{1}{N_{u(\pi)}}$ / $\frac{-1}{N_\pi}$ can be separated using the triangle inequality, and is $\mathcal{O}\left(\frac{1}{\nu_m N}\right)$.

Both the initialization and the inheritance parts of the recurrence involve the quantity N_π, but the only influence of this impurity is in the expectation's conditioning (or in the summation indices). We see that this change replaces the $\frac{1}{N_\pi}$ term by $\frac{1}{N_\pi \pm 1}$, and therefore ultimately only changes the bounds by a constant amount. If this leads to a policy which is not played anymore (that is, $N_\pi = 1$ before the deviation), then we can use the previously-developed argument regarding the $N_\pi = 0$ case, noting that the probability of $N_\pi = 1$ is $N\nu(\pi)(1-\nu(\pi))^{N-1} = O\left(\frac{1}{\nu_m N}\right)$.

Thus we also have that

$$\mathbb{E}\left[\|\mu^N_{-\pi,u(\pi)} - \mu(\nu)\|_2\right] = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

If the deviated-to policy is *not* in the support of ν: Then it creates a single new term in the local-to-global development (We note N'_π the "updated" number of players playing π: either it is equal to N_π for the non-deviating policies, or it is equal to $N_\pi - 1$ for the policy the deviating player played; and $u(\pi)$ the deviated-to policy):

$$\mathbb{E}\left[\|\mu_N(t) - \mu(\nu)(t)\|_2^2\right] = \mathbb{E}\left[\|\frac{1}{N}\mu_{N,u(\pi)}(t) + \sum_\pi \frac{N'_\pi}{N}\mu_{N,\pi}(t) - \nu(\pi)\mu_\pi(\nu)(t)\|_2^2\right]$$

$$\leq \mathbb{E}\left[\left(\frac{1}{N}\underbrace{\|\mu_{N,u(\pi)}(t)\|_2}_{\leq 1} + \sum_\pi \|\frac{N'_\pi}{N}\mu_{N,\pi}(t) - \nu(\pi)\mu_\pi(\nu)(t)\|_2\right)^2\right]$$

We see that this new impurity adds a $\frac{1}{N}$ term within the sum, which does not alter the end result regarding the closeness of μ^N to $\mu(\nu)$.

Integrating over $\Delta(\Pi)$: The bound derived above depends on ν, yet to compute expected deviation payoffs, we must integrate over $\Delta(\Pi)$ following ρ's distribution. In the current case, ρ is a sum of finitely many diracs.

Then

$$\mathbb{E}_{\nu\sim\rho}\left[\frac{1}{\sqrt{\nu_m N}}\right] = \sum_{\nu|\rho(\nu)>0} \frac{\rho(\nu)}{\sqrt{\nu_m N}}$$

i.e.

$$\mathbb{E}_{\nu\sim\rho}\left[\frac{1}{\sqrt{\nu_m N}}\right] = \frac{1}{\sqrt{N}}\sum_{\nu|\rho(\nu)>0} \frac{\rho(\nu)}{\sqrt{\nu_m}}$$

and we keep the $\frac{1}{\sqrt{N}}$ bound, with an added term representing the non-optimality of each ν in the discrete support of ρ weighted by ρ. □

Proof of Theorem 46

We recall Theorem 46:

Theorem. *Let ρ be an $\epsilon \geq 0$-Mean-Field (coarse) correlated equilibrium. Then, if*

- *The reward and transition functions are Lipschitz in μ for the L_2 norm, and*
- *ρ is not a finite sum of diracs,*

then ρ is an $\epsilon + O\left(\frac{1}{N^{\frac{1}{3}}}\right)$ (coarse) correlated equilibrium of the corresponding N-player game.

Proof. The proof follows the very same path as the proof of Theorem 45 until the last step: integration over ρ.

In the case when ρ is not a finite sum of diracs, given the bound $\frac{1}{\nu_m N}$, it could be that ρ assigns mass on a sequence of ν for which $\nu(\pi) \to 0$. We however note that, given a threshold $\alpha \in\ ]0, 1[$, we can separate, using the triangular inequality, policies whose selection probability according to ν is lower than α from those for which it is higher than α:

$$\mathbb{E}\left[\|\mu^N - \mu(\nu)\|_2\right] = \mathbb{E}\left[\|\sum_{\pi} \frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right]$$
$$\leq \mathbb{E}\left[\|\sum_{\pi|\nu(\pi)>\alpha} \frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right] + \sum_{\pi|\nu(\pi)\leq\alpha} \mathbb{E}\left[\|\frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right]$$

We examine the second term for a given policy π, which contains only policies whose selection probability is lower than α.

$$\mathbb{E}\left[\|\frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right] \leq \mathbb{E}\left[\left|\frac{N_\pi}{N} - \nu(\pi)\right| \underbrace{\|\mu_\pi^N\|_2}_{\leq\sqrt{T}} + \nu(\pi) \underbrace{\|\mu_\pi^N - \mu_\pi(\nu)\|_2}_{\leq\sqrt{2T}}\right]$$
$$\leq \mathbb{E}\left[\sqrt{T}\left|\frac{N_\pi}{N} - \nu(\pi)\right| + \sqrt{2T}\nu(\pi)\right]$$
$$\leq \frac{\sqrt{T}}{N}\sqrt{\mathbb{E}\left[\left(N_\pi - N\nu(\pi)\right)^2\right]} + \sqrt{2T}\nu(\pi)$$
$$\leq \frac{\sqrt{T}}{N}\sqrt{N\nu(\pi)(1-\nu(\pi))} + \sqrt{2T}\nu(\pi)$$
$$\leq \sqrt{\frac{\alpha T}{N}} + \sqrt{2T}\alpha$$

We have of course that for each $t \in \mathcal{T}$, using similar steps, and assuming $\alpha \leq \frac{K}{\sqrt{N}}$ with $K \in \mathbb{R}_+^*$ a given constant,

$$\mathbb{E}\left[\|\frac{N_\pi}{N}\mu_\pi^N(t) - \nu(\pi)\mu_\pi(\nu)(t)\|_2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

We now make the point that the whole demonstration can be done while only considering policies whose play probabilities is $> \frac{1}{\sqrt{N}}$, up to a $\mathcal{O}\left(\frac{|\Pi|}{\sqrt{N}}\right)$ term. It is not straightforward to find the minimum value of $\frac{1}{\sqrt{N\alpha}} + |\Pi|\left(\sqrt{\frac{\alpha T}{N}} + \sqrt{2T}\alpha\right)$ when varying α. However, we are only interested in $\mathcal{O}$ relationships. Now, proceeding by degree analysis, we realize that, writing $\alpha = N^x$, the value of x for which $\frac{1}{\sqrt{N\alpha}}$, α and $\sqrt{\frac{\alpha}{N}}$ have the same degree is $x = \frac{-1}{3}$. We therefore take α to be $\frac{K}{N^{\frac{1}{3}}}$, which transforms term into $\mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right)$.

Indeed, at each step of the proof, for each time $t \in \mathcal{T}$, for policies whose play probability is $> \frac{K}{\sqrt{N}}$, the only term which involves other policies is $\mathbb{E}\left[\|\mu^N(t) - \mu(\nu)(t)\|_2\right]$.

This term can be separated in two using the triangular inequality, between policies whose play probability is lower than $\frac{K}{\sqrt{N}}$, and policies whose play probability isn't. The first group adds at most a $\mathcal{O}\left(\frac{|\Pi|}{\sqrt{N}}\right)$ term. The second term, a $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ with partial dependency on some ν for which all interesting components are $> \frac{1}{\sqrt{N}}$.

More specifically, writing $\nu_{m,\alpha} = \min_{\pi|\nu(\pi)>\alpha} \nu(\pi)$ and noting that $\frac{1}{\sqrt{\nu_{m,\alpha}}} < \frac{1}{\sqrt{\alpha}}$,

$$\mathbb{E}_{\nu\sim\rho}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[\|\mu^N - \mu(\nu)\|_2\right]\right] = \mathbb{E}_{\nu\sim\rho}[\underbrace{\mathbb{E}\left[\|\sum_{\pi|\nu(\pi)>\alpha}\frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right]}_{\leq\frac{1}{\sqrt{N\nu_{m,\alpha}}}} + \underbrace{\sum_{\pi|\nu(\pi)\leq\alpha}\mathbb{E}\left[\|\frac{N_\pi}{N}\mu_\pi^N - \nu(\pi)\mu_\pi(\nu)\|_2\right]}_{\leq|\Pi|(\sqrt{\frac{\alpha T}{N}}+\sqrt{2T\alpha})}]$$

$$\leq \frac{K'}{\sqrt{N\nu_{m,\alpha}}} + |\Pi|(\sqrt{\frac{\alpha T}{N}} + \sqrt{2T\alpha})$$

$$\leq \frac{K'}{\sqrt{N\alpha}} + |\Pi|(\sqrt{\frac{\alpha T}{N}} + \sqrt{2T\alpha})$$

We look for the optimal value of α while remembering that we must have $\alpha \leq \frac{K}{\sqrt{N}}$ with K independent of N for the above developments to remain true.

It is not straightforward to find the minimum value of $\frac{1}{\sqrt{N\alpha}} + |\Pi|\left(\sqrt{\frac{\alpha T}{N}} + \sqrt{2T\alpha}\right)$ when varying α. However, we are only interested in $\mathcal{O}$ relationships. Now, proceeding by degree analysis, we realize that, writing $\alpha = N^x$, the value of x for which $\frac{1}{\sqrt{N\alpha}}$ and α have the same degree is $x = \frac{-1}{3}$, and $\sqrt{\frac{\alpha}{N}}$ has a lower degree than both in this case. However, this doesn't respect the constraint that $\alpha \leq \frac{K}{\sqrt{N}}$. Looking at the order relationship between different terms' exponents, we therefore take α to be $\frac{K}{N^{\frac{1}{4}}}$, which transforms term into $\mathcal{O}\left(\frac{1}{N^{\frac{1}{4}}}\right)$.

Going back to the initial developments of the proof, we have that, if ρ is discrete

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))\right] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right),$$

which means that

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \epsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

If ρ is continuous,

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[J(u(\pi),\mu(\nu)) - J(\pi,\mu(\nu))\right] + \mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right),$$

which means that

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[\mathbb{E}_{\mu^N\sim\mu(\nu)}\left[J(u(\pi),\mu^N_{-\pi,u(\pi)}) - J(\pi,\mu^N)\right]\right] \leq \epsilon + \mathcal{O}\left(\frac{1}{N^{\frac{1}{3}}}\right),$$

which concludes the proof. □

4.5 Limitations and Future Work

Correlated and coarse-correlated equilibria are an important class of equilibria; however, there exists a plethora of other, derived equilibria especially designed for extensive-form games [121], which are easier to learn and reach, yet carry the same stability flavour, and could widely benefit from being introduced to the Mean-Field setting. Another question of interest, treated in the next chapter, relates to reaching these equilibria: what algorithms can reach correlated and coarse-correlated equilibria? How quickly do they do so? Under which condition? We investigate these questions in the next chapter.

Chapter 5

Learning to man sails: Learning Equilibria in Mean-Field Games

We have defined, in Chapter 4, Mean-Field correlated and coarse-correlated equilibria, yet have not yet provided algorithms to find these in our setting. This section is concerned with this question. It first generalizes the notion of *regret* to the Mean-Field setting in Section 5.1, and links it with correlated and coarse-correlated equilibria. Once this is done, in Section 5.2, we prove that both Mean-Field Online Mirror Descent, and an alteration of Fictitious Play, Joint Fictitious Play, are external-regret minimizing in *all games*, which means that their *empirical distribution*, a term we rigorously define in Section 5.1, converges to a coarse-correlated equilibrium. Section 5.3 introduces Mean-Field PSRO, which converges to Nash, correlated and coarse-correlated equilibria in *all games*.

5.1 Regret Minimization and Empirical Play

There are strong connections between game-theoretic equilibria and regret minimisation in online learning. A core result [21] states that if all players follow a regret-minimizing algorithm to select their strategy, then the (learning-)time average of their joint behaviour converges to the set of coarse correlated equilibria. This connection provides a means of computing approximate equilibria which has been fundamental to recent advances in the state-of-the-art of games such as heads-up no-limit poker [30, 119].

Regret minimization has surprisingly been understudied in the Mean Field Games literature. In this section, we describe a corresponding connection between regret-minimizing algorithms and Mean-Field coarse correlated equilibria, which serve as the basis for deriving convergence results of learning equilibria in Section 5.2.

5.1.1 Empirical Play

A continuous-time learning algorithm generates a continuous-time, measurable sequence of policies $(\pi_s)_{0\leq s\leq t}$. A correlation device is extracted from this sequence by recommending a policy from a uniformly-selected moment of play: it is the *empirical play*.

Definition 36 (Empirical Play)**.** The empirical play $\hat{\rho} \in \mathcal{P}(\Delta(\Pi))$ of the sequence of policies $(\pi_s)_{0\leq s\leq t}$ is the correlation device resulting from uniformly recommending each deterministic component of one stochastic policy selected at random among $(\pi_s)_{0\leq s\leq t}$.

More formally, in the continuous case, this yields

$$\hat{\rho}(A) = \frac{1}{t}\int_0^t \mathbb{1}\{\nu \in A \mid \pi_s = \pi(\nu)\}\mathrm{d}s\,.$$

In the discrete case, this yields

$$\hat{\rho}(\nu) = \frac{1}{t}\sum_{s=1}^{t} \delta_{\pi_s = \pi(\nu)}\,,$$

The motivation for introducing the notion of empirical play is that several key results in this section establish that if each member of a population that played the sequence of policies $(\pi_s)_{0\leq s\leq t}$ is relatively happy with their choice of policies in hindsight, in a sense made precise below, then the corresponding empirical play correlation device is an approximate equilibrium for the Mean-Field game under consideration.

To evaluate how close to optimal the empirical play is, we define policy alterations which characterize the expected deviation payoffs when one follows it.

Definition 37 (Policy Alterations)**.** The set of **Policy Alterations** $\mathcal{U}_A$ is the set of functions $\bar{\Pi} \to \bar{\Pi}$ such that $u \in \mathcal{U}_A$ is a policy alteration if there exists a function $u' \in \mathcal{U}_{CE}$ such that for all $\bar{\pi} = \sum_{\pi\in\Pi} \alpha_\pi \pi$, $u(\bar{\pi}) = \sum_{\pi\in\Pi} \alpha_\pi u'(\pi)$

Informally, a policy alteration of $\bar{\pi} \in \bar{\Pi}$ is a function that swaps around deterministic policies' mass in the composition of $\bar{\pi}$.

The set of **Coarse Policy Alterations** $\mathcal{U}_{CA}$ is the subset of $\mathcal{U}_A$ composed only of constant functions.

The remainder of this section is devoted to formalising the relationship between regret minimization and both correlated equilibria and coarse correlated equilibria, and the question of how such sequences of policies can be generated algorithmically is addressed in Section 5.2.

5.1.2 External Regret and Coarse Correlated Equilibria

Consider a representative agent in a Mean-Field game, using policy π_s at time s, against a population distribution μ_s. The cumulative return of the agent over a time interval $[0, t]$ is given by

$$\int_0^t J(\pi_s, \mu^s)\mathrm{d}s\,.$$

A natural question to consider is how better the agent could have done in hindsight by sticking with a fixed policy π throughout the interval $[0, t]$, in contrast to using the sequence $(\pi_s)_{0\leq s\leq t}$. The increase in payoff that the agent could have received is referred to as the *regret* of not having played π. The *external regret* of a policy sequence codifies the worst-case regret against a fixed policy.

Definition 38 (External regret)**.** Given a sequence of population distributions $(\mu_s)_{0\leq s\leq t}$, the *external regret* of a policy sequence $(\pi_s)_{0\leq s\leq t}$ is given by

$$\mathrm{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu_s)_{0\leq s\leq t}) = \sup_{\pi\in\Pi} \int_0^t J(\pi, \mu^s)\mathrm{d}s - \int_0^t J(\pi_s, \mu^s)\mathrm{d}s\,.$$

Alternatively, an equivalent definition is

$$\mathrm{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu_s)_{0\leq s\leq t}) = \sup_{u\in\mathcal{U}_{CA}} \int_0^t J(u(\pi_s), \mu^s)\mathrm{d}s - \int_0^t J(\pi_s, \mu^s)\mathrm{d}s\,,$$

where the equivalence is immediate when equating $\mathcal{U}_{CA}$ to Π.

For a bounded reward function J, an immediate upper bound on the external regret of a policy sequence $(\pi_s)_{0\leq s\leq t}$ given a population sequence $(\mu_s)_{0\leq s\leq t}$ is $O(t)$. Of particular interest are methods for selecting policy sequences $(\pi_s)_{s\geq 0}$ in the presence of a population sequence $(\mu_s)_{s\geq 0}$ such that the external regret grows as $o(t)$; such a policy sequence is said to be *no-regret*, or

regret-minimising. This interest, in the context of game theory, stems from the close connection between external regret and coarse correlated equilibria; both notions encode the value of deviation to a fixed policy in certain circumstances.

This connection is well-known in non-Mean-Field game theory, and forms the basis for many algorithms for computing equilibria. The following result makes this connection precise in the case of Mean-Field games, and serves as a key motivation for the use of regret-minimisation algorithms for computing coarse correlated equilibria in Mean-Field games, following similar results in non-Mean-Field game theory.

Proposition 49. *Let $\epsilon > 0$ and $(\pi_s)_{0\leq s\leq t}$ be a sequence of policies. Then the following two propositions are equivalent.*

1. $\frac{1}{t}\operatorname{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu^{\pi_s})_{0\leq s\leq t}) \leq \epsilon$
2. *The Empirical Play of $(\pi_s)_{0\leq s\leq t}$ is an ϵ-Mean Field Coarse Correlated Equilibrium.*

Proof. Let select $\epsilon > 0$ and $(\pi_s)_{0\leq s\leq t}$, and name $\hat{\rho}$ the correlation device recommending the empirical play of $(\pi_s)_{0\leq s\leq t}$. Observe that

$$\frac{1}{t}\operatorname{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu^{\pi_s})_{0\leq s\leq t}) = \sup_{\pi'\in\Pi} \mathbb{E}_{\nu\sim\hat{\rho},\pi\sim\nu}[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))]\,,$$

following the definition of $\hat{\rho}$ as recommending the empirical play uniformly: each ν recommended by ρ is derived from uniformly recommending ν_{π_s} over s. We deduce that

$$\frac{1}{t}\operatorname{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu^{\pi_s})_{0\leq s\leq t}) = \sup_{u\in\mathcal{U}_{CCE}} \mathbb{E}_{\nu\sim\hat{\rho},\pi\sim\nu}[J(u(\pi), \mu(\nu)) - J(\pi, \mu(\nu))]\,,$$

hence providing the connection with the coarse correlated equilibrium characterization stated in Definition 28.

Hence, $\hat{\rho}$ is an ϵ-Mean-Field Coarse Correlated Equilibrium if and only if the Average External Regret of $(\pi_s)_{0\leq s\leq t}$ is smaller than ϵ. □

The correspondence between ϵ-external regret and ϵ-coarse correlated equilibria is now established. However, in general, algorithms never really reach 0 regret, and we now wonder: does an asymptotically no-regret algorithm indeed get closer to the set of coarse correlated equilibria as it minimizes regret, or could it actually remain "away" from this set? The following proposition proves that no-external-regret learners do approach the set of CCEs!

Proposition 50. *Let $(\pi_s)_{0\leq s\leq t}$ be such that $\lim_{t\to\infty}\frac{1}{t}\operatorname{ExtReg}((\pi_s)_{0\leq s\leq t}, (\mu^{\pi_s})_{0\leq s\leq t}) = 0$, and assume the reward function r is bounded and the set of coarse correlated equilibria is non-empty. Then the empirical play of π, $\hat{\rho}^t_\pi$, converges to the set of coarse correlated equilibria $\mathcal{C}$,* i.e. $\inf_{\rho_0\in\mathcal{C}} d_{\mathcal{W}_2}(\hat{\rho}^t_\pi, \rho_0) \to 0$, *where $d_{\mathcal{W}_2}$ is the Wasserstein-2 distance.*

Proof. First, notice that, since we are in a finite-time, finite-state setting, r being bounded implies directly that J is bounded. Let us denote by $\mathcal{C}_\epsilon$ is the set of ϵ-CCE, while $\mathcal{C}$ is the set of CCE.

We will prove by contradiction that

$$\forall\alpha > 0,\, \exists\epsilon > 0, \forall\rho \in \mathcal{C}_\epsilon,\ \inf_{\rho_0\in\mathcal{C}} d_{\mathcal{W}_2}(\rho, \rho_0) < \alpha\,. \tag{5.1}$$

Let us suppose that

$$\exists\alpha > 0,\, \forall\epsilon > 0, \exists\rho \in \mathcal{C}_\epsilon,\ \inf_{\rho_0\in\mathcal{C}} d_{\mathcal{W}_2}(\rho, \rho_0) \geq \alpha\,. \tag{5.2}$$

We take a sequence $(\rho_n)_n$ such that

$$\forall n,\ \rho_n \in \mathcal{C}_{\frac{1}{2^n}},\ d_{\mathcal{W}_2}(\rho_n, \rho_0) \geq \alpha\,.$$

Correlation devices are distributions over distributions over $|\Pi|$ elements. The set of distributions over $|\Pi|$ elements is the set of vectors in $\mathbb{R}_+^{|\Pi|}$ which sum to 1. It is compact as a closed and bounded subset of $\mathbb{R}^{|\Pi|}$. All measures over the set of population distributions are therefore, by definition, tight. Since their set is tight, Theorem 5.1 in Billingsley [19] indicates that the set of correlation devices is relatively compact.

Hence, there exists a subsequence of $(\rho_n)_n$, denoted $(\rho'_n)_n$ converging weakly towards a point $\bar{\rho}$. Since $\mathbb{R}^{|\Pi|}$ is Polish, $(\rho'_n)_n$ converges towards $\bar{\rho}$ with respect to the Wasserstein distance $d_{\mathcal{W}_2}$.

We note that the deviation-payoff function $\rho \to \max_{\pi \in \Pi} \int_\nu \rho(d\nu) \left(J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\right)$ is continuous (It is the max over the integral over the finite set Π of continuous functions of ρ) provided J is bounded. Hence, since $\rho^n \in \mathcal{C}_{\frac{1}{2^n}}$, $\bar{\rho}$ must be a coarse correlated equilibrium. This contradicts equation 5.2 so that equation 5.1 holds.

Moreover, equation 5.1 directly implies that

$$\forall \alpha > 0,\ \exists \epsilon_\alpha > 0,\ \forall \epsilon \leq \epsilon_\alpha,\ \forall \rho \in \mathcal{C}_\epsilon,\ \inf_{\rho_0 \in \mathcal{C}} d_{\mathcal{W}_2(\rho, \rho_0)} < \alpha\ ,$$

since the sets $(\mathcal{C}_\epsilon)_{\epsilon \leq \epsilon_\alpha}$ are included into the set of ϵ_α-coarse correlated equilibria.

We define a sequence α_n which converges to 0, and a subsequence $\phi(n)$ such that, $\forall n$, $\phi(n)$ is the first n from which $(\hat{\rho}^n_\pi)_n$ is an ϵ_{α_n}-CCE and after which it never becomes a worse equilibrium. We know that $\forall t \geq \phi(n)$, $\hat{\rho}^t_\pi$ is also an ϵ_{α_n}-CCE, and therefore $\forall t \geq \phi(n)$, $\inf_{\rho_0 \in \mathcal{C}} d_{\mathcal{W}_2}(\hat{\rho}^t_\pi, \rho_0) < \alpha_n$ as well.

Thus $\forall \epsilon > 0$, $\exists N \geq 0$, $\forall t \geq N$, $\inf_{\rho_0 \in \mathcal{C}} d_{\mathcal{W}_2}(\hat{\rho}^t_\pi, \rho_0) < \epsilon$, and thus $\inf_{\rho_0 \in \mathcal{C}} d_{\mathcal{W}_2}(\hat{\rho}^t_\pi, \rho_0) \to 0$.

□

5.1.3 Swap Regret and Correlated Equilibria

A second naturally arising question is: given the output of an algorithm over several timesteps, had the agent swapped its policies for other policies (that is, every time it was recommended to play π_1, it chose to play π_2 instead), could they have received a higher payoff? This is the definition of swap regret [21, 69]: given a policy alteration u, what is the difference between our received payoff and the maximal payoff, were we to have altered our play using the best possible u?

More formally,

Definition 39 (Swap Regret)**.** Given a sequence of policies $(\pi_s)_{1 \leq s \leq t}$ and a sequence of population distributions $(\mu_s)_{1 \leq s \leq t}$, we define swap regret as

$$\text{SwapReg}((\pi_s)_{1 \leq s \leq t}) = \sup_{u \in \mathcal{U}_A} \int_s J(u(\pi_s), \mu_s) - J(\pi_s, \mu_s) ds$$

Proposition 51. *Let $\epsilon > 0$ and $(\pi_s)_{0 \leq s \leq t}$ be a sequence of policies. Then the following two propositions are equivalent.*

1. $\frac{1}{t}\,\text{SwapReg}((\pi_s)_{0 \leq s \leq t}, (\mu^{\pi_s})_{0 \leq s \leq t}) \leq \epsilon$;
2. *The Empirical Play of $(\pi_s)_{0 \leq s \leq t}$ is an ϵ-Mean Field Correlated Equilibrium.*

Proof. Let $\epsilon > 0$ and $(\pi_s)_{0 \leq s \leq t}$ a history of policies. We begin this proof by noting that for all $0 \leq s \leq t$, all policies π and Mean-Field flow μ,

$$J(\pi, \mu) = \sum_{\tilde{\pi} \in \Pi} \nu_\pi(\tilde{\pi}) J(\tilde{\pi}, \mu),$$

where we recall that $\forall \bar{\pi} \in \bar{\Pi}$, $\pi(\nu_{\bar{\pi}}) = \bar{\pi}$.

We thus have

$$\begin{aligned}\frac{1}{t}\,\mathrm{SwapReg}((\pi_s)_{0\le s\le t},(\mu^{\pi_s})_{0\le s\le t}) &= \frac{1}{t}\sup_{u\in\mathcal{U}_A}\int_s J(u(\pi_s),\mu^{\pi_s})-J(\pi_s,\mu^{\pi_s})ds\\ &= \frac{1}{t}\sup_{u\in\mathcal{U}_{CE}}\int_s\sum_{\pi\in\Pi}\nu_{\pi_s}(\pi)\left(J(u(\pi),\mu^{\pi_s})-J(\pi,\mu^{\pi_s})\right)ds\\ &= \sup_{u\in\mathcal{U}_{CE}}\mathbb{E}_{\pi\sim\nu_{\pi_s},\,\nu_{pi_s}\sim Uniform((\pi_t)_t)}[J(u(\pi),\mu^{\pi_s})-J(\pi,\mu^{\pi_s})]\\ &= \sup_{u\in\mathcal{U}_{CE}}\mathbb{E}_{\pi\sim\nu,\nu\sim\hat{\rho}}[J(u(\pi),\mu^{\pi_s})-J(\pi,\mu^{\pi_s})],\end{aligned}$$

with $\hat{\rho}$ the empirical play of $(\pi_s)_{0\ge s\ge t}$, which concludes the proof.

□

Once again, we may wonder what happens when a no-regret algorithm learns: does it go closer to the set of correlated equilibria? The following proposition answers this question positively.

Proposition 52. *Let $(\pi_s)_{0\le s\le t}$ be such that $\lim_{t\to\infty}\frac{1}{t}\,\mathrm{SwapReg}((\pi_s)_{0\le s\le t},(\mu_s)_{0\le s\le t})=0$. Then the empirical play of $(\pi_s)_{0\le s\le t}$ converges to the set of correlated equilibria,* i.e. $\min_{\rho_0\in\mathcal{C}} d_{\mathcal{W}_2}(\rho_\nu,\rho_0)\to 0$.

Proof. The proof follows the same steps as that of Proposition 50, the only change being the set of deviations considered and the deviation payoff function. Since the deviation payoff function remains continuous, the proof remains unchanged. □

5.2 Learning Coarse-Correlated Equilibria in Mean Field Games

Now that we have introduced new equilibrium concepts for Mean-Field games, a new question must be asked: how can they be algorithmically reached? This section provides new insights on various learning algorithms that are known to efficiently learn Nash equilibria in Mean Field Games under certain conditions, including Nash unicity.

More specifically, we focus on three algorithms, which we apply to Mean Field games that do not necessarily satisfy monotonicity or contractivity properties. We study Online Mirror Descent [147]'s convergence properties without assuming monotonicity; we also present a new version of Fictitious Play [149], *Joint Fictitious Play*, and prove that both Online Mirror Descent and Joint Fictitious Play are no-external-regret. As we proved in Section 5.1, this means that their empirical plays converge towards the set of coarse correlated equilibria.

We remark once again that these results do not require *any condition* on the games played, provided they fit our framework, in particular, they do not require any monotonicity or contractivity properties to be true.

5.2.1 Mean-Field Joint Fictitious Play

Using Fictitious play algorithms to learn Nash equilibria in games dates back to the seminal papers of Brown [28] and Robinson [155]. Its extension to Mean Field games has been considered in [36, 75], while its rate of convergence has been discussed in [149] when learning in continuous time and in [66] when learning in discrete time. We focus here on frameworks of games for which several Nash equilibria may exist and present a variant of Fictitious Play in continuous learning time.

Continuous-Time Joint Fictitious Play Algorithm

In Joint Fictitious Play (Joint FP), at every step, the agents all play simultaneously the same policy which is sampled from the past best responses. In continuous time, at time s, each best

response is computed as:

$$\pi_\tau^{BR} = \arg\max_{\pi' \in \Pi} \int_{s=0}^{\tau} \langle \mu^{\pi'}, r^{\pi'}(\cdot, \mu^{\pi_s}) \rangle ds \ ,$$

$$\mu^{\pi_\tau}(x) = \frac{1}{\tau} \int_0^{\tau} \mu^{\pi_s^{BR}}(x) ds \ .$$

Remark 10. *An intuition for the reason why we need a different algorithm for no-regret-learning while in N-player games, traditional fictitious play is no-regret comes from the Mean-Field non-linearity problem, highly highlighted by [126]: while Joint FP and FP are the same in the N-player setting - in those, the reward against an averaged policy is the same as the average reward against each policy -, they are different here, and only Joint FP directly minimizes external regret. It is unclear whether FP also minimizes external regret, or if there are cases where it would not.*

Regret Minimization

The convergence of continuous-time FP to the set of mixed Nash equilibria in the context of monotone Mean Field Games has been derived in [149]. It can encompass the presence of common noise in the dynamics and the derived convergence rate is of order $O(1/\tau)$. This convergence property requires the consideration of Mean Field Games satisfying the classical monotonicity condition, ensuring in particular the uniqueness of Nash equilibrium.

Whenever the monotonicity condition is not satisfied, we verify that a small alteration to continuous-time FP, continuous-time Joint FP, converges to a coarse correlated equilibrium. This is proven from the external regret minimization property of Joint FP.

Following a similar line of argument as in [143], we now demonstrate that continuous time JFP converges to a MF-CCE (observe that the monotonicity assumption is not required).

Proposition 53. *For continuous time JFP, at time τ, the regret* $\text{ExtReg}\left((\pi(s))_{0\leq s\leq \tau}, (\mu^{\pi(s)})_{0\leq s\leq \tau}\right)$ *of the continuous time FP policy is of order $O(1/t)$.*

Proof. For $\tau > 0$ and by definition of $\pi^{BR(\tau)}$, the envelope theorem [58] ensures that

$$\frac{d}{d\tau}\left[\max_{\pi'} \int_{s=0}^{\tau} \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds\right] = \langle \mu^{\pi^{BR(\tau)}}, r^{\pi^{BR(\tau)}}(., \mu^{\pi^\tau}) \rangle.$$

Integrating between an arbitrary time $\tau_0 > 0$ and T, this directly implies

$$\begin{aligned}
&\max_{\pi'} \int_{s=0}^{T} \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds - \max_{\pi'} \int_{s=0}^{\tau_0} \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds \\
&= \int_{\tau_0}^{T} \frac{d}{d\tau}\left[\max_{\pi'} \int_{s=0}^{\tau} \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds\right] d\tau \\
&= \int_{\tau_0}^{T} \langle \mu^{\pi^{BR(\tau)}}, r^{\pi^{BR(\tau)}}(., \mu^{\pi^\tau}) \rangle d\tau \\
&= \int_{\tau=0}^{T} \langle \mu^{\pi^{BR(\tau)}}, r^{\pi^{BR(\tau)}}(., \mu^{\pi^\tau}) \rangle d\tau - \int_{\tau=0}^{\tau_0} \langle \mu^{\pi^{BR(\tau)}}, r^{\pi^{BR(\tau)}}(., \mu^{\pi^\tau}) \rangle d\tau \, .
\end{aligned}$$

Finally, we deduce that

$$\max_{\pi'} \int_0^T \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds - \int_0^T \langle \mu^{\pi^{BR(s)}}, r^{\pi^{BR(s)}}(., \mu^{\pi^s}) \rangle ds$$
$$= \max_{\pi'} \int_0^{\tau_0} \langle \mu^{\pi'}, r^{\pi'}(., \mu^{\pi^s}) \rangle ds - \int_0^{\tau_0} \langle \mu^{\pi^{BR(s)}}, r^{\pi^s}(., \mu^{\pi^{BR(s)}}) \rangle ds \,.$$

Hence, the previous left hand side expression is $O(1)$ implying that the external regret is $O(1/t)$. □

Discrete-Time Joint Fictitious Play algorithm

We describe here a discretization of the above continuous algorithm in Algorithm 17, whose empirical convergence properties are illustrated in Section 5.2.4.

Algorithm 17 Joint Fictitious Play in Mean Field Games

Require: Initial policy π_0
1: $\bar{\pi}_0 = \pi_0$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $\pi_t^{BR} = \arg\max_{\pi_t^{BR} \in \Pi} \sum_{i=0}^t \langle \mu^{\pi_t^{BR}}, r^{\pi_t^{BR}}(\cdot, \mu^{\pi_t}) \rangle$.
4: Compute $\bar{\mu}_t = \frac{t-1}{t} \bar{\mu}_{t-1} + \frac{1}{t} \mu^{\pi_t^{BR}}$.
5: Compute $\pi_t = \pi(\bar{\mu}_t)$.
6: **end for**
7: **return** Collection of policies $(\pi_t)_t$

Dominated Strategy Exclusion

Finally, we investigate the relationship between Joint FP's empirical play and dominated strategies: do we have guarantees that Joint FP's computed equilibrium will not include dominated strategies? How about pre-asymptotic behavior, how quickly are dominated strategies eliminated from play?

Proposition 54 (Fictitious Play Pareto-Optimality)**.** *Let $(\pi_t)_{t \in]0;T]}$ be the policies produced by Fictitious Play by time $T > 0$. Then a policy sampled from this set will asymptotically almost-surely never be dominated as $T \to \infty$, and the probability of sampling a dominated strategy is $\leq \frac{1}{T}$.*

Proof. We begin the proof by recalling the definition of a dominated policy: $\pi \in \Pi$ is dominated if there exists $\pi' \in \Pi$, $\forall \mu,\ J(\pi', \mu) > J(\pi, \mu)$).

We note that $\forall t > 0,\ \pi^{BR}(t)$ can by definition not be dominated, since it is defined as $\arg\max_{\pi'} \int_{s=0}^t J(\pi', \mu_s) ds$: if π' dominated $\pi^{BR}(t)$, then $\int_{s=0}^t J(\pi', \mu_s) ds > \int_{s=0}^t J(\pi_t, \mu_s) ds$, which is contradictory.

Therefore, the only potentially dominated strategy among the mixture that defines π^t is π_0 : the probability that π^t plays according to a dominated strategy is therefore at most the probability that π^t plays π_0.

The policy-mixing distribution is continuous, so this probability is null for all $t > 1$, and potentially equal to 1 for $t \in [0; 1]$. We therefore have $\mathbb{P}\left(\text{Sampling actions following } \pi_0 \text{ from } \pi_t \mid t\right) = \frac{1}{T}$ if $t \geq 1$.

All in all, we have

$$\mathbb{P}\left(\text{Playing dominated strategy in a game}\right) \leq \mathbb{P}\left(\text{Playing according to } \pi_0 \text{ in a game.}\right)$$
$$\leq \frac{1}{T}$$

□

5.2.2 Mean-Field Online Mirror Descent

We now turn to Online Mirror descent algorithms for mean field games as studied in [147].

Continuous-Time Mean-Field Online Mirror Descent

Algorithm 18 Discrete-Time Online Mirror Descent

Require: N number of actions, $\eta > 0$ learning rate, τ_{max} max learning steps.
1: $\tau = 0$.
2: $y_0 = 0$.
3: $\pi_0 =$ Uniform policy.
4: **while** $t = 1, \dots, T$ **do**
5: Observe current Q-value $Q^{\pi_t}(x, \cdot) \quad \forall x$.
6: Set $y_{t+1}(x, \cdot) = y_t(x, \cdot) + \eta Q^{\pi_t}(x, \cdot) \quad \forall x$.
7: Compute $\pi_t(x, \cdot) = \text{softmax}(y_t(x, \cdot))$.
8: **end while**
9: **return** Collection of policies $(\pi_t)_t$

For the Online Mirror Descent algorithm, [147] introduce a regularizer $h : \Delta(\mathcal{A}) \to \mathbb{R}$, that is assumed to be ρ-strongly convex for some constant $\rho > 0$. Its conjugate $h^* : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ is defined as $h^*(y) = \max_{\pi \in \Pi}[< y, \pi > -h(\pi)]$. When h has good properties we have

$$\Gamma(y) := \nabla h^*(y) = \arg\max_{\pi}[< y, \pi > -h(\pi)] \, . \tag{5.3}$$

The continuous-time Online Mirror Descent dynamics are defined as

$$y_t(x, a, \tau) = \int_0^\tau Q_t^{\pi(s), \mu^{\pi(s)}}(x, a) ds, \quad t \in \mathcal{T} \tag{5.4}$$

$$\pi_t(. \mid x, \tau) = \Gamma(y_n(x, ., \tau)), \quad t \in \mathcal{T} \tag{5.5}$$

where we define $Q^{\pi,\mu} = (Q_t^{\pi,\mu})_{t \in \mathcal{T}}$ and, with $T = \max_{t \in \mathcal{T}} t$:

$$\begin{cases} Q_T^{\pi,\mu}(x, a) = 0 \\ Q_t^{\pi,\mu}(x, a) = r(x, a, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, a, \mu_t) \sum_{a'} \pi_t(x, a') Q_{t+1}^{\pi,\mu}(x', a'), \\ \qquad t = T-1, T-2, \dots, 0. \end{cases}$$

where we assume, without loss of generality, that $\mathcal{T}$ is the sequence 0, ..., T.

Convergence Properties

We characterize the regret-minimizing properties of Online Mirror Descent.

Theorem 55. *Online Mirror Descent is a regret minimizing strategy in Mean Field games (no monotonicity required):*

$$\frac{1}{\tau} \text{ExtReg}((\pi(s))_{0 \le s \le \tau}; (\mu^{\pi(s)})_{0 \le s \le \tau}) = O(\frac{1}{\tau})$$

Proof. We introduce $t \in \mathcal{T}$ the game time. In the following arguments, we draw the reader's attention towards the distinction between game time t and learning time τ.

We define, for all $\pi \in \bar{\Pi}$, and for y, Q and $\pi(\tau)$ the quantities defined above,

$$\mathcal{L}(\pi, y(., ., \tau)) = \sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}} \mu_t^\pi(x)[h^*(y_t(x, ., \tau)) - h^*(y_{\pi,t}(x, .)) - \langle \pi_t, y_t(x, ., \tau) - y_{\pi,t}(x, .) \rangle]$$

where y_π is such that $\pi(. \mid x) = \Gamma(y_\pi(x, .))$.

We can deduce that $\frac{d}{d\tau}\mathcal{L}(\pi, y(., ., \tau)) = V_0^{\pi(\tau),\mu^{\pi(\tau)}} - V_0^{\pi,\mu^{\pi(\tau)}}$.

Indeed:

$$
\begin{aligned}
&\frac{d}{d\tau}\mathcal{L}(\pi, y(., ., \tau)) \\
&= \frac{d}{d\tau}\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu^\pi(x)_t[h^*(y_t(x, ., \tau)) - h^*(y_{\pi,t}(x, .)) - < \pi_t, y_t(x, ., \tau) - y_{\pi,t}(x, .) >] \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)\frac{d}{d\tau}[h^*(y_t(x, ., \tau)) - h^*(y_{\pi,t}(x, .)) - < \pi_t, y_t(x, ., \tau) - y_{\pi,t}(x, .) >] \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)[\frac{d}{d\tau}h^*(y_t(x, ., \tau)) - < \pi_t, \frac{d}{d\tau}y_t(x, ., \tau) >] \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)[< \frac{d}{d\tau}y_t(x, ., \tau), \nabla h^*(y_t(x, ., \tau)) > - < \pi_t, \frac{d}{d\tau}y_t(x, ., \tau) >] \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)[\underbrace{< Q_t^{\pi(\tau),\mu_\tau}(x, .), \pi(x, ., \tau) >}_{V_t^{\pi(\tau),\mu_\tau}(x)} - < \pi_t, Q_t^{\pi(\tau),\mu_\tau}(x, .) >] \\
&< \pi_t, Q_t^{\pi(\tau),\mu_\tau}(x, .) > \\
&= \sum_a \pi_t(x, a)[r(x, a, \mu_\tau) + \sum_{x'\in\mathcal{X}} p(x' \mid x, a, \mu_t)V_{t+1}^{\pi(\tau),\mu_\tau}(x')] \\
&= \underbrace{\sum_a \pi_t(x, a)[r(x, a, \mu_\tau) + \sum_{x'\in\mathcal{X}} p(x' \mid x, a, \mu_\tau)V_{t+1}^{\pi,\mu_\tau}(x')]}_{=V_t^{\pi,\mu_\tau}(x)} + \sum_a \pi_t(x, a)\sum_{x'\in\mathcal{X}} p(x' \mid x, a, \mu_t)[V_{t+1}^{\pi(\tau),\mu_\tau}(x') - V_{t+1}^{\pi,\mu_\tau}(x')] \\
&= V_t^{\pi,\mu_\tau}(x) + \sum_a \pi_t(x, a)\sum_{x'\in\mathcal{X}} p(x' \mid x, a, \mu_t)[V_{t+1}^{\pi(\tau),\mu_\tau}(x') - V_{t+1}^{\pi,\mu_\tau}(x')] \\
&\frac{d}{d\tau}\mathcal{L}(\pi, y(., ., \tau)) \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)[V_t^{\pi(\tau),\mu_\tau}(x) - V_t^{\pi,\mu_\tau}(x)] - \sum_{t\in\mathcal{T}}\underbrace{\sum_{x\in\mathcal{X}}\mu_t^\pi(x)\pi_t(x, a)\sum_{x'\in\mathcal{X}} p(x' \mid x, a, \mu_t)}_{=\sum_{x'\in\mathcal{X}}\mu_{t+1}^\pi(x')}[V_{t+1}^{\pi(\tau),\mu_\tau}(x') - V_{t+1}^{\pi,\mu_\tau}(x')] \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_t^\pi(x)[V_t^{\pi(\tau),\mu_\tau}(x) - V_t^{\pi,\mu_\tau}(x)] - \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\mu_{t+1}^\pi(x)[V_{t+1}^{\pi(\tau),\mu_\tau}(x) - V_{t+1}^{\pi,\mu_\tau}(x)] \\
&= \sum_{x\in\mathcal{X}}\mu_0^\pi(x)[V_0^{\pi(\tau),\mu_\tau}(x) - V_0^{\pi,\mu_\tau}(x)] \\
&= V_0^{\pi(\tau),\mu_\tau} - V_0^{\pi,\mu_\tau} \qquad (5.6)
\end{aligned}
$$

The proof is concluded by saying:

$$
\begin{aligned}
\text{ExtReg}((\pi(\tau))_{0\leq\tau\leq\tau_0}; (\mu^{\pi(\tau)})_{0\leq\tau\leq\tau_0}) &= \max_\pi \int_0^{\tau_0} V_0^{\pi,\mu_\tau} - V_0^{\pi(\tau),\mu_\tau} d\tau \\
&= \max_\pi \int_0^{\tau_0} -\frac{d}{d\tau}\mathcal{L}(y(., ., \tau))d\tau \\
&= \max_\pi [\mathcal{L}(\pi, y(., ., 0)) - \mathcal{L}(\pi, y(., ., \tau_0))] \qquad (5.7)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathcal{L}(\pi, y(.,.,0)) - \mathcal{L}(\pi, y(.,.,\tau_0)) \\
&= \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}} \mu_t^\pi(x)[h^*(y_t(x,.,0)) - \langle \pi_t, y_t(x,.,0)\rangle \underbrace{-h^*(y_t(x,.,\tau_0)) + \langle \pi_t, y_t(x,.,\tau_0)\rangle}_{\leq h(\pi_t)}] \\
&\leq \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}} \mu_t^\pi(x) \left[\underbrace{h^*(y_t(x,.,0)) - \langle \pi_t(0)(x,.), y_t(x,.,0)\rangle}_{=-h(\pi_t(0)(x,.)} + h(\pi_t) \underbrace{-\langle \pi_t - \pi_t(0)(x,.), y_t(x,.,0)\rangle}_{\leq \|y(.,.,0)\|_{+\infty}}\right] \\
&\leq \sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}} \mu_t^\pi(x) \left[h(\pi_t) - h(\pi_t(0)(x,.)) + \|y(.,.,0)\|_{+\infty}\right] \\
&\leq \underbrace{\left(\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}} \mu_t^\pi(x)\right)}_{T} [h_{\max} - h_{\inf} + \|y(.,.,0)\|_{+\infty}]
\end{aligned}
$$

In the end by combining those two results we have

$$
\frac{1}{\tau_0}\,\mathrm{ExtReg}((\pi(\tau))_{0\leq\tau\leq\tau_0}; (\mu^{\pi(\tau)})_{0\leq\tau\leq\tau_0}) \leq \frac{T}{\tau_0}\left(h_{\max} - h_{\min} + \|y(.,.,0)\|_{+\infty}\right)
$$

□

We note that we obtain convergence bounds for the external regret of Online Mirror Descent, which is in stark contrast with its exploitability, for which, even in the monotonic case, we do not have such bounds. This is due to the fact that what makes average external regret converge is the averaging, and external regret being strictly bounded thanks to it being the sum of past Online Mirror Descent plays: whereas a single Online Mirror Descent policy may be more or less exploitable in ways that are difficult to evaluate, its sequence of policies is difficult to exploit "all at the same time", leading to bounded external regret.

Dominated Strategy Exclusion

Similarly to Joint FP, we investigate OMD's exclusion of dominated strategies and its speed in doing so. Just like Joint FP, OMD's elimination of dominated strategies in its empirical play is $\mathcal{O}\left(\frac{1}{T}\right)$-quick due to the empirical play's uniform average over all previous timesteps.

Proposition 56 (Online-Mirror Descent Optimality)**.** *As t tends to infinity, a policy π uniformly sampled from $(\pi_t)_{t\in[0;T]}$ produced by OMD with entropy regularizer almost-surely never takes $\epsilon > 0$-dominated actions.*

Proof. Let x be a state, a_1 an action ϵ-dominated by a_2, *i.e.* $\forall \mu \in \mathcal{P}(\mathcal{X})$, $\forall \pi \in \Pi$, $Q^{\pi,\mu}(x, a_1) \leq Q^{\pi,\mu}(x, a_2) - \epsilon$ with $\epsilon > 0$. We have that $\pi_t(x) = \mathrm{softmax}(y)$, and $y = \int_0^T Q^{\pi_s,\mu^{\pi_s}} ds$. Directly, $y(x, a_1, t) \leq -\epsilon t + y(x, a_2, t)$, thus $\pi_t(x, a_1) \leq e^{-t\epsilon}\pi_t(x, a_2)$.

Whether a_2 keeps being selected or not, we have necessarily that $\pi_t(x, a_1) \to 0$.

Let $\epsilon' > 0$, $t' > 0$ such that $\forall t > t'$, $\pi_t(x, a_1) < \frac{1}{2}\epsilon'$. Finally, take T such that $\frac{t'}{T} \leq \frac{1}{2}\epsilon'$, and randomly sample π_t from $(\pi_t)_{t\in[0;T]}$.

$$
\begin{aligned}
\mathbb{P}(\pi_t \text{ plays } a_1) &= \underbrace{\mathbb{P}(\pi_t \text{ plays } a_1 \mid t < t')}_{\leq 1}\mathbb{P}(t < t') + \underbrace{\mathbb{P}(\pi_t \text{ plays } a_1 \mid t \geq t')}_{< \frac{1}{2}\epsilon'}\mathbb{P}(t \geq t') \\
&\leq \frac{t'}{T} + \frac{\epsilon'}{2}\underbrace{\frac{T - t'}{T}}_{\leq 1} \\
&\leq \epsilon'
\end{aligned}
$$

There are only a finite amount of states and actions, thus there are only a finite amount of dominated actions. Taking a sup over all possible times T, we have that for all $\epsilon, \epsilon' > 0$, $\exists T' > 0$ such that $\forall T \geq T'$, $\mathbb{P}(\text{Sampled } \pi_t \text{ from } (\pi_t)_{t\in[0;T]} \text{ plays } \epsilon\text{-dominated action}) \leq \epsilon'$, which concludes the proof. □

Neither algorithm presented above converges towards a Mean-Field correlated equilibrium, and one could legitimately wonder whether such an algorithm does exist. Mean-Field PSRO, introduced by Muller et al. [126], and presented below, answers this question by the affirmative.

We show examples of OMD's, JFP's and Mean-Field PSRO's behavior in different games in Section 5.2.4.

5.2.3 Links between Regret and Mean-Field Regret

Given the two above examples, which show that no-N-player-regret algorithms applied to Mean-Field games are no-Mean-Field regret, we are led to the following question: if an algorithm is no-regret, is it no-Mean-Field-regret when applied to a Mean-Field game ?

We call algorithm a function $\mathcal{A} : \left(\mathbb{R}^{\Pi}\right)^{\mathbb{N}} \to \bar{\Pi}$ which takes in a (partial) sequence of payoff vectors, the payoff that each deterministic policy receives when it is played, and returns a new mixed policy to play.

We start by defining clearly what we mean by no-regret.

Definition 40 (No-Regret Algorithm)**.** An algorithm $\mathcal{A}$ is no-internal-regret iff, for all sequence of payoff vectors $(J_t)_t \in \left(\mathbb{R}^{\Pi}\right)^{\mathbb{N}}$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \langle u(\mathcal{A}(J_{\tau\leq t-1})) - \mathcal{A}(J_{\tau\leq t-1}), J_t\rangle = 0, \quad \forall u \in \mathcal{U}_A.$$

If u is restricted to $\mathcal{U}_{CA}$, then the algorithm is no-external-regret.

Given this definition, we have the following property, inspired by the idea of Muller et al. [126] to use no-adversarial-regret algorithms:

Proposition 57. *Let $\mathcal{A}$ be a no-regret learning algorithm. Then $\mathcal{A}$ applied with no modification on a Mean-Field game is no-Mean-Field-regret.*

Proof. Since $\mathcal{A}$ is no-regret, since no-regretness is true for any sequence of payoff functions $(J_t)_t$, and finally, since at time T, $\mathcal{A}$ only depends on $(J_t)_{t\leq T-1}$, we can say that at each time T, $J_T = J(\cdot, \mu^{\mathcal{A}((J_t)_{t\leq T-1})})$, vectorized over all deterministic policies.

Let $u \in \mathcal{U}_A$ for swap regret, $u \in \mathcal{U}_{CA}$ for external regret.

We define the policy sequence $(\pi_t)_t$ by $\pi_T = \mathcal{A}((J_t)_{t\leq T-1})$.

We compute the average Mean-Field regret of $(\pi_t)_t$ for deviation u at time T.

$$\begin{aligned}
\frac{1}{T}\int_0^T J(u(\pi_t), \mu^{\pi_t}) - J(\pi_t, \mu^{\pi_t})\mathrm{d}t &= \frac{1}{T}\sum_{t=0}^{T} J(u(\pi_t), \mu^{\pi_t}) - J(\pi_t, \mu^{\pi_t}) \\
&= \frac{1}{T}\sum_{t=1}^{T} \langle u(\pi_t) - \pi_t, J(\cdot, \mu^{\pi_t})\rangle \\
&= \frac{1}{T}\sum_{t=1}^{T} \langle u(\mathcal{A}(J_{\tau\leq t-1})) - \mathcal{A}(J_{\tau\leq t-1}), J_t\rangle,
\end{aligned}$$

and we know that the last term, by definition, tends to 0, hence $\mathcal{A}$ used in Mean-Field games is no-Mean-Field-regret. □

5.2.4 Experimental Results

The following section presents several experimental results of the algorithms presented so far in this section, Online Mirror Descent (OMD) and Joint Fictitious Play (JFP) on normal form games. Openspiel [101] was used to produce all the figures.

Games of Interest

In order to illustrate the approximation of coarse correlated equilibria by the 3 algorithms described above, we focus our attention on three normal-form, 3-actions (A, B and C) Mean-Field games:

- The dominated-action game, with reward structure

$$\begin{aligned} r(A, \mu) &= \mu(A) + \mu(C)\,, \\ r(B, \mu) &= \mu(B)\,, \\ r(C, \mu) &= \mu(A) + \mu(C) - 0.05\mu(B)\,; \end{aligned}$$

We will use this game to characterize how action C, which is strictly dominated by action A, will be eliminated by different algorithms. It is also interesting to see conditions for algorithms to converge towards playing A only vs. playing B only. We will see that all algorithms eliminate action C, but in different ways and with different speeds.

- The almost-dominated-action game, with reward structure

$$\begin{aligned} r(A, \mu) &= \mu(A) + \mu(C)\,, \\ r(B, \mu) &= \mu(B)\,, \\ r(C, \mu) &= \mu(A) + \mu(C) - 0.05\mu(B)\,; \end{aligned}$$

We use this game as an example which shows that an action needs to be *strictly* dominated to be eliminated - action C is dominated by action A whenever $\mu(B) > 0$, but this domination goes to 0 as $\mu(B)$ tends to 0. We will see that, while Joint FP and Mean-Field PSRO eliminate C in this setting, OMD does not.

- The biased rock-paper-scissors game, with reward structure

$$\begin{aligned} r(A, \mu) &= 0.5 * \mu(B) - 0.3 * \mu(C)\,, \\ r(B, \mu) &= 0.3 * \mu(C) - 0.7 * \mu(A)\,, \\ r(C, \mu) &= 0.7 * \mu(A) - 0.5 * \mu(B)\,; \end{aligned}$$

This game is an example of a non-monotonic game where OMD and Joint FP do not converge to a single point, but instead *cycle*: it shows that pointwise convergence is not always obtained with these two algorithms, and cycling is possible.

Online Mirror Descent

Figures 5.1a and 5.1b show OMD on the almost-dominated action game with different initializations, Figure 5.1c shows OMD on the dominated-action game, Figure 5.1d shows OMD on the Biased RPS game, and Figure 5.2 shows OMD's final policies (determined by a color) as a function on its initial policy (the color's position) on the almost-dominated game. Figures 5.1a, 5.1b, 5.1c and 5.1d show OMD's current policy at different learning steps, one red circle per step. Heavily red areas are areas where OMD spent a lot of learning time; very light-red areas are areas not much visited

by OMD. Each circle in Figure 5.2 represents the initial policy played by OMD via its position, and the final policy played by OMD via its color. Colors are computed as $\pi(A)\mathrm{B} + \pi(B)\mathrm{G} + \pi(C)\mathrm{R}$, where π is the final policy, and R, G and B are the primary colors.

We see that, on the dominated-action game of Figures 5.1a and 5.1b, OMD eliminates action C, which is 0.05-dominated by action A, and converges to either A or B depending on its initialization. However, we also see what happens when there exists a 0-dominated action: in the almost-dominated-action game, Figure 5.1c, C is 0-dominated by A, and we do see that once OMD has eliminated action B from its distribution of play, it finds an equilibrium where it does not eliminate C, since A and C are in that case equivalent. This empirically shows that the condition $\epsilon > 0$-dominated condition must be true for a dominated action to be systematically eliminated by OMD.

On the biased rock-paper-scissors game, Figure 5.1d, which is not a monotonic game, we see that the last iterate of OMD does not actually converge to a fixed policy, and instead cycles, yielding an approximate coarse correlated equilibrium. We note that, since its last iterate is only proven to converge in the monotonic case, this does not contradict the theory behind OMD, and instead enriches it with cases where OMD does reach a Mean-Field coarse correlated equilibrium without last-iterate convergence.

Figure 5.2 provides a lower-granularity view of OMD's behavior when varying its starting points through initial q-value change on the almost-dominated action game. We see that when $\mu(B) > 50\%$, OMD converges towards B; whereas its behavior is much more nuanced when the probability of playing B is lower than 50%: in this case, OMD converges towards a location-dependent, continuous-looking mixture between A and C.

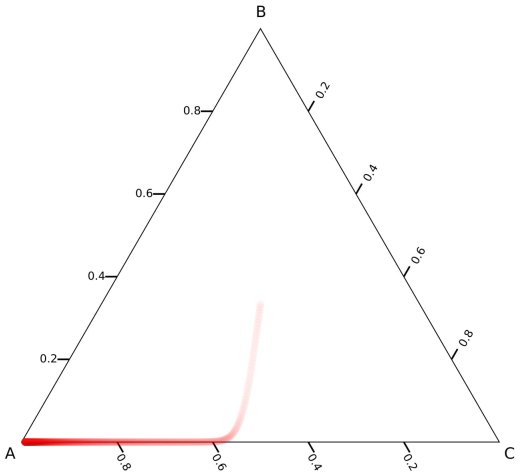
Joint Fictitious Play

Figures 5.3a and 5.3b show Joint FP on the almost-dominated action game with different initializations, Figure 5.3c shows Joint FP on the dominated-action game, Figure 5.3d shows Joint FP on the Biased RPS game, and Figure 5.4 shows Joint FP's converged-to policies (shown by a color) as a function of its initial policies (the color's position) on the almost-dominated game. Figures 5.3a, 5.3b, 5.3c and 5.3d show Joint FP's current policy at different learning steps, one red circle per step. Heavily red areas are areas where Joint FP spent a lot of learning time; very light-red areas are areas not much visited by Joint FP. Each circle in Figure 5.4 represents the initial policy played by Joint FP via its position, and the final policy played by Joint FP via its color. Colors are computed as $\pi(A)\mathrm{B} + \pi(B)\mathrm{G} + \pi(C)\mathrm{R}$, where π is the final policy, and R, G and B are the primary colors.

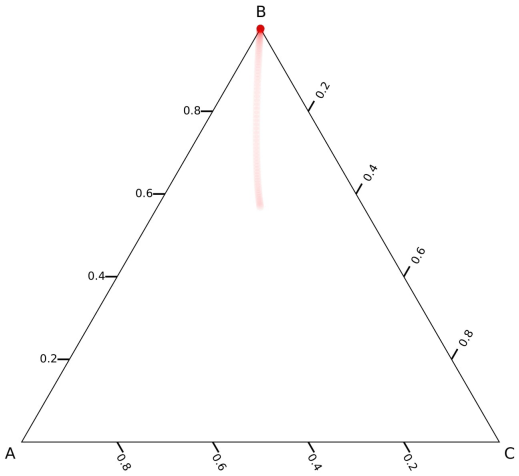
Figures 5.3a and 5.3c demonstrate that Joint Fictitious Play is much faster and harsher in eliminating dominated actions: indeed, action C is never even considered by the algorithm - it is eliminated directly. However, we note that if the algorithm had started in a region where A and C were equivalent (where $\mu_B = 0$), it would indeed have kept their proportions equal. As expected and shown in Figure 5.3b, Joint FP converges to action B when it starts close enough to it.

On Biased Rock-Paper-Scissors, Figure 5.3d, we notice that Joint FP behaves similarly as OMD: Joint FP does not manage to converge, but instead cycles around the optimal policy, yielding an approximate coarse correlated equilibrium. Very interestingly, but also unsurprisingly, JFP walks "in straight lines", because its new policies are always best responses; its decreasing speed is due to the $\frac{1}{N}$ factor in its update.

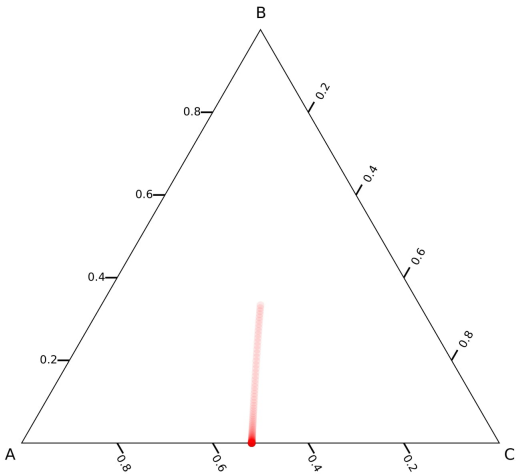
Figure 5.4 provides a lower-granularity view of JFP's behavior when varying its starting policy on the Almost-dominated action game. We see that as soon as the proportion of population playing B exceeds 50%, JFP will converge towards B, whereas, contrarily to OMD and in accordance with Proposition 54, it will completely eliminate action C and only focus on action A - since A is everywhere better than C.



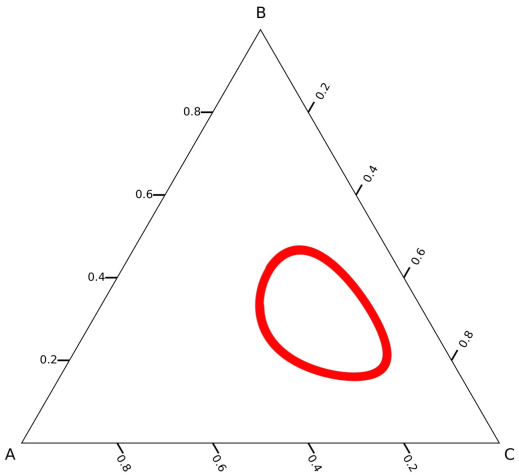
(a) Online Mirror Descent (OMD) on the Dominated Strategy Game - Center start.



(b) OMD on the Dominated Strategy Game - Biased start towards B.



(c) OMD on the Almost-Dominated Strategy Game.



(d) OMD on the Biased Rock-Paper-Scissors Game.

Figure 5.1: Online Mirror Descent (OMD) on several Normal-Form Mean-Field Games. Each red circle represents OMD's policy at a given step.

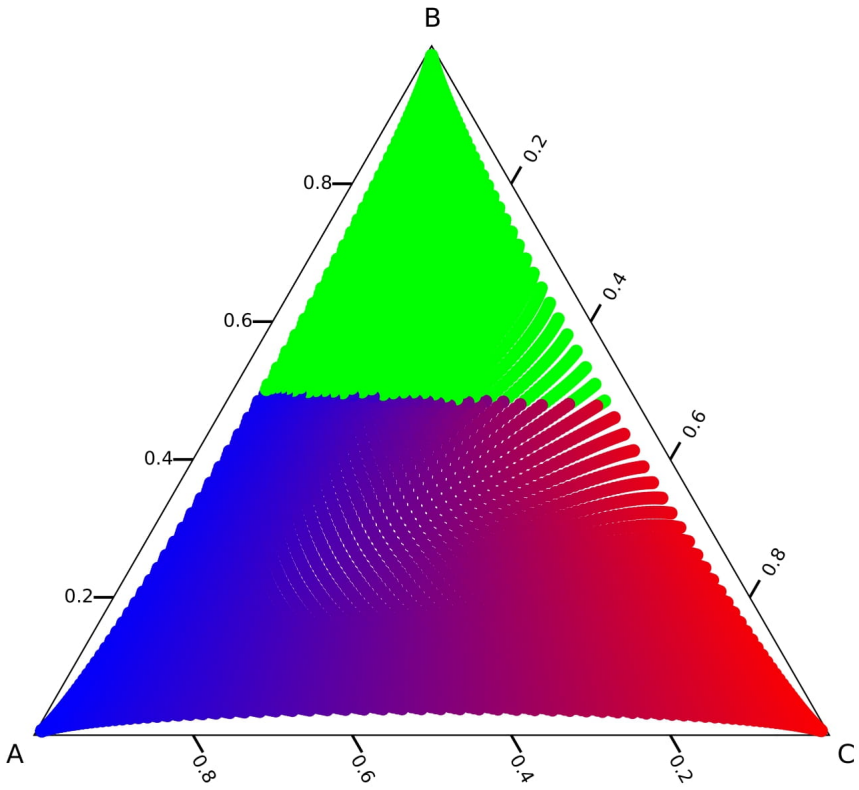
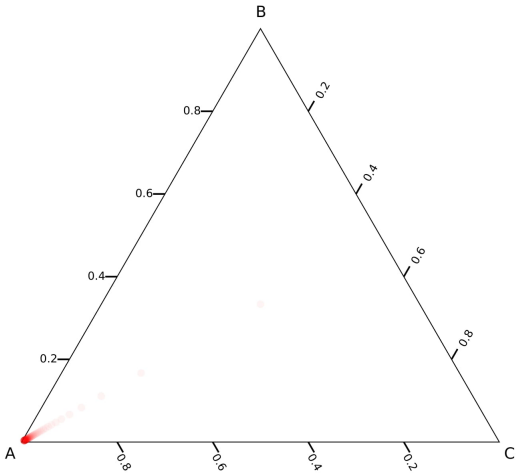
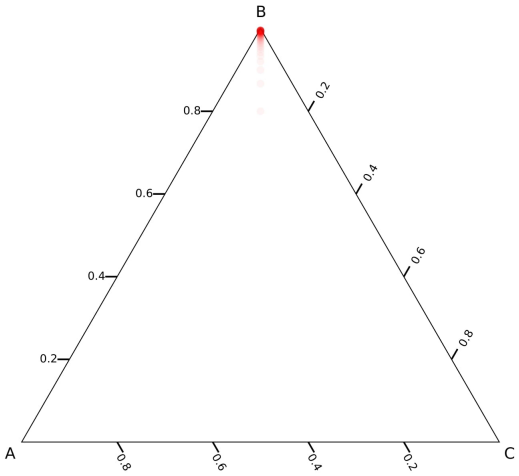


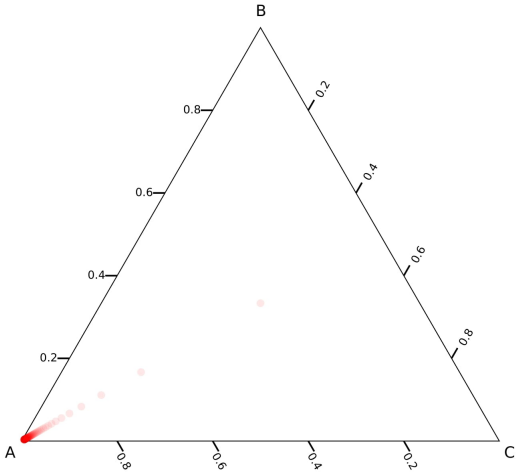
Figure 5.2: Online Mirror Descent (OMD) map of starting-points to converged-points on the Almost-dominated Strategy game. Each point represents the starting policy of OMD, and each color, its final policy. Colors are computed as $\pi(A)\mathrm{B} + \pi(B)\mathrm{G} + \pi(C)\mathrm{R}$, where π is the final policy, and R, G and B are the primary colors.



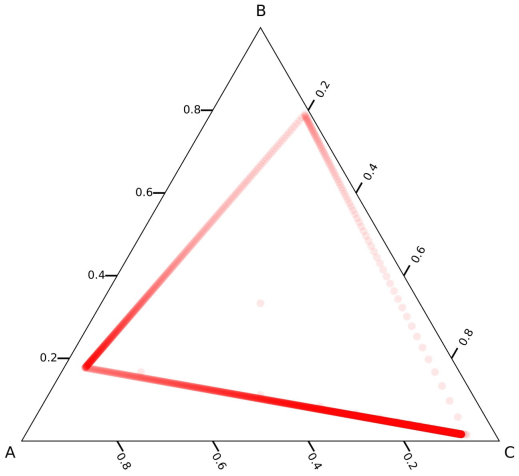
(a) Joint Fictitious Play on the Dominated Strategy Game - Center start.



(b) Joint Fictitious Play on the Dominated Strategy Game - Biased start towards B.



(c) Joint Fictitious Play on the Almost-Dominated Strategy Game.



(d) Joint Fictitious Play on the Biased Rock-Paper-Scissors Game.

Figure 5.3: Joint Fictitious Play on several Normal-Form Mean-Field Games. Each red circle represents Joint FP's policy at a given step.

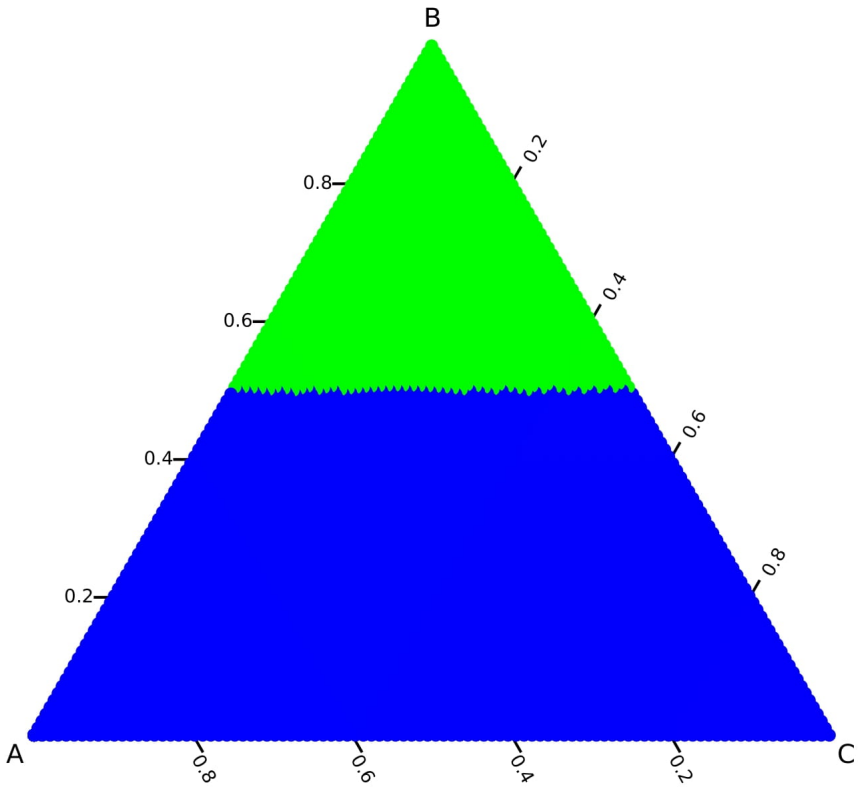


Figure 5.4: Joint Fictitious Play (JFP) map of starting-points to converged-points on the Almost-dominated Strategy game. Each point represents the starting policy of JFP, and each color, its final policy. Colors are computed following $\pi(A)\mathrm{B} + \pi(B)\mathrm{G} + \pi(C)\mathrm{R}$, where π is the final policy, and R, G and B are the primary colors.

5.3 Learning Equilibria in Mean-Field Games: Mean-Field PSRO

To define Mean-Field PSRO, we must first define the notion of restricted games and meta-games in a Mean-Field setting.

Given policies $\pi_1, ..., \pi_n \in \Pi$, we call **restricted game** the stateless game where players choose one policy among $\{\pi_i | 1 \leq i \leq n\}$ at the beginning of the game, then keep playing it until the end.

We also define **meta-games**, which are normal-form games whose payoff matrix for player 1 is, at row i and column j, $J(\pi_i, \mu^{\pi_j})$ - and the transpose thereof for player 2. The complex relationship between these notions, which are equivalent in N-player games, is explored in Section 5.3.1.

We also define the notion of Diff-Affinity, a linear-like property which is a sufficient property to use PSRO as-is in the Mean-Field setting:

Definition 41 (Diff-Affinity)**.** We say that a function $f : x, z \to f(x, z)$ is diff-affine in z, or z-diff-affine, if $\forall x, y,\ \Delta_{x,y}(f) : z \to f(x, z) - f(y, z)$ is affine.

We provide below conditions on r which make J diff-affine.

If r is of the form $r(x, a, \mu) = C(\mu) + r_1(x, a)^t \mu + r_2(x, a)$, with C any function of μ, then J is diff-affine in μ. Provided r is $\mathcal{C}^2$, this property is also necessary.

We note that our following proofs' logic can also be applied with an approximate version of diff-affinity, where

$$J(\pi', \mu(\nu)) - J(\pi, \mu(\nu)) \leq \sum_{\pi} \nu(\pi) \left(J(\pi', \mu^{\pi}) - J(\pi, \mu^{\pi})\right) + \epsilon$$

this is for example the case when $r = f+g$, with f a diff-affine function in μ, and $\forall (x, a, \mu, \mu'),\ |g(x, a, \mu') - g(x, a, \mu)| \leq \epsilon$. In this case, Mean-Field PSRO converges to ϵ variants of our equilibria.

Remark 11. *Requiring that $f : x, z \to f(x, z)$ to be such that $\forall x, y,\ \Delta_{x,y}(f) : z \to f(x, z) - f(y, z)$ is convex is equivalent to requiring that f be diff-affine.*

Proof. Let f be diff-convex. Then we know that, since f is scalar, $\Delta_{x,y}(f)$ is as well for all x, y. If f is twice-differentiable in z, so is $\Delta_{x,y}(f)$ for all values of x, y. The convexity condition on $\Delta_{x,y}(f)$ can be rewritten, if f is twice-differentiable, as

$$\begin{aligned} \forall x, y, \frac{d^2 \Delta_{x,y}(f)}{dz^2} &\geq 0 \\ \frac{d^2 f(x, z)}{dz^2} &\geq \frac{d^2 f(y, z)}{dz^2} \end{aligned}$$

Inverting x and y, we find that we have necessarily, $\forall x, y, \frac{d^2 f(x,z)}{dz^2} = \frac{d^2 f(y,z)}{dz^2} = c(z)$, therefore we know that $\forall x, z, f(x, z) = C(z) + a(x)z + b(x)$ □

A game is monotonic if and only if all its restricted games are.

Proof. Assume all restricted games are monotonic, take π_1, π_2 two policies of the game, and take the monotonic game containing only π_1, π_2. By assumption, it is monotonic, *i.e.* $\forall \nu_1,\ \nu_2 \in \Delta(\{\pi_1, \pi_2\})$,

$$J_r(\nu_1, \nu_1) + J_r(\nu_2, \nu_2) - J_r(\nu_1, \nu_2) - J_r(\nu_2, \nu_1) \leq 0$$

with $J_r(\nu, \nu') = J(\pi(\nu), \mu(\nu'))$. It suffices to take $\nu_1 = \delta_{\pi_1}$ and $\nu_2 = \delta_{\pi_2}$ to directly have

$$J(\pi_1, \mu^{\pi_1}) + J(\pi_2, \mu^{\pi_2}) - J(\pi_2, \mu^{\pi_1}) - J(\pi_1, \mu^{\pi_2}) \leq 0$$

and since this is true for all $\pi_1,\ \pi_2$, the game is monotonic.

Assume the game is monotonic. Take $\pi_1, ..., \pi_N$ with $N > 0$ and consider their derived restricted game. Let $\nu_1, \nu_2 \in \Delta(\{\pi_1, ..., \pi_N\})$.

$$J_r(\nu_1, \nu_1)+J_r(\nu_2, \nu_2)-J_r(\nu_2, \nu_1)-J_r(\nu_1, \nu_2) = J(\pi(\nu_1), \mu(\nu_1))+J(\pi(\nu_2), \mu(\nu_2))-J(\pi(\nu_2), \mu(\nu_1))-J(\pi(\nu_1), \mu(\nu_2))$$

given that $\forall \nu,\ \mu(\nu) = \mu^{\pi(\nu)}$. Since $\pi(\nu_1)$ and $\pi(\nu_2)$ are both policies of the true game, and the true game is monotonic,

$$J(\pi(\nu_1), \mu(\nu_1)) + J(\pi(\nu_2), \mu(\nu_2)) - J(\pi(\nu_2), \mu(\nu_1)) - J(\pi(\nu_1), \mu(\nu_2)) \leq 0$$

and thus

$$J_r(\nu_1, \nu_1) + J_r(\nu_2, \nu_2) - J_r(\nu_2, \nu_1) - J_r(\nu_1, \nu_2) \leq 0$$

which concludes the proof. □

5.3.1 Challenges in Scaling to Mean-Field Games

Our central proposal in this section is a generalisation of PSRO to the Mean-Field setting. We introduce the two distinct, abstract algorithms for the computation of either Mean-Field Nash equilibria or Mean-Field (coarse) correlated equilibria that we need to get in Algorithms 19 and 20.

Algorithm 19 Mean-Field PSRO(Nash)

Require: Optimizer $\sigma(J, \Pi, \Pi_n) = \arg\min_{\nu \in \Delta(\Pi_n)} \max_{i=1,...,n} J(\pi_i, \mu(\nu)) - J(\pi(\nu), \mu(\nu))$.
1: $\Pi_1 = \{\pi_1\}$ with π_1 any policy in Π.
2: $\nu_1(\pi_1) = 1$.
3: $n = 1$.
4: **while** $\Pi_{n+1} \setminus \Pi_n \neq \emptyset$ **do**
5: $\quad \Pi_{n+1} = \Pi_n \cup \{BR(\mu^{\pi(\nu_n)})\}$
6: $\quad n = n + 1$.
7: $\quad \nu_n = \sigma(J, \Pi, \Pi_n) = \arg\min_{\nu \in \Delta(\Pi_n)} \max_{i=1,...,n} J(\pi_i, \mu(\nu)) - J(\pi(\nu), \mu(\nu))$.
8: **end while**
9: **return** Π_n, ν_n, Nash equilibrium $\pi(\nu_n)$.

Algorithm 20 Mean-Field PSRO((C)CE)

Require: No-regret learner $\mathbb{A}$, best-response function BR.
1: $\Pi_1 = \{\pi_1\}$ with π_1 any policy in Π.
2: $\rho(\delta_{\pi_1}) = 1.0$.
3: $n = 1$.
4: **while** $\Pi_{n+1} \setminus \Pi_n \neq \emptyset$ **do**
5: $\quad \Pi_{n+1} = \Pi_n \cup BR(\Pi_n, \rho)$.
6: $\quad n = n + 1$.
7: $\quad \rho_n = \mathbb{A}(\Pi_n)$.
8: **end while**
9: **return** Π_n, ϵ-Mean-Field (coarse) correlated equilibrium $\rho^* \in \Delta(\Delta(\Pi^*))$

Note that for Algorithm 20 to compute the correct equilibria, $\mathbb{A}$ and BR must satisfy certain computational requirements:

CE To compute correlated equilibria, we must have

1. $BR(\Pi_n, \rho) = \{BR_{CE}(\pi_i, \rho_n) \mid \pi_i,\ \rho_n(\pi_i) > 0\}$, and
2. $\mathbb{A}(\Pi_n) = \arg\min_{\rho \in \Delta(\Delta(\Pi_n))} \mathbb{E}_{\nu \sim \rho, \pi \sim \nu}[\max_{i=1..n} J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$.

CCE To compute coarse-correlated equilibria, we must have

1. $\Pi_{n+1} = \Pi_n \cup BR_{CCE}(\rho_n)$, and
2. $\mathbb{A}(\Pi_n) = \underset{\rho \in \Delta(\Delta(\Pi_n))}{\arg\min} \underset{i=1,...,n}{\max} \mathbb{E}_{\nu \sim \rho, \pi \sim \nu}[J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$.

where BR_{CCE} and BR_{CE} are defined below.

These two algorithms have a very similar structure to the PSRO as described for N-player games in Section 2.3.4; within the inner loop, a distribution is computed for the restricted game under consideration (either a Nash equilibrium, or a (coarse) correlated equilibrium), and new policies are derived as certain types of best response against the computed equilibrium. Keeping the same insight as [112], we define two different Best Responder functions BR_{CE} and BR_{CCE}, for use with MF-PSRO in computing CEs and CCEs, respectively:

- $BR_{CCE}(\rho) := \underset{\pi^* \in \Pi}{\arg\max} \sum_\nu \rho(\nu) J(\pi^*, \mu(\nu))$;
- $BR_{CE}(\pi_k, \rho) := \underset{\pi^* \in \Pi}{\arg\max} \sum_\nu \rho(\nu|\pi_k) J(\pi^*, \mu(\nu))$.

We note that $BR_{CCE}(\rho)$ is the Best Response corresponding to a unilateral deviation from ρ, *i.e.* deviating before having been given a recommendation, whereas $BR_{CE}(\pi_k, \rho)$ is the best response generated by deviating from recommendation π_k.

Given these proto-algorithms, several important questions are immediately raised. First, are these algorithms guaranteed to return instances of the equilibria they seek to find? This is a purely mathematical question. Second, how should the restricted game equilibria in the inner loop be computed? As described in Section 2.3.4, the restricted game in usual applications of PSRO satisfies a 'linearity of evaluation' property: the payoff obtained when playing against a mixture of deterministic policies is equal to the mixture of the payoffs against each deterministic policy : $J(\pi', \sum_i \sigma_i \pi_i) = \sum_i \sigma_i J(\pi', \pi_i)$, where $\pi_i, \pi' \in \Pi$, $\sigma \in \Delta(\Pi)$.

Unfortunately, this linearity property is lost in the case of Mean-Field games, in which the representative player's payoff is generally non-linear as a function of the population occupancy measure. Indeed, even when assuming that Mean-Field dynamics are independent of the Mean-Field distribution (which is a restrictive assumption), despite the fact that distributions are linear in $\nu \in \Delta(\Pi)$, *i.e.* $\mu^{\pi(\nu)} = \sum_i \nu(\pi_i) \mu^{\pi_i}$, we typically do not have that $J(\pi', \sum_i \nu_i \mu^{\pi_i}) = \sum_i \nu_i J(\pi', \mu^{\pi_i})$, nor the property that is needed to use a payoff table's equilibrium, $J(\pi', \sum_i \nu_i \mu^{\pi_i}) - J(\pi, \sum_i \nu_i \mu^{\pi_i}) \leq \sum_i \nu_i \left(J(\pi', \mu^{\pi_i}) - J(\pi, \mu^{\pi_i})\right)$, as we show in Section 5.3.5. This lack of linearity presents a serious barrier in directly applying PSRO to Mean-Field games, and an important contribution of this section is how to circumvent this barrier: indeed, it is now typically impossible to correctly use payoff tables ! We however note that, for a limited class of Mean-Field games, linearity is preserved; we describe the details of this case in Section 5.3.5.

The next two sections treat the theoretical and implementation questions raised above for Nash equilibria, and for (coarse) correlated equilibria, in turn.

5.3.2 Convergence to Nash Equilibria

Existence and Computation of Restricted Game Equilibria

In the inner loop of MF-PSRO(Nash), an important subroutine is the computation of a Mean-Field Nash equilibrium for the restricted game; namely, a distribution $\nu \in \Delta(\Pi_n)$ such that

$$J(\pi', \mu(\nu)) - J(\pi(\nu), \mu(\nu)) \leq 0\,, \ \ \forall \pi' \in \{\pi_1, ..., \pi_n\}.$$

We note that if at least one such ν exists, then the following optimization problem in the inner loop of MF-PSRO(Nash), which minimizes exploitability, will return a Nash equilibrium

$$\nu^* = \underset{\nu \in \Delta_n}{\arg\min} \underset{i=1...n}{\max} J(\pi_i, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\,. \tag{5.8}$$

Fortunately, the conditions of existence for a Nash equilibrium of the restricted game - so called restricted Nash equilibrium - only require continuity of r with respect to μ, as shown in the following theorem.

Theorem 58 (Existence of restricted Nash equilibria)**.** *If the reward function of the game is continuous with respect to μ, then there always exists a restricted game Nash equilibrium.*

Proof. Let $\phi : \Delta(\Pi_n) \to 2^{\Delta(\Pi_n)}$ be the best-response map in the restricted game characterized by policies in the set Π_n:

$$\forall \nu \in \Delta(\Pi_n), \quad \phi(\nu) := \underset{\nu' \in \Delta(\Pi_n)}{\arg\max}\, J(\pi(\nu'), \mu(\nu)).$$

$\Delta(\Pi_n)$ is non-empty and convex, together with closed and bounded in a finite-dimensional space, and therefore compact.

For all $\nu \in \Delta(\Pi_n)$, $\underset{\nu' \in \Delta(\Pi_n)}{\arg\max}\, J(\pi(\nu'), \mu(\nu)) \subseteq \Delta(\Pi_n)$ because $\Delta(\Pi_n)$ is closed, and $\phi(\nu)$ is therefore non-empty.

Let $\nu_1,\ \nu_2 \in \phi(\nu)$, $t \in [0, 1]$.

$$J(\pi(t\nu_1 + (1-t)\nu_2), \mu(\nu)) = tJ(\pi(\nu_1), \mu(\nu)) + (1-t)J(\pi(\nu_2), \mu(\nu))$$

so $t\nu_1 + (1-t)\nu_2 \in \phi(\nu)$ and $\phi(\nu)$ is therefore convex.

$\text{Graph}(\phi) = \{(\nu, \nu') \in \Delta(\Pi_n) \times \Delta(\Pi_n) \mid \nu' \in \phi(\nu)\}$. Let $(\nu_k^1, \nu_k^2)_k$ be a sequence of elements of $\text{Graph}(\phi)$ which converges towards $(\nu_*^1, \nu_*^2) \in \Delta(\Pi_n) \times \Delta(\Pi_n)$.

r is continuous in μ, therefore J is also continuous in μ. Since $J : (\nu_1, \nu_2) \to J(\pi(\nu_1), \mu(\nu_2))$ is linear in ν_1 because $J\left(\pi(\nu_1), \mu(\nu_2)\right) = \sum_i \nu_1^i J\left(\pi_i, \mu(\nu_2)\right)$, it is also bicontinuous.

Since J is bicontinuous, let $\epsilon > 0$ and $\alpha > 0$ be such that $\forall (\nu_1, \nu_2) \in \Delta(\Pi_n) \times \Delta(\Pi_n)$ such that $d\left((\nu_1, \nu_2), (\nu_*^1, \nu_*^2)\right) \leq \alpha$,

$$\left|J\left(\pi(\nu_1), \mu(\nu_2)\right) - J(\pi(\nu_*^1), \mu(\nu_*^2))\right| \leq \epsilon$$

with d a metric over $\Delta(\Pi_n) \times \Delta(\Pi_n)$ under which J is continuous. Let $N_0 > 0$ be such that $\forall n \geq N_0,\ d\left((\nu_k^1, \nu_k^2), (\nu_*^1, \nu_*^2)\right) \leq \alpha$, and let $n \geq N_0$.

By bicontinuity and triangle inequality,

$$J\big(\pi(\nu), \mu(\nu_*^2)\big) \leq \epsilon + J\big(\pi(\nu), \mu(\nu_n^2)\big)$$

$$-J\big(\pi(\nu_*^1), \mu(\nu_*^2)\big) \leq \epsilon - J\big(\pi(\nu_n^1), \mu(\nu_n^2)\big),$$

and by optimality of ν_n^1 against $\mu(\nu_n^2)$, $\forall \nu \in \Delta(\Pi_n)$,

$$J\big(\pi(\nu), \mu(\nu_n^2)\big) - J\big(\pi(\nu_n^1), \mu(\nu_n^2)\big) \leq 0.$$

We then have, $\forall \nu \in \Delta(\Pi_n)$,

$$\begin{aligned} J(\pi(\nu), \mu(\nu_*^2)) - J(\pi(\nu_*^1), \mu(\nu_*^2)) &\leq 2\epsilon + J(\pi(\nu), \mu(\nu_n^2)) - J(\pi(\nu_n^1), \mu(\nu_n^2)) \\ &\leq 2\epsilon \end{aligned}$$

This is true for all ν, so also for their sup:

$$\sup_{\nu} J(\pi(\nu), \mu(\nu_*^2)) - J(\pi(\nu_*^1), \mu(\nu_*^2)) \leq 2\epsilon.$$

Finally, this is true for all $\epsilon > 0$. Taking ϵ to 0, we have that $J(\pi(\nu_*^1), \mu(\nu_*^2)) = \sup_\nu J(\pi(\nu), \mu(\nu_*^2))$, and thus $(\nu_*^1, \nu_*^2) \in \text{Graph}(\phi)$. Therefore $\text{Graph}(\phi)$ is closed.

We have all the hypotheses required to apply Kakutani's fixed point theorem [91]: there thus exists $\nu^* \in \Delta(\Pi_n)$ such that $\nu^* \in \phi(\nu^*)$, *i.e.* $\nu^* = \arg\max_{\nu'} J(\pi(\nu'), \mu(\nu^*))$, which means that $\forall \nu' \in \Delta(\Pi_n), J(\pi(\nu'), \mu(\nu^*)) \leq J(\pi(\nu^*), \mu(\nu^*))$, in other words: ν^* is a Nash equilibrium of the restricted game. □

Having established the existence of Nash equilibria for the restricted Mean-Field game in the inner loop of MF-PSRO(Nash), we now turn to the problem of how such an equilibrium can be (approximately) computed. As remarked earlier, due to the non-linearity of the restricted game, this problem is a non-linear (and potentially non-convex) optimisation problem over $\Delta(\Pi_n)$. Thus, the optimal solution of Equation equation 5.8 can be, in the absence of any additional assumptions on the game, found via Black-Box optimization approaches, such as random search [168], Bayesian optimization [64], evolutionary search (our experiments use CMA-ES [78]), or any other appropriate method for the considered game.

Convergence to Nash

The termination condition of PSRO is the following: if at step $N+1$, the new policy π_{n+1} produced by the algorithm is in Π_n, then the algorithm terminates. Given that each π_i is a deterministic policy, and that the set of deterministic policies is finite, PSRO will therefore necessarily terminate. We must only prove one thing:

Proposition 59 (Termination-optimality)**.** *If MF-PSRO(Nash) terminates, it stops at a Nash equilibrium of the true game.*

Proof. If MF-PSRO(Nash) terminates at step n, then $\pi^* = \arg\max_{\pi \in \Pi} J(\pi, \mu(\nu))$ is a member of Π_n.

Since ν is a Nash equilibrium of the restricted game by assumption, then necessarily $J(\pi^*, \mu(\nu)) \leq J(\pi(\nu), \mu(\nu))$, and thus $\forall \pi \in \Pi, J(\pi, \mu(\nu)) \leq J(\pi(\nu), \mu(\nu))$, which concludes the proof. □

Using the former discussion and this property, we deduce

Theorem 60 (Mean-Field PSRO convergence to Nash equilibria)**.** *Mean-Field PSRO(Nash) converges to a Nash equilibrium of the true game.*

5.3.3 Convergence to (Coarse) Correlated Equilibria

We now turn our attention to the versions of MF-PSRO that aim to compute Mean-Field correlated equilibria and Mean-Field coarse correlated equilibria.

Overview

Computing restricted MF(C)CEs is potentially more involved than computing restricted Mean-Field Nash equilibria; while the optimisation problem defining restricted Nash equilibria is over the finite-dimensional space $\Delta(\Pi_n)$, the optimisation problem defining restricted MF(C)CEs is over the infinite-dimensional space $\Delta(\Delta(\Pi_n))$. One could resort to computing an approximate Mean-Field Nash equilibrium (a special case of both Mean-Field coarse-correlated and correlated equilibria) using the black-box optimisation approach described in the previous section, but it is possible to exploit the structure of the Mean-Field game to compute approximate MF(C)CEs more efficiently. The approach we pursue is fundamentally based on no-regret learning; we also find opportunities to increase the quality of the approximate equilibrium by post-processing the output of the regret-minimisation algorithm via linear programming; see Figure 5.6 for an overview of the techniques at play.

Approximate (Coarse) Correlated Equilibria via Regret Minimisation

Our goal is to approximate an MF(C)CE for the restricted MFG based on the policy set $\Pi_n = \{\pi_1, \ldots, \pi_n\}$, as required within the inner loop of Algorithm 20. Recall that this amounts to solving the optimisation problem

$$\rho_n = \arg\min_{\rho \in \Delta(\Delta(\Pi_n))} \max_{i=1,\ldots,n} \mathbb{E}_{\nu \sim \rho, \pi \sim \nu}[J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$$

in the case of coarse correlated equilibria, and

$$\rho_n = \underset{\rho \in \Delta(\Delta(\Pi_n))}{\arg\min} \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}[\max_{i=1..n} J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$$

in the case of correlated equilibria. In principle, similar black-box techniques described for approximating Nash equilibria in the previous section may be applied to solve these problems too. However, such an approach is likely to be inefficient in practice, and instead we build on regret-minimisation theory, a classical approach to computing (C)CEs in N player games.

The overall approach relies on the fact that if the population distribution μ is fixed, the payoff function $\mathbb{E}_{\pi\sim\nu}[J(\pi, \mu)]$ is linear in the distribution $\nu \in \Delta(\Pi_n)$, and we are in fact considering online linear optimisation problems. Focusing first on the case of coarse correlated equilibria, we will make use of Algorithms **A** achieving $O(\sqrt{T})$ external regret in online linear optimisation, of the form described in Algorithm 21.

Algorithm 21 Generic form of regret-minimisation algorithm for online linear optimisation on the domain $\Delta(\Pi_n)$.

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Algorithm proposes a distribution $\nu_t \in \Delta(\Pi_n)$.
3: Algorithm observes a linear reward function $R_t : \Delta(\Pi_n) \to \mathbb{R}$.
4: Algorithm receives the reward $R_t(\nu_t)$.
5: **end for**
6: **return** Sequence of predictions $(\nu_t)_{t=1}^T$ such that $\max_{\nu\in\Delta(\Pi_n)} \sum_{t=1}^T R_t(\nu) - \sum_{t=1}^T R_t(\nu_t) = O(\sqrt{T})$.

We may apply such an algorithm for MF(C)CE computation as shown in Algorithm 22.

Algorithm 22 Protocol for computing an approximate MF(C)CE via regret-minimisation.

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Representative player selects distribution $\nu_t \in \Delta(\Pi_n)$ using a regret-minimisation algorithm $\mathbb{A}$ based on past loss function $(R_s)_{s=1}^{t-1}$. .
3: Representative player observes reward function $R_t(\nu) = \mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(v_t)]$.
4: Representative player receives reward $R_t(\nu_t) = \mathbb{E}_{\pi\sim\nu_t}[J(\pi, \mu(v_t)]$.
5: **end for**
6: **return** Empirical average $\rho = \frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$.

This algorithm returns the empirical average $\frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$, which is in fact an approximate MF(C)CE for the restricted game, as the following result shows.

Proposition 61. *The empirical average $\rho = \frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$ returned by Algorithm 22 using a regret-minimisation algorithm $\mathbb{A}$ of the form described in Algorithm 21, is a $O(1/\sqrt{T})$-MF(C)CE for the restricted Mean-Field game.*

Proof. This is a direct computation. The benefit of the representative player deviating to π_i under the correlation device ρ is

$$\begin{aligned}
&\mathbb{E}_{\nu\sim\rho}[J(\pi_i, \mu(\nu)) - \mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(\nu))]] \\
=&\frac{1}{T}\sum_{t=1}^T (J(\pi_i, \mu(\nu_t)) - \mathbb{E}_{\pi\sim\nu_t}[J(\pi, \mu(\nu_t))]) \\
=&\frac{1}{T} O\left(\sqrt{T}\right) = O\left(1/\sqrt{T}\right) ,
\end{aligned}$$

where the penultimate equality follows from the regret-minimising property of algorithm $\mathbb{A}$. The proof for CEs is similar. □

This result establishes a rigorous means of approximating an MF(C)CE in the restricted game considered within the inner loop of Mean-Field PSRO, and therefore provides an implementable version of Mean-Field PSRO. By strengthening the regret minimisation algorithm described above to minimise *internal* regret, we obtain a time-average strategy that is an approximate Mean-Field correlated equilibrium. In both cases, we have the following correctness guarantee for MF-PSRO.

Theorem 62 (MF-PSRO Convergence to MF(C)CEs)**.** *MF-PSRO using a no-internal-regret (Respectively no-external-regret) algorithm to compute its Mean-Field correlated equilibrium (Respectively Mean-Field coarse correlated equilibrium) with average regret threshold ϵ and Best-Response Computation BR_{CE} (Respectively BR_{CCE}) converges to an ϵ-CE (Respectively an ϵ-CCE).*

Proof. Since there are only a finite number of deterministic strategies in the game, we know that PSRO must necessarily terminate.

If PSRO terminates when using a restricted MFCCE, we must have

$$\pi^* = \arg\max_{\pi} \sum_{\nu} \rho(\nu) J(\pi, \mu(\nu)) \in \Pi_n \ .$$

By definition of ρ, $\sum_\nu \rho(\nu)\Big(J(\pi^*, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\Big) \leq \epsilon$, and therefore $\forall \pi \in \Pi$, $\sum_\nu \rho(\nu)\Big(J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\Big) \leq \epsilon$, ergo: ρ is a Mean-Field ϵ-coarse correlated equilibrium.

If PSRO terminates when using a restricted Mean-Field correlated equilibrium, then it means that $\forall \pi_k,\ \rho(\pi_k) > 0,\ \pi^*(\pi_k) = \arg\max_{\pi} \sum_\nu \rho(\nu|\pi_k) J(\pi, \mu(\nu)) \in \Pi_n$. By definition of ρ, $\forall \pi \in \Pi_n,\ \rho(\pi) \sum_\nu \rho(\nu|\pi)\Big(J(\pi^*(\pi), \mu(\nu)) - J(\pi, \mu(\nu))\Big) \leq \epsilon$, and therefore $\forall \pi' \in \Pi$, $\sum_\nu \rho(\nu)\Big(J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\Big) \leq \epsilon$, ergo: ρ is a Mean-Field ϵ-correlated equilibrium. □

As we will see in the next section, it is often possible to improve upon the uniform mixture of $(\nu_t)_{t=1}^T$ output by the regret-minimisation algorithm to obtain a more accurate approximation to an MF(C)CE.

Improving the Bandit: Speed

One could use no-regret learners directly to converge towards MF(C)CE, but their equilibrium contains T different distributions. This potentially means a very high amount of different ν_t recommended by our (C)CE, which can lead to learning difficulties on the part of best-responders (since every separate ν must be taken into account), implementation difficulties of equilibria in the real world, and inefficiencies: Indeed, changing per-timestep weights $\frac{1}{T}$ to potentially non-uniform ρ_t can lead to converging to ϵ'-MF(C)CE instead of ϵ ones, with $\epsilon' \ll \epsilon$, which is illustrated in Figure 5.5, computed at the first iteration of PSRO, in the Crowd Modelling [149] game. We define $(\rho_t)_t$ as the optimal solution of the following optimization problem:

$$\begin{aligned} &\min_{\rho} \ \max_{i} \ \rho^t \text{Regret}_i \\ &\text{s.t. } \forall t \ \ \rho_t \geq 0, \ \ \sum_t \rho_t = 1 \end{aligned} \tag{5.9}$$

with $\text{Regret}_i[t] := J(\pi_i, \mu(\nu_t)) - J(\pi(\nu_t), \mu(\nu_t))$.

We note that Problem (5.9) can be interpreted as finding the row player's Nash equilibrium distribution in a zero-sum normal-form game whose payoff matrix for player 1 is regret. We note that this objective can be expressed linearly.

A similar problem can be solved to find better restricted Mean-Field correlated equilibria. First, define

$$\text{Regret}_{i,j}(t) = \nu_t(i)\Big(J(\pi_j, \mu(\nu_t)) - J(\pi_i, \mu(\nu_t))\Big).$$

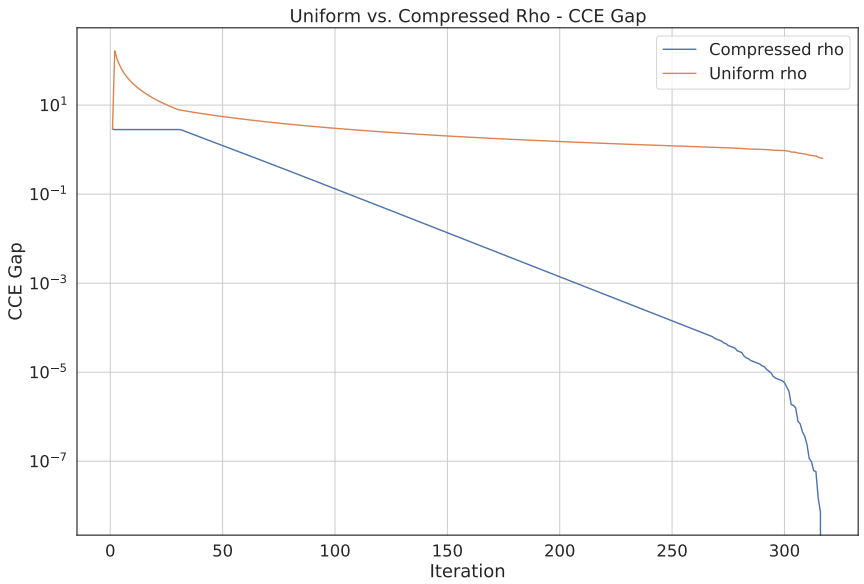


Figure 5.5: Uniform vs. Compressed ρ - CCE Gap / Time

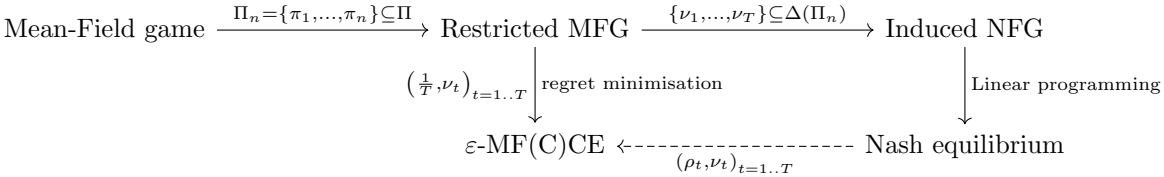


Figure 5.6: Reductions involved in approximation equilibrium computation in MF-PSRO.

The following problem gives optimal temporal weights ρ for restricted Mean-Field correlated equilibria

$$\min_{\rho} \max_{i,j} \rho^t \text{Regret}_{i,j} \tag{5.10}$$
$$\text{s.t. } \forall t \;\; \rho_t \geq 0, \;\; \sum_t \rho_t = 1.$$

This problem can similarly be expressed linearly. The following theorem confirms the optimality of ρ, the solution of Problem (5.9) or Problem (5.10):

Theorem 63 (Optimality of ρ)**.** *If $\rho = \frac{1}{T}\sum_{t=1}^{T} \delta_{\nu_t}$ is a restricted ϵ-CCE (respectively ϵ-CE), then $(\rho_t^*, \nu_t)_t$, with ρ^* the optimal solution of Problem 5.9 (respectively 5.10), yields a restricted ϵ'-MF(C)CE of the restricted game, with $\epsilon' \leq \epsilon$; and no other ρ distribution over $(\nu_t)_t$ can yield an ϵ''-MF(C)CE with $\epsilon'' < \epsilon'$.*

Proof. For restricted CCEs, the deviation incentive against the correlation device sampling ν_t with probability ρ_t in the restricted game is

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\right] = \max_i \rho^t \text{Regret}_i \,.$$

Since the uniform distribution is a possible value for ρ, we necessarily have $\max_i \rho^t \text{Regret}_i \leq \max_i \frac{1}{T}\sum_t \text{Regret}_i[t] = \epsilon$, which concludes that part of the proof.

For restricted CEs, the deviation incentive against policy π_i recommended by the correlation device sampling ν_t with probability ρ_t in the restricted game is

$$\begin{aligned}
&\max_{\pi,\pi'} \rho(\pi)\mathbb{E}_{\nu\sim\rho(\cdot|\pi)}\left[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\right] \\
&\qquad = \max_{i,j} \sum_t \rho_t \nu_t(i)\Big(J(\pi_j, \mu(\nu_t)) - J(\pi_i, \mu(\nu_t))\Big) \\
&\qquad = \max_{i,j} \rho^t \text{Regret}_{i,j}
\end{aligned}$$

Since the uniform distribution is a possible solution of Problem 5.10, we thus have that the average max deviation incentive against ρ^* the solution of Problem 5.10 is lower than or equal to that of the uniform distribution, which concludes this part of the proof.

Optimality of the solutions of problems (5.9) and (5.10) directly follows from their definitions together with the above derivations. □

Given the empirical tendency of this approach to compress temporal distribution, we name it **bandit compression**. Empirically, it allows us to find much more accurate (Figure 5.5) and sparser (Figure 5.7) distributions than uniformly averaging over $\big(\nu_t\big)_t$, and in a much lower number of steps. Yet, this algorithm is only exact in the case where the regret used by the algorithm uses exact value computations, which are impossible to obtain in large-enough games. In such games, we must typically contend with empirical, which means noisy, estimates. The next question is therefore, how sensitive is bandit compression to value-estimation noise in the regret matrix?

We provide bounds on computed average regret differences when J is perturbed by an additive random variable ϵ: $\tilde{J}(\pi, \mu) = J(\pi, \mu) + \epsilon$, giving rise to notation Regret_i^ϵ, and to the identity, if we write $\tilde{\epsilon}_t = \epsilon_t - (\nu_t)^t \epsilon_t$, $\text{Regret}_i^\epsilon = \text{Regret}_i + \tilde{\epsilon}_i$.

We write

$$\text{Regret}_* = \min_\rho \max_i \rho^t \text{Regret}_i, \;\; \text{Regret}_*^\epsilon = \min_\rho \max_i \rho^t \text{Regret}_i^\epsilon$$

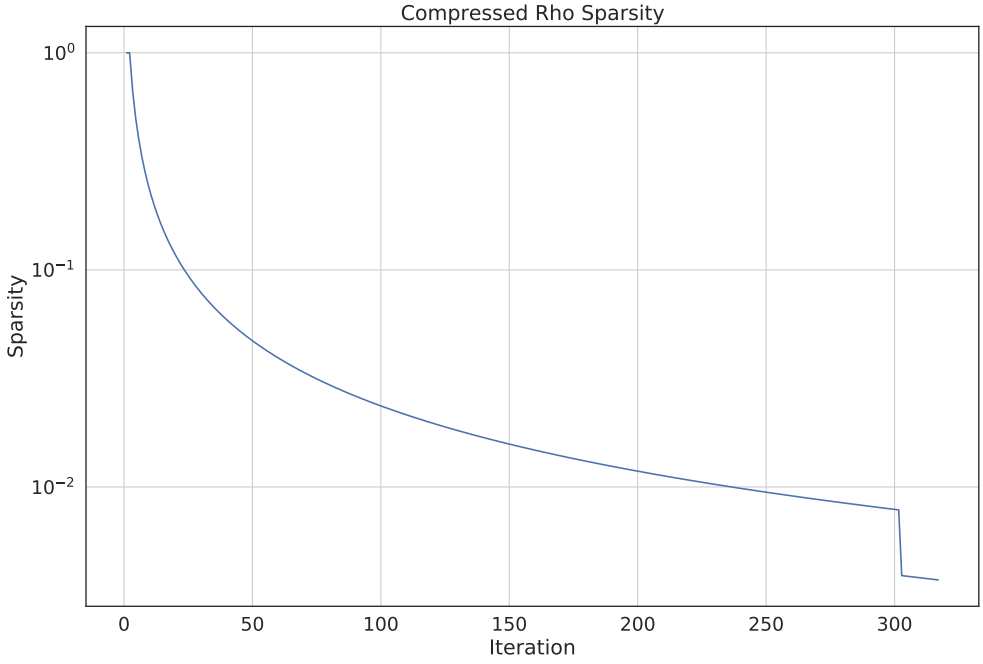


Figure 5.7: Bandit compression: sparsity / time.

We name i_* and ρ_* terms such that $\text{Regret}_* = (\rho_*)^t \text{Regret}_{i_*}$, and i^ϵ_* and ρ^ϵ_* the same values for Regret^ϵ_*.

The quantity we wish to bound is how much additional regret we experience in expectation (*i.e.* without noise) when using the noisy mixture weight ρ^ϵ_* instead of ρ_*, which we name $\Delta_O = \max_i (\rho^\epsilon_*)^t \text{Regret}_i - (\rho_*)^t \text{Regret}_{i_*}$.

Proposition 64 (Value-continuity of min-max)**.** *The optimality gap Δ_O is bounded in the following way:*

$$0 \leq \Delta_O \leq (\rho_*)^t \tilde{\epsilon}_{i^\epsilon_*} - \min_i (\rho^\epsilon_*)^t \tilde{\epsilon}_i \leq 2||\tilde{\epsilon}||_\infty \leq 4||\epsilon||_\infty \,.$$

Proof. By optimality of ρ_*, we already have that $\Delta_O \geq 0$.

$$\begin{aligned}
\Delta_O &= \max_i (\rho^\epsilon_*)^t \text{Regret}_i - (\rho_*)^t \text{Regret}_{i_*} \\
&= \max_i (\rho^\epsilon_*)^t (\text{Regret}_i + \tilde{\epsilon}_i) - (\rho^\epsilon_*)^t \tilde{\epsilon}_i - (\rho_*)^t \text{Regret}_{i_*} \\
&\leq (\rho^\epsilon_*)^t (\text{Regret}_{i^\epsilon_*} + \tilde{\epsilon}_{i^\epsilon_*}) - \min_i (\rho^\epsilon_*)^t \tilde{\epsilon}_i - (\rho_*)^t (\text{Regret}_{i^\epsilon_*} + \tilde{\epsilon}_{i^\epsilon_*}) + (\rho_*)^t \tilde{\epsilon}_{i^\epsilon_*} \\
&\leq (\rho^\epsilon_* - \rho_*)^t (\text{Regret}_{i^\epsilon_*} + \tilde{\epsilon}_{i^\epsilon_*}) + (\rho_*)^t \tilde{\epsilon}_{i^\epsilon_*} - \min_i (\rho^\epsilon_*)^t \tilde{\epsilon}_i \\
&\leq (\rho_*)^t \tilde{\epsilon}_{i^\epsilon_*} - \min_i (\rho^\epsilon_*)^t \tilde{\epsilon}_i \leq 2||\tilde{\epsilon}||_\infty
\end{aligned}$$

$\forall t,\ \tilde{\epsilon}_t = \epsilon_t - (\nu_t)^t \epsilon_t$, and $\epsilon_t \leq ||\epsilon||_\infty$ and $-(\nu_t)^t \epsilon_t \leq ||\epsilon||_\infty$, therefore $||\tilde{\epsilon}||_\infty \leq 2||\epsilon||_\infty$, which concludes the proof. □

The tightness of this bound can be verified via noting that if $\rho_* = \rho^\epsilon_*$ and the minimum of $(\rho_*)^t \epsilon_i$ is reached for $i = i^\epsilon_*$, then the optimality gap is null.

Assuming each ϵ_i is a random Gaussian variable with variance $\sigma > 0$, the term $\max_i \rho_\Delta^t \epsilon_i$ is such that $\mathbb{P}(\max_i \rho_\Delta^t \epsilon_i \leq y) = \Phi^n \left(\frac{y}{\sigma \sqrt{\rho_\Delta^t \rho_\Delta}} \right)$ where Φ is a standard Gaussian CDF, and the term on the right $\rho_\epsilon^t \epsilon_{i_\Delta} \sim \backslash(0, \frac{\sigma^2}{\rho_\epsilon^t \rho_\epsilon})$.

To get an estimation of the magnitude of this gap's distribution, we assume N to be high enough that we can ignore the term $\rho_\epsilon^t \epsilon_{i_\Delta}$ for our numerical application. Using the 5σ rule, we find that $\mathbb{P}\left(\Delta_O \leq 5\sigma \sqrt{\rho_\Delta^t \rho_\Delta}\right) \geq 0.9999994^N$. Assuming $\sigma = 0.1$, $N = 50$ and $\rho_\Delta^t \rho_\Delta = \frac{1}{50}$ (Fully uniform distribution), $\mathbb{P}\left(\Delta_O \leq 0.07\right) \geq 0.99997$. When ρ_Δ is fully focused on one point, $\rho_\Delta^t \rho_\Delta = 1$, and the former equation becomes $\mathbb{P}\left(\Delta_O \leq 0.5\right) \geq 0.99997$.

We note that this is a pessimistic estimate for several reasons

- ρ should presumably not be focused on a single point, and therefore the term $5\sigma\sqrt{\rho^t \rho}$ will be quite lower.
- This does not take into account the complex relationship and dependence between the two terms $\max_i \rho_\Delta^t \epsilon_i$ and $\rho_\epsilon^t \epsilon_{i_\Delta}$.

We add bandit compression onto Algorithm 22, accompanied with a few optimization criteria, yielding Algorithm 23. Sped-Up PSRO includes the following new features:

- ρ_{tol}: This term is a regret threshold. If the optimal solution of Problem 5.9 or 5.10 yields regret lower than this term, we consider the equilibrium search as successful.
- ρ_{lim} and new loop conditions: The PSRO loop does not terminate anymore when $\Pi_{n+1} == \Pi_n$, what it does then is *refine* its current equilibrium by halving ρ_{tol} at every iteration where $\Pi_{n+1} == \Pi_n$, until $\rho_{tol} == \rho_{lim}$, with ρ_{lim} set to a very low value, typically 10^{-12}.
- $\tau_{Compress}$: Typically set to 1, this value allows one not to optimize problems 5.9 or 5.10 at every regret minimization step. This can be especially useful when computing MFCEs, for which the problem is much slower to solve than for MFCCEs.

Remark 12 (Use of the Algorithm for Nash-Convergence)**.** *We note that one can also use Algorithm 23 for convergence towards Mean-Field Nash equilibria if one uses an iterative solver for computing the Nash equilibrium - in that case, $\mathbb{A}$ is the Nash solver, and Regret$_*$ is the exploitability. Since a Nash equilibrium only uses a single distribution, one can either bypass solving Problem 5.9, or solve it trivially with $\rho(\nu_*) = 1$.*

We discuss below the effects of Sped-up Mean-Field PSRO's parameters:

PSRO Parameter	Effect when Increased
T	Improved asymptotic convergence, lowers speed
M	Lower noise, improves convergence at fixed T, lower speed
ρ_{Tol}	Lower precision, higher speed
$\tau_{Compress}$	Higher speed if costly compression, otherwise lower

Dominated Strategy Exclusion

Just like we did for Joint FP and OMD, we examine the relationship between Mean-Field PSRO and dominated strategies. Perhaps surprisingly, we find that PSRO does not necessarily eliminate dominated strategies, at least when computing coarse-correlated equilibria. The only guarantee we find is that, when computing correlated equilibria, it always asymptotically eliminates them. To counteract this undesirable property, we propose two different alterations of the algorithm which guarantee that Mean-Field PSRO **never** recommends a dominated strategy *at any time during training.*

Proposition 65 (Mean-Field PSRO's CE-optimality)**.** *Mean-Field PSRO used to compute Mean-Field correlated equilibria can never recommend a dominated strategy at convergence.*

Algorithm 23 Sped-up Mean-Field PSRO((C)CE).

Require: ρ_{tol}, $\rho_{lim} < \rho_{tol}$, No-Regret learner $\mathbb{A}$, $\tau_{Compress}$.
$\Pi_0 = \{\}$, with π_0 any policy in Π.
$\Pi_1 = \{\pi_1\}$, with π_1 any policy in Π.
$\rho(\delta_{\pi_1}) = 1.0$.
$n = 0$
while $(\Pi_{n+1} \setminus \Pi_n) \neq \emptyset$ or $\rho_{tol} > \rho_{lim}$ **do**
  $n+ = 1$.
  $\Pi_{n+1} = \Pi_n \cup \{BR_{(C)CE}(\pi_i, \rho_n) \mid \pi_i,\ \rho(\pi_i) > 0\}$.
  if $\Pi_{n+1} == \Pi_n$ **then**
    $\rho_{tol} = \frac{\rho_{tol}}{2}$.
  end if
  Initialize $\mathbb{A}(\Pi_n)$.
  Step Count = 0.
  while Regret$_*$ ¿ ρ_{tol} **do**
    Step Count $+ = 1$.
    Do one step of $\mathbb{A}(\Pi_{n+1})$
    if Step Count $\equiv 0[\tau_{Compress}]$ **then**
      Compute ρ_* optimal solution of Problem 5.9 (CCE) / 5.10 (CE).
      Compute ρ_*'s associated regret Regret$_*$.
    end if
  end while
  $\rho_{n+1} = \rho_*$.
end while
return Empirical average $\rho = \frac{1}{T}\sum_{t=1}^{T} \delta_{\nu_t}$.

Proof. The proof results from the fact that a correlated equilibrium can, by definition, never recommend a strictly dominated strategy (if it did, then deviating to the strategy which dominates the dominated strategy would always yield payoff improvements, and therefore the correlation device in question would not be a correlated equilibrium). At convergence, Mean-Field PSRO has found a correlated equilibrium, and hence cannot recommend strictly dominated strategies. □

However, we note that PSRO could potentially recommend strictly dominated strategies when computing Mean-Field coarse correlated equilibria (which can contain dominated strategies, as shown in Section 4.2.6), or in the process of computing a Mean-Field correlated equilibrium. This is due to the initial policies present in the initialization pool of PSRO, of which we cannot guarantee optimality / non-suboptimality. It is however possible to slightly modify the algorithm to obtain an optimality-guaranteeing result:

Proposition 66 (Mean-Field PSRO: Optimality Modification)**.** *Either of the following two PSRO modifications ensures that PSRO* ***never*** *recommends strictly dominated strategies, while keeping PSRO's convergence guarantees:*

- *Ensure that all of PSRO's initial policies are not strictly dominated,*

  or

- *After PSRO's first iteration, remove all initial policies from the pool and only keep the best-responses (only PSRO's first step can then contain strictly-dominated strategies).*

Proof. Mean-Field PSRO can never add to its pool a strictly dominated strategy, since it only adds best-responses and best-responses can never be strictly dominated. Only the initial policies present in PSRO's pool could potentially be. If they are not (First modification), then PSRO's pool never contains dominated strategies, and therefore PSRO never recommends strictly dominated strategies. If we cannot be certain that they are not, we note that the best response against them

can never be strictly dominated; hence, removing them from the pool and only keeping these best-responses empties the pool from potentially strictly dominated strategies, thus preventing PSRO from recommending strictly dominated strategies □

Complexity Discussion

The use of traditional solvers, as has been the case in PSRO so far, requires filling a payoff table. At a given iteration n, this means estimating n match results for the newly added Best Response (the other match results being stored).

[Payoff matrix estimation complexity] When match payoff estimation is done via sampling match outcomes, the number of matches T necessary to reach within-ϵ estimation precision with probability α is $T = O(\frac{n}{\alpha\epsilon^2})$.

Proof. If we have T episodes to gather on $2n+1$ matches, the most natural (though not necessarily most efficient) way to distribute our compute budget is to give each match $\frac{T}{2n+1}$ episodes.

The variance of an estimated match score $\hat{J}$ is therefore $\text{Var}(\hat{J}) = \frac{\text{Var}(J)}{\frac{T}{2n+1}} = (2n+1)\frac{\text{Var}(J)}{T}$ where $\text{Var}(J)$ is the variance of the random variable representing match outcomes for J.

Using Chebyshev's inequality, we have $\mathbb{P}(|\hat{J} - J| \geq \epsilon) \leq \frac{\text{Var}(\hat{J})}{\epsilon^2} = (2n+1)\frac{\text{Var}(J)}{T\epsilon^2}$. If we aim to be within ϵ-precision of J with probability α, i.e. $\mathbb{P}(|\hat{J} - J| \geq \epsilon) = \alpha$, we need $T = O(\frac{n}{\alpha\epsilon^2})$. □

We contrast this with the complexity of using no-external- and internal-regret learners, given that one chooses an efficient algorithm:

[Bandit ϵ-Regret Complexity] The number of game matches T necessary to reach within-ϵ average regret is $T = O\left(\frac{n^3\ log(n)}{\epsilon^2}\right)$ for no-internal-regret learners, and $T = O\left(\frac{n\ log(n)}{\epsilon^2}\right)$ for no-external-regret learners. In the case of additively noisy evaluation, where samples are evaluated M times and averaged, these complexities become $T = O\left(\frac{M\ n^3\ log(n)}{\epsilon^2}\right)$ for internal-regret, and $T = O\left(\frac{n\ M\ log(n)}{\epsilon^2}\right)$ for external regret; both with probability $\delta \geq 1 - n\frac{4\sigma^2}{TM\epsilon^2}$, where σ^2 is the noise variance.

Proof. The Hedge Algorithm [21] adapted for the partial-information setting [22] has average regret bound $\epsilon = O\left(\sqrt{\frac{n\ log(n)}{T}}\right)$, therefore $T = O\left(\frac{n\ log(n)}{\epsilon^2}\right)$ when returns are exact.

Optimal Swap-regret minimizers can be derived from optimal external-regret minimizers by running N instances of them in parallel, as shown in [21], therefore $\epsilon = O\left(n\sqrt{\frac{n\ log(n)}{T}}\right)$ and $T = O\left(\frac{n^3\ log(n)}{\epsilon^2}\right)$.

In the case where payoffs are additively noisy, regret can be decomposed into two terms: $\text{Regret}_i = R_i + \tilde{R}_i$, where R is the true, noiseless regret, and $\tilde{R}_i$ is the noise-derived regret. Given $\alpha > 0$, after $O\left(\frac{M\ n^3\ log(n)}{\alpha^2}\right)$ steps, we know that Regret $\leq \frac{\alpha}{2}$. We have that True Regret$-$Regret $= \max_j R_j - (\max_i R_i + \tilde{R}_i) \leq \max_i -\tilde{R}_i$, and $\mathbb{P}(\max_i -\tilde{R}_i \geq \alpha) \leq \sum_i \mathbb{P}(-\tilde{R}_i \geq \alpha)$. Given that $\tilde{R}_i = \frac{1}{T}\sum_t \epsilon_i[t] - \nu_t^t\epsilon[t]$, and ϵ_t is averaged over M samples and thus has $\frac{\sigma^2}{M}$ variance, then $\text{Var}(-\tilde{R}_i) \leq \frac{\sigma^2}{TM}$ and Chebyshev's inequality yields $\mathbb{P}(-\tilde{R}_i \geq \frac{\alpha}{2}) \leq \frac{4\sigma^2}{TM\alpha^2}$, thus yielding $\mathbb{P}(\max_i -\tilde{R}_i \geq \alpha) \leq n\frac{4\sigma^2}{TM\alpha^2}$. We then have $\mathbb{P}(\text{True Regret} \leq \alpha) \geq \mathbb{P}(\text{Regret} + \max_i -\tilde{R}_i \leq \alpha) \leq \mathbb{P}(\max_i -\tilde{R}_i \leq \frac{\alpha}{2}) \leq 1 - n\frac{4\sigma^2}{TM\alpha^2}$ The probability of the true regret being lower than α after $O\left(\frac{M\ n^3\ log(n)}{\alpha^2}\right)$ steps is therefore $\delta \geq 1 - n\frac{\sigma^2}{TM\alpha^2}$. Since each regret steps is now composed of M times as many rollouts, the total rollout-complexity of the algorithm must be multiplied by M, which concludes the proof. □

We provide below a commentary of these results:

- Minimizing swap regret directly has higher complexity than payoff matrix estimation by a factor $n^2 log(n)$ in worst cases.

- Minimizing external regret directly has higher complexity than payoff matrix estimation by a factor $log(n)$ in worst cases.
- Using regret minimizers directly provides the user with a useable distribution over policies.
- Estimating the payoff matrix means the user still has to run an algorithm over said payoff matrix, which could have large complexity (Linear solvers have complexity $O(n^{2+\gamma})$ with $\gamma > 0$, for example).
- The relationship between payoff uncertainty and solver output uncertainty is difficult to analyze in general, due to the strong non-linearities of solvers. Indeed, picture the following 0-sum game: three strategies face off, π_1 has payoff 100 against π_2 and π_3, and π_2 has payoff 1 against π_3. Any reasonable ϵ-error in estimating the payoff obtained by playing π_1 would not change e.g. its Nash distribution, or the distribution of a correlated equilibrium. In contrary, in a game where all average payoffs are very close to 0 (π_1 barely beats π_2 and π_3, and π_2 barely beats π_3), an ϵ-error could lead to a reversal of these interactions (In the noisy payoff matrix, it could be that π_2 beats π_3 which beats π_1), thus completely changing the computed distribution.
- We do not yet fully understand the complexity reduction granted by Bandit Compression, which could greatly lower asymptotic complexity of the Bandit approach.
- This complexity insight can be transferred to the N-player case. In this case, one needs to compute $(n+1)^N - n^N = O(n^{N-1})$ matches, and the estimation complexity is therefore $T = O\left(\frac{n^{N-1}}{\alpha\epsilon^2}\right)$. The number of different actions is n^N, so the complexity of minimizing internal regret is $O\left(\frac{Nn^{3N}log(n)}{\epsilon^2}\right)$ and it is $O\left(\frac{Nn^{N}log(n)}{\epsilon^2}\right)$ for external regret minimization.
- If we can observe all policies' payoffs at no additional cost, the regret bounds become $O\left(\frac{n\ log(n)}{\epsilon^2}\right)$ for internal regret, and $O\left(\frac{log(n)}{\epsilon^2}\right)$ for external regret. The complexity for payoff matrix estimation nevertheless remains $O(\frac{n+1}{\alpha\epsilon^2})$, as one is interested in $(J(\pi_n, \mu^{\pi_k}))_k$ and $(J(\pi_k, \mu^{\pi_n}))_k$.

On the Complexity of Computing Maximum Welfare Equilibria

We have so far introduced a method that learns Nash, correlated and coarse correlated equilibria in Mean-Field games. A subsequent question for correlated and coarse correlated equilibria is, could we influence the learning process for it to find high-welfare equilibria instead of low-welfare ones?

This problem of high-welfare convergence was shown by [15] to be NP-hard in general, with the notable exception of succinct aggregate games, for which the existence of polynomial algorithms converging to high-welfare equilibria is proven. However, their method relies on a discretization of and grid-search over the aggregate space, the space of statistics summarizing the behavior of other players.

At the n-th step of Mean-Field PSRO, discretizing the n-dimensional probability vector space with step size $\frac{1}{M}$ amounts to considering matrices of size $\geq n\left(\frac{M}{2}\right)^n$, a complexity exponential in the number of iterations, therefore prohibitive.

We therefore leave open the question of high-welfare convergence for now.

5.3.4 Evaluation

To demonstrate the viability of our approach, we use three different metrics presented in Section 5.3.4, which we evaluate when running MF-PSRO on four different Mean-Field games, which are described in Section 5.3.4. Evaluation methods are detailed in Section 5.3.4, and evaluation results are discussed in Section 5.3.4.

Metrics

For a given correlation device ρ, we define

$$\text{CCEGap}(\rho) := \max_{\pi} \sum_{\nu} \rho(\nu)\big(J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\big)$$

By construction, we directly have that $\text{CCEGap}(\rho) = 0$ is equivalent to ρ being an MFCCE. In the same fashion, we define

$$\text{CEGap}(\rho) := \max_{\pi'} \max_{\pi|\rho(\pi)>0} \sum_{\nu} \rho(\nu|\pi)\big(J(\pi', \mu(\nu)) - J(\pi(\nu), \mu(\nu))\big)$$

for MFCE characterisation. Finally, for a given population distribution $\nu \in \Delta(\Pi)$, we introduce

$$\text{Exploitability}(\nu) := \max_{\pi} J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))$$

so that $\text{CCEGap}(\rho) = 0$, which reaches 0 if and only if ν is a Mean-Field Nash equilibrium.

Games

The four games we use to evaluate convergence include two complex games available in OpenSpiel [101], Predator-Prey [147] and Crowd Modeling [149], and two new small normal-form Mean-Field games, *Coop / Betray / Punish* and *Mean-Field biased Rock-Paper-Scissors*. Both games have 3 actions, A, B and C, whose rewards depend on the action distribution of the population.

Mean-Field biased Rock-Paper-Scissors is a classic biased Rock-Paper-Scissors game, where one gets as reward for playing rock the proportion of players playing scissors minus that playing paper, all distributions multiplied by different coefficients. Its payoff function is

$$\begin{aligned} r(A, \mu) &= 0.5 * \mu(B) - 0.3 * \mu(C) \\ r(B, \mu) &= 0.3 * \mu(C) - 0.7 * \mu(A) \\ r(C, \mu) &= 0.7 * \mu(A) - 0.5 * \mu(B) \end{aligned}$$

This game is meant to be a trap for Online-Mirror Descent and Fictitious Play methods, making their last iterate adopt a cyclic behavior, just like they do in N-player games.

Coop / Betray / Punish is a 3-action normal-form game where agents can choose to either Cooperate, and all get a good reward; betray and take advantage of others; or punish the betrayers. But punishing agents also take some reward away from cooperators (they must support the punishers). Payoffs are non-linear (quadratic) in distributions. Its payoff function is

$$\begin{aligned} r(A, \mu) &= \mu(A) - \frac{20}{9}(\mu(A) - \mu(C)) * \mu(C) - 2\mu(B) \\ r(B, \mu) &= 2(\mu(A) - \mu(B)) - 238\mu(C) \\ r(C, \mu) &= \frac{200}{9}(\mu(A) - \mu(C)) * \mu(C) \end{aligned}$$

This game is meant to showcase the optimality of PSRO in games with non-linear payoffs, with a flavour of prisoner's dilemma and resource attribution.

Methods

The regret minimizer used by Mean-Field PSRO((C)CE) is regret matching [174], and the Black-Box Optimization method used by Mean-Field PSRO(Nash) is CMA-ES [78]. As per Remark 12,

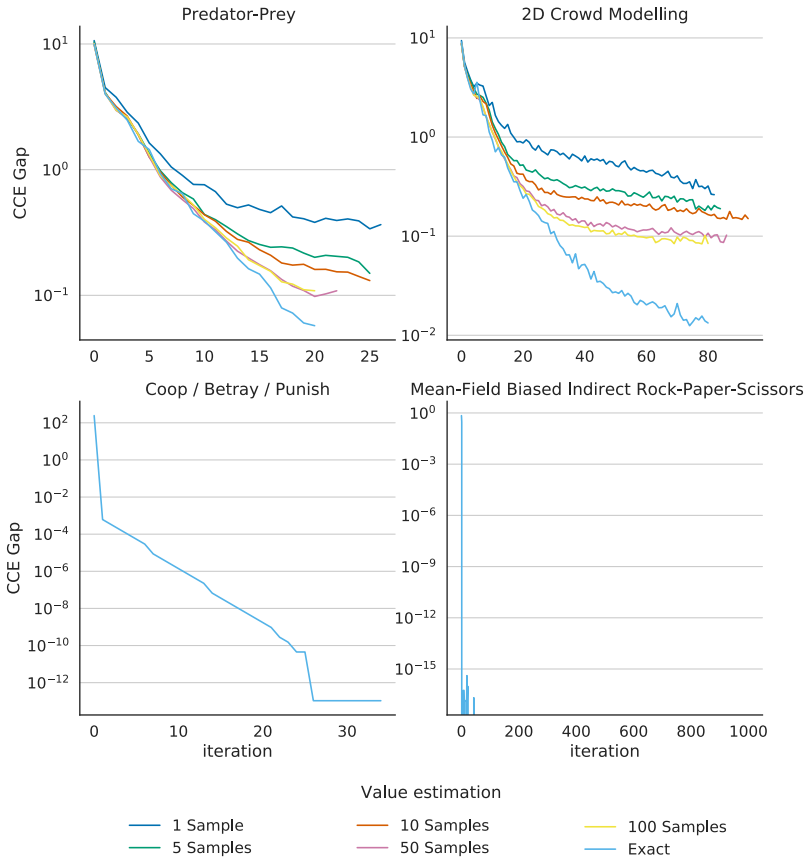


Figure 5.8: CCE Gap of Mean-Field PSRO(CCE). The two curves at the bottom only show Exact value estimation because the games they characterize are Normal-Form games, hence return exact values directly.

we use Algorithm 23 for both Mean-Field PSRO((C)CE) and Mean-Field PSRO(Nash), since the Nash solver CMA-ES is iterative.

Regarding convergence to MFCE, since there exists, to the best of our knowledge, no other algorithm known to converge towards these weaker equilibria, we only investigate the convergence behavior of Mean-Field PSRO(CE) with additional payoff noise, with no other baseline.

Regarding convergence towards Mean-Field Nash equilibria, we compare Mean-Field PSRO to OMD with several different learning rates, and Fictitious Play, both algorithms available on OpenSpiel.

Results

Figure 5.11 shows exploitability of MF-PSRO(Nash), OMD for several learning rates, and Fictitious Play.

Figure 5.8 presents the CCE-Gap of Mean-Field PSRO(CCE), 5.9, the CE-Gap of Mean-Field PSRO(CE), while Figure 5.10 exposes the Exploitability of Mean-Field PSRO(Nash) on the four Mean-Field game environments described above. We note that in both normal-form games, Mean-Field PSRO converges within numerical precision towards Mean-Field correlated, coarse correlated and Nash equilibria after only a few iterations.

Nash-wise, OMD seems capable of following PSRO at a similar speed on *Coop / Betray / Punish*, but fails utterly to converge on *Mean-Field biased Rock-Paper-Scissors*. We note that OMD's convergence is strongly affected by its learning rate. Fictitious play does not manage to find good equilibria in these games.

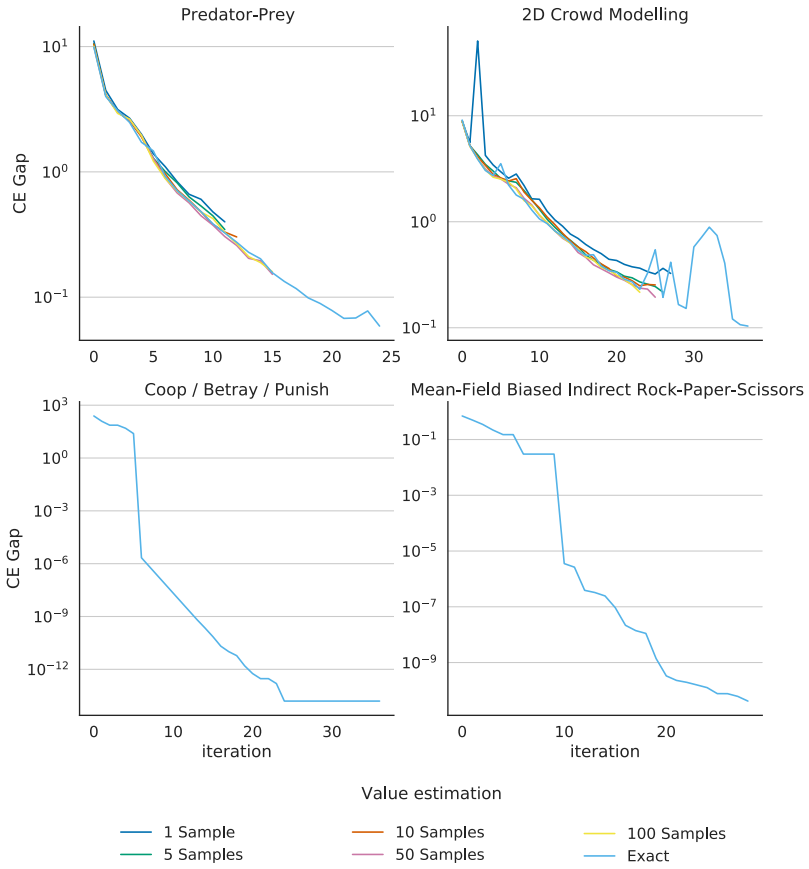


Figure 5.9: CE Gap of Mean-Field PSRO(CE). The two curves at the bottom only show Exact value estimation because the games they characterize are Normal-Form games, hence return exact values directly.

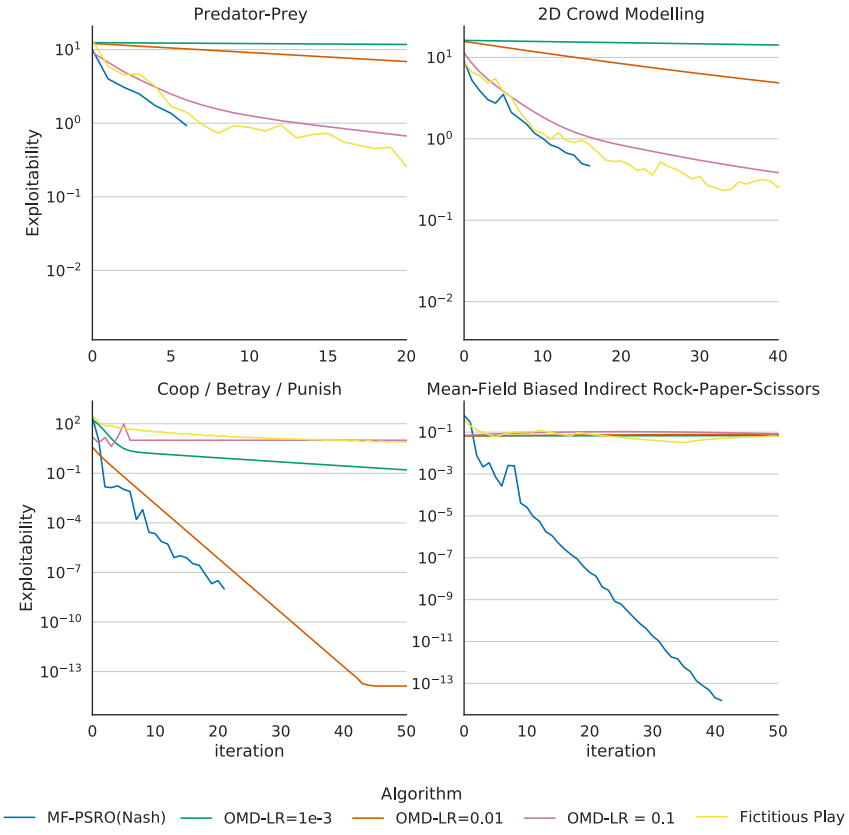


Figure 5.10: Exploitability of Mean-Field PSRO(Nash).

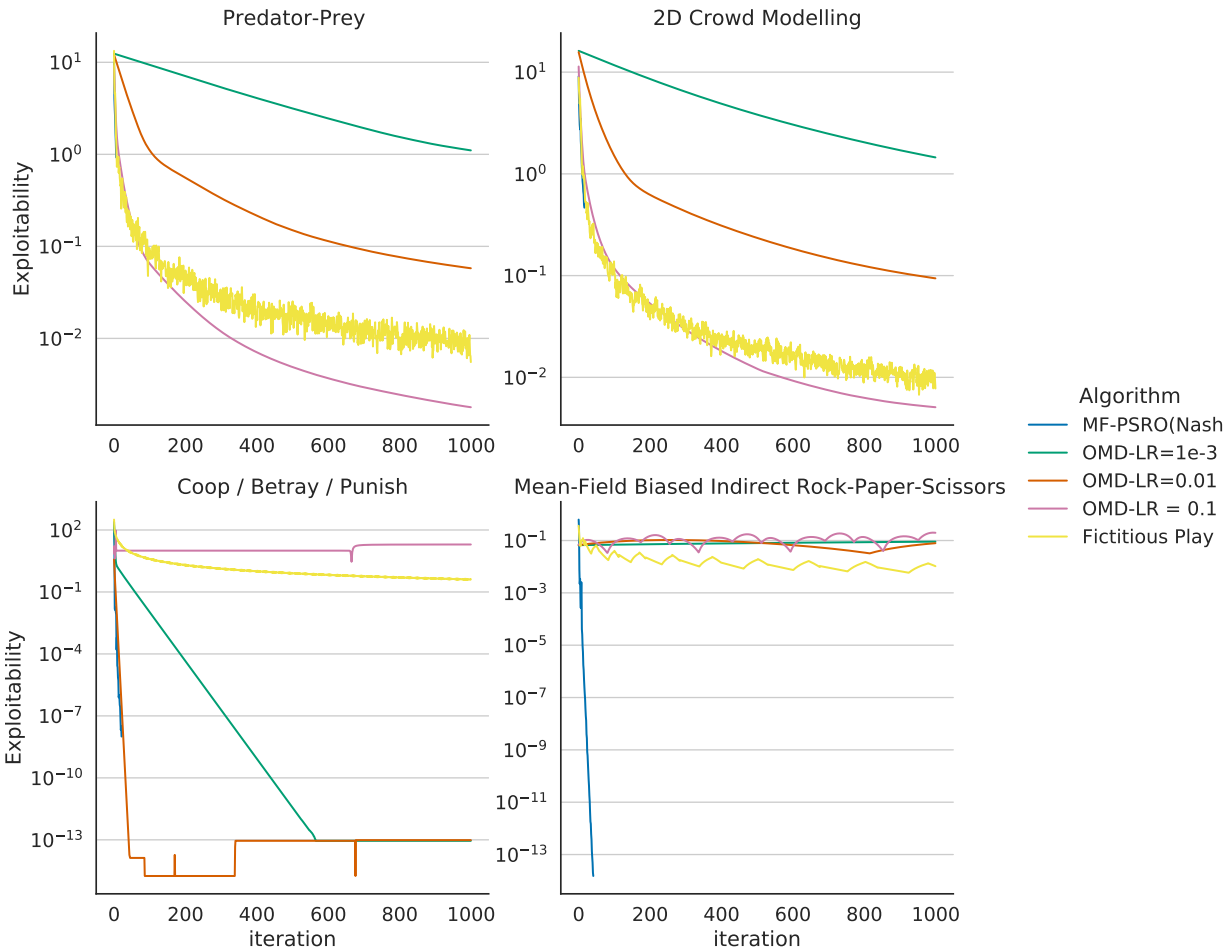


Figure 5.11: Algorithm Exploitabilities

On more complex games, Mean-Field PSRO quickly converges towards very good correlated (CCE Gap $\approx 10^{-1}$), coarse correlated equilibria (CE Gap $\approx 10^{-1}$), and Mean-Field PSRO(Nash) seems to quickly minimize exploitability - but it does much more slowly (time-wise) than both OMD and FP. This hints at a strong potential direction of improvement for Mean-Field PSRO. We note that in this zoomed-in plot, FP seems to outperform OMD. We provide a zoomed-out version in Figure 5.11 where we see that OMD, with the correct learning rate, outperforms Fictitious Play as expected.

We also show on Figure 5.12 several qualitative results regarding PSRO's behavior on the games identified in Section 5.2.4.

The first row shows the equilibria found by Mean-Field PSRO(CCE). We represent each policy played by the equilibrium, and change the policy's color from black to red the more present it is in the mixture. We notice that Mean-Field PSRO removes the dominated action in the dominated action game, and yields an interesting equilibrium for biased Rock-Paper-Scissors.

The second row shows the same results as the first for PSRO(CE). Here, we see exactly the same behavior as PSRO(CCE) for both the dominated and the almost-dominated action games; however, equilibrium shape changes drastically for the biased RPS game: instead of recommending three almost pure strategies, as did PSRO(CCE) - deviations wouldn't be able to tell which strategy is being recommended, so this is a sensible CCE -, PSRO(CE) is forced to recommend strategies closer to optimality (though not necessarily optimal) so as to reach an actual CE.

On the third and fourth row, we represent the trajectories that respectively the polynomial weights algorithm and the regret matching algorithm, both no-adversarial-regret algorithms, produce when starting with all three actions on different games. We note how much faster regret matching is at finding equilibria, a property that has already been empirically shown in N-player games in [174]. We note that these trajectories were generated without bandit compression, a speedup algorithm introduced in [126]. We notice that despite speed and trajectory differences, regret matching and the polynomial weights algorithm yield similar results on the first two (almost-)dominated-action games, whereas their behavior fundamentally differs on biased RPS.

5.3.5 The Linear Special Case

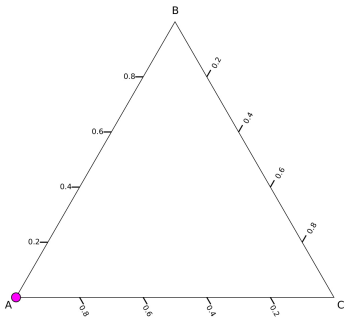
Normal-form games equilibria and links to Mean-Field restricted games

This section presents results linking restricted games with normal-form representations under the μ-diff-affinity condition. We name $\Pi_n = \{\pi_1, ..., \pi_n\}$ the set of policies used by the restricted game in the following.

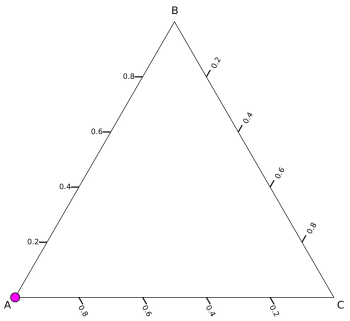
Nash Equilibrium We wish to compute the restricted Mean-Field Nash equilibrium of given policies $\pi_1, ..., \pi_n$. To do this, we store values $(J(\pi_i, \mu^{\pi_j}))_{i,j}$ in a payoff matrix and compute the Nash equilibrium of the two-player game defined as follows : Player 1 receives the payoff received when player 1 chooses a deviating policy i and player 2 chooses the population-generating policy j; Player 2 receives the transposed payoff (*i.e.* Player 1 picks the population-generating policy i and Player 2 picks the deviating policy j).

$$\begin{pmatrix} J(\pi_1, \mu^{\pi_1}) & ... & J(\pi_1, \mu^{\pi_n}) \\ J(\pi_2, \mu^{\pi_1}) & ... & J(\pi_2, \mu^{\pi_n}) \\ ... & ... & ... \\ J(\pi_{n-1}, \mu^{\pi_1}) & ... & J(\pi_{n-1}, \mu^{\pi_n}) \\ J(\pi_n, \mu^{\pi_1}) & ... & J(\pi_n, \mu^{\pi_n}) \end{pmatrix}, \begin{pmatrix} J(\pi_1, \mu^{\pi_1}) & ... & J(\pi_n, \mu^{\pi_1}) \\ J(\pi_1, \mu^{\pi_2}) & ... & J(\pi_n, \mu^{\pi_2}) \\ ... & ... & ... \\ J(\pi_1, \mu^{\pi_{n-1}}) & ... & J(\pi_n, \mu^{\pi_{n-1}}) \\ J(\pi_1, \mu^{\pi_n}) & ... & J(\pi_n, \mu^{\pi_n}) \end{pmatrix}$$

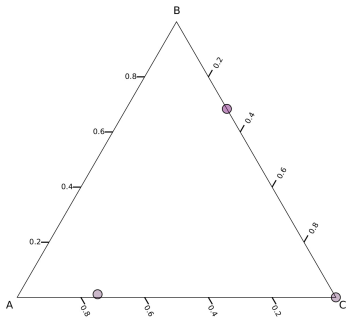
Theorem 67 (Normal-form and restricted game equivalence)**.** *If J is μ-diff-affine, then any symmetric Nash equilibrium of the symmetric two-player game defined above is also a Nash-equilibrium of the restricted Mean-Field game defined by $\pi_1, ..., \pi_n$.*



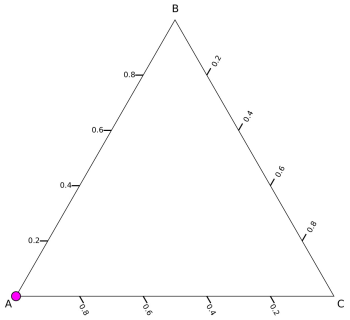
(a) PSRO(CCE) on the Dominated Action game.



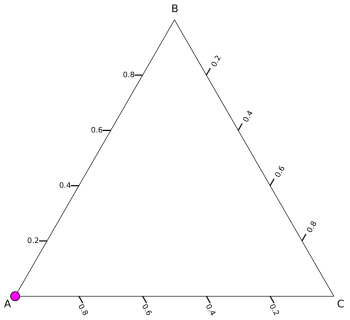
(b) PSRO(CCE) on the Almost-Dominated Action game.



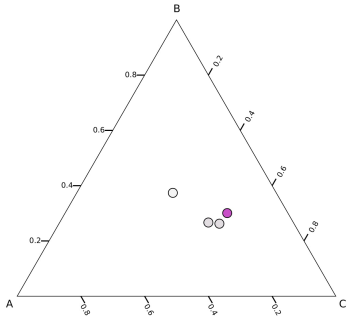
(c) PSRO(CCE) on the Biased Rock-Paper-Scissors game.



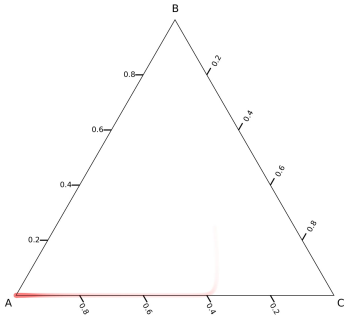
(d) PSRO(CE) on the Dominated Action game.



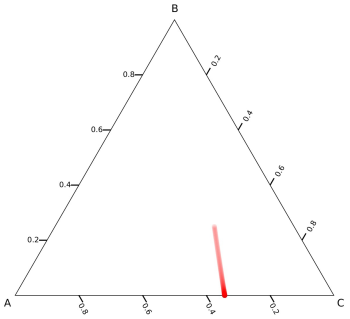
(e) PSRO(CE) on the Almost-Dominated Action game.



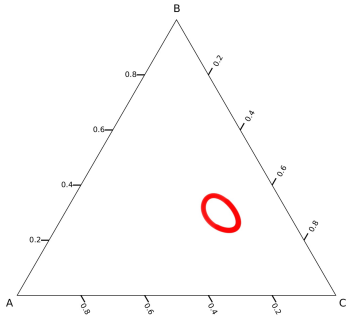
(f) PSRO(CE) on the Biased Rock-Paper-Scissors game.



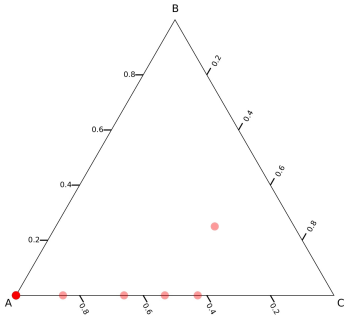
(g) Polynomial Weights (PW) on the Dominated Action game.



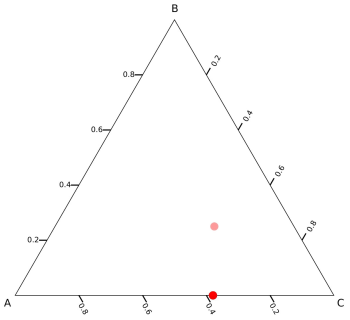
(h) PW on the Almost-Dominated Action game.



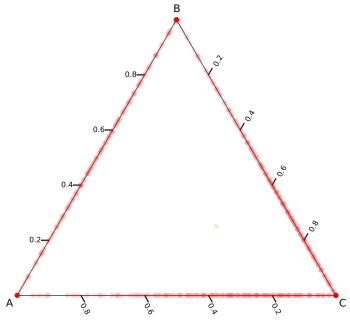
(i) PW on the Biased Rock-Paper-Scissors game.



(j) Regret Matching (RM) on the Dominated Action game.



(k) RM on the Almost-Dominated Action game.



(l) RM on the Biased Rock-Paper-Scissors game.

Figure 5.12: PSRO and PSRO regret minimizers on several Normal-Form Mean-Field Games. The PSRO plots show the final equilibrium learnt by PSRO. For PW and RM, each red circle represents OMD's policy at a given step.

Proof. Let ν be a symmetric Nash equilibrium of the normal-form game. Then we have that

$$\forall \pi' \in \{\pi_1, ..., \pi_n\}, \sum_i \sum_j \nu_i \nu_j \left(J(\pi', \mu^{\pi_j}) - J(\pi_i, \mu^{\pi_j})\right) \leq 0$$

$$\sum_i \nu_i \Delta_{\pi', \pi_i}(J) \underbrace{\left(\sum_j \nu_j \mu^{\pi_j}\right)}_{=\mu(\nu)} \leq \sum_i \sum_j \nu_i \nu_j \Delta_{\pi, \pi_i}(J)(\mu_j^{\pi}) \leq 0$$

$$J(\pi, \mu(\nu)) - \sum_i \nu_i J(\pi_i, \mu(\nu)) \leq 0$$

$$J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu)) \leq 0,$$

where the last line comes from the fact that $\pi(\nu)$ is exactly the policy resulting from sampling π from ν at the start of every episode. Therefore $\pi(\nu)$ is a Nash equilibrium of the game if we restrict deviations to be within the set of the $(\pi_i)_i$. □

We must note one important corollary: since the Nash equilibrium of a μ-diff-affine restricted game can be expressed as the symmetric Nash equilibrium of a 2-player symmetric normal-form game, then, according to the Nash Theorem, this Nash equilibrium always exists. This, in turn, guarantees the existence of correlated and coarse-correlated equilibria in μ-diff-affine games.

[Restricted game equilibrium existence] In a μ-diff-affine restricted game, Nash, correlated and coarse-correlated equilibria always exist.

Restricted-Game Coarse Correlated Equilibrium We define a restricted game coarse correlated equilibrium as a recommendation device ρ which recommends population distributions $\nu \in \Delta(\Pi_n)$ such that

$$\max_{\pi_k} \sum_{\nu} \rho(\nu) \sum_i \sum_j \nu_i \nu_j \left(J(\pi_k, \mu_j) - J(\pi_i, \mu_j)\right) \leq 0$$

$$\text{i.e.} \quad \max_{e_k} \sum_{\nu} \rho(\nu) \left(e_k - \nu\right)^t J \, \nu \leq 0$$

We note that although the set $\Delta(\Pi)$ is not discrete in general, the above equation is written using a sum (Though we note it could also be written using an integral), the reason being algorithmic. Indeed, given K the number of searched different ν with non-zero support in ρ, our optimization process searches for K different $\nu_k \in \mathbb{R}^N$, and their distribution $\rho \in \mathbb{R}^K$, instead of searching over the infinite-dimensional space $\mathcal{P}(\Delta(\Pi))$.

We propose below a maximum-margin solution with parameter K, which one can solve using Quadratic Programming by introducing intermediary variables

$$\begin{aligned} \text{Objective:} \quad & \min_{\rho, \nu_1, ..., \nu_K} \max_{e_l} \sum_k \rho_k \left(e_l - \nu_k\right)^t J \, \nu_k \\ \text{Probability constraint:} \quad & \rho \geq 0, \;\; \underline{1}^t \rho = 1 \\ & \forall k, \;\; \nu_k \geq 0, \;\; \underline{1}^t \nu_k = 1. \end{aligned}$$

Though note that other objectives and formulations are possible, for example a maximum-entropy one. We note that the CCE constraint is quadratic, therefore QCP-capable solvers are required, though expect usual simplifications to hold.

$$\text{Objective:} \quad \min_{\rho,\nu_1,\ldots,\nu_K} \sum_{k=1}^{K} \rho_k \; \nu_k^t \log(\nu_k) \tag{5.11}$$

$$\text{CCE-Constraint:} \quad \forall l, \;\; \sum_k \rho_k \left(e_l - \nu_k\right)^t J \; \nu_k \leq 0 \tag{5.12}$$

$$\text{Probability constraint:} \quad \rho \geq 0, \;\; \underline{1}^t \rho = 1 \tag{5.13}$$

$$\forall k, \;\; \nu_k \geq 0, \;\; \underline{1}^t \nu_k = 1.$$

Or maximum-entropy with KL-regularization imposing differences between population recommendations, for $0 \leq \lambda \leq 1$,

$$\text{Objective:} \quad \min_{\rho,\nu_1,\ldots,\nu_K} \sum_{k=1}^{K} \rho_k \left(\nu_k^t \log(\nu_k) - \lambda \sum_{k'=1}^{K} \nu_k^t \log(\frac{\nu_k}{\nu_{k'}}) \right) \tag{5.14}$$

$$\text{CCE-Constraint:} \quad \forall l, \;\; \sum_k \rho_k \left(e_l - \nu_k\right)^t J \; \nu_k \leq 0 \tag{5.15}$$

$$\text{Probability constraint:} \quad \rho \geq 0, \;\; \underline{1}^t \rho = 1 \tag{5.16}$$

$$\forall k, \;\; \nu_k \geq 0, \;\; \underline{1}^t \nu_k = 1.$$

Restricted-Game Correlated Equilibrium We define a restricted game correlated equilibrium as a recommendation device ρ which recommends population distributions $\nu \in \Delta(\Pi_n)$ such that

$$\sum_i \max_{\pi_k} \sum_\nu \rho(\nu) \sum_j \nu_i \nu_j \left(J(\pi_k, \mu_j) - J(\pi_i, \mu_j)\right) \leq 0$$

$$\sum_i \max_k \sum_\nu \rho(\nu) \nu_i \left(e_k - e_i\right)^t J \; \nu \leq 0$$

And thus, we have

$$\forall i, k, \;\; \sum_\nu \rho(\nu) \nu_i \left(e_k - e_i\right)^t J \; \nu \leq 0.$$

As before, given a fixed number of different population distributions K, we suggest three different optimization objectives : A maximum-margin, quadratic optimization one

$$\text{Objective:} \quad \min_{\rho,\nu_1,\ldots,\nu_K} \sum_i \max_k \sum_\nu \rho(\nu) \nu_i \left(e_k - e_i\right)^t J \; \nu$$

$$\text{Prob. constraint:} \quad \rho \geq 0, \;\; \underline{1}^t \rho = 1$$

$$\forall k, \;\; \nu_k \geq 0, \;\; \underline{1}^t \nu_k = 1.$$

A maximum-entropy one

$$\text{Objective:} \quad \max_{\rho,\nu_1,\ldots,\nu_K} \sum_{k=1}^{K} \rho_k \; \nu_k^t \log(\nu_k) \tag{5.17}$$

$$\text{CE-Constraint:} \quad \forall i, k, \;\; \sum_\nu \rho(\nu) \nu_i \left(e_k - e_i\right)^t J \; \nu \leq 0 \tag{5.18}$$

$$\text{Probability constraint:} \quad \rho \geq 0, \;\; \underline{1}^t \rho = 1 \tag{5.19}$$

$$\forall k, \;\; \nu_k \geq 0, \;\; \underline{1}^t \nu_k = 1.$$

Or a maximum-entropy with KL-regularization imposing differences between population recommendations, when $0 \leq \lambda \leq 1$

$$\text{Objective:} \quad \max_{\rho, \nu_1, \ldots, \nu_K} \sum_{k=1}^{K} \rho_k \left(\nu_k^t \log(\nu_k) - \lambda \sum_{k'=1}^{K} \nu_k^t \log(\frac{\nu_k}{\nu_{k'}}) \right) \tag{5.20}$$

$$\text{CE-Constraint:} \quad \forall i, k, \ \sum_{\nu} \rho(\nu) \nu_i \left(e_k - e_i\right)^t J \, \nu \leq 0 \tag{5.21}$$

$$\text{Probability constraint:} \quad \rho \geq 0, \ \underline{1}^t \rho = 1 \tag{5.22}$$

$$\forall k, \ \nu_k \geq 0, \ \underline{1}^t \nu_k = 1.$$

Mean-Field PSRO: Convergence to Nash equilibria in diff-affine games

MF-PSRO is defined in a very similar way to standard PSRO in diff-affine games: start with a restricted policy set Π_0, and, at each step n, compute the Π_n restricted Nash equilibrium ν_n, and compute a best response π_{n+1} to this Π_n mixed according to ν_n. If $\pi_{n+1} \in \Pi_n$, then Π_n mixed according to ν_n is a Nash equilibrium of the true game, otherwise the algorithm continues, as shown in Algorithm 24.

Algorithm 24 MF-PSRO(Nash) (Diff-Affine case).

$\Pi_1 = \{\pi_1\}$, with π_1 any policy in Π.
$\nu(\pi_1) = 1.0$.
$n = 0$
while $J(BR(\mu^{\pi(\nu)}), \mu^{\pi(\nu)}) > J(\pi(\nu), \mu^{\pi(\nu)})$ **do**
  $\Pi_{n+1} = \Pi_n \cup \{BR(\mu^{\pi(\nu)})\}$.
  $n = n + 1$.
  $\forall i, j \leq n, M_{i,j} = \mathbb{E}[J(\pi_i, \mu^{\pi_j})]$.
  $\nu =$ Matrix Nash Solver$([M, M^t])$.
end while
return Nash equilibrium $\pi(\nu)$.

Mean-Field PSRO: Convergence to (Coarse) Correlated Equilibria in Diff-Affine Games

When the game is μ-diff-affine, we have the following property

In a μ-diff-affine game, any (coarse) correlated equilibrium of the restricted game is a (coarse) correlated equilibrium of the True game when deviations are restricted to the set of known policies $(\pi_n)_n$

Proof. Since the game is μ-diff-affine, for all $\nu \in \Delta(\Pi_n)$, $\pi_k \in \Pi_n$ we have

$$J(\pi_k, \mu(\nu)) - J(\pi_i, \mu(\nu)) = \sum_j \nu_j \left(J(\pi_k, \mu_j) - J(\pi_i, \mu_j) \right).$$

Let ρ be the correlation device of a Mean-Field correlated equilibrium of the restricted game. Then

$$\sum_i \max_{\pi_k \in \Pi_n} \sum_{\nu} \rho(\nu) \nu_i \underbrace{\sum_j \nu_j \left(J(\pi_k, \mu_j) - J(\pi_i, \mu_j) \right)}_{\geq J(\pi_k, \mu(\nu)) - J(\pi_i, \mu(\nu))} \leq 0$$

$$\sum_i \max_{\pi_k \in \Pi_n} \sum_{\nu} \rho(\nu) \nu(\pi_i) \left(J(\pi_k, \mu(\nu)) - J(\pi_i, \mu(\nu)) \right) \leq 0$$

therefore ρ is a Mean-Field correlated equilibrium.

Let ρ be the correlation device of an MFCCE of the restricted game. Then

$$\max_{\pi_k \in \Pi_n} \sum_i \sum_\nu \rho(\nu)\nu_i \underbrace{\sum_j \nu_j \left(J(\pi_k, \mu_j) - J(\pi_i, \mu_j)\right)}_{\geq J(\pi_k, \mu(\nu)) - J(\pi_i, \mu(\nu))} \leq 0$$

$$\max_{\pi_k \in \Pi_n} \sum_\nu \rho(\nu) \left(J(\pi_k, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\right) \leq 0,$$

which concludes the proof. □

We present the modified PSRO version for diff-affine games in Algorithm 25.

Algorithm 25 MF-PSRO((C)CE) (Diff-Affine case).

$\Pi_1 = \{\pi_1\}$, with π_1 any policy in Π.
$\rho(\delta_{\pi_1}) = 1.0$.
$n = 0$
while $\Pi_{n+1} \neq \Pi_n$ **do**
  $\Pi_{n+1} = \Pi_n \cup \{BR_{(C)CE}(\Pi_n, \rho)\}$.
  $n = n + 1$.
  $\forall i, j \leq n, M_{i,j} = \mathbb{E}[J(\pi_i, \mu^{\pi_j})]$.
  $\rho =$ Restricted-game Mean-Field (C)CE Solver($[M, M^t]$).
end while
return Nash equilibrium $\pi(\nu)$.

5.4 Limitations and Future Work

Despite their modularity, several improvements on our algorithms can be envisioned for further research. First, none of them can efficiently select higher-welfare (C)CEs over lower ones, and it is not clear how to modify them to reliably choose some (C)CEs over others. The specific problem of finding high**er**-welfare (C)CEs is known to be NP-Hard in general, but learning approaches could hold the key to unlocking these possibilities.

It is also unclear how to adapt any of these methods to general equilibria, as was done at the end of Chapter 3. This is due to the difficulty of explicitly computing normal-form equilibria in Mean-Field games. It is not clear either how they could, more straightforwardly and as was hinted at the end of Chapter 4, be used to converge to extensive-form correlated equilibria. Looking into Mean-Field CFR methods could be the key to reaching those.

Regarding Mean-Field PSRO, Mean-Field PSRO(Nash) relies on a black-box algorithm, whose characteristics strongly impact the speed and equilibrium accuracy of the algorithm. Finding a principled, general and fast Nash solver in complex restricted games, like we have for Mean-Field (C)CEs, could yield great improvements, both theoretically and performance-wise.

Finally, our methods are much slower than last-iterate OMD or Fictitious Play because they either rely on using empirical distributions (OMD, JFP), which yields extremely complex equilibria (finite correlation devices with a large number of recommended ν) where each component is of little importance - this makes finding best responses very difficult, thus making JFP even slower, since a best-response must be computed by taking into account every ν recommended by the correlation device -, or use bandit methods to compute equilibria, which induces a combination of slow payoff evaluation (be it sampled payoff or exact payoff) and relatively large amounts of steps needed to find a restricted equilibrium. Speeding up these algorithms would be a great improvement.

Chapter 6

Send the World Flying: Penalty Kicks and Applications of Equilibria

We start with an appetizer to show that Game Theory can be applied to real life situations, by taking the example of Penalty Kicks from [177]. The question is simple. In a penalty kick situation, goalkeepers do not have time to read where the ball is going - if they do and have not jumped to a side of the goal by the time the ball is flying, they will almost certainly not manage to catch it.

Goalkeepers must therefore decide on which side to jump *before the ball has been kicked* by the kicker, and jump *as the ball is kicked by the kicker*.

This situation can be formalized as a normal-form-game where both players, the kicker and the goalkeeper, can be considered to have three actions, Kicking / Jumping Left or Right, or in the Middle, which they choose simultaneously.

Prior work from Palacios-Huerta [144] examines penalty kick scenarios from a game-theoretic perspective, using empirical payoff tables to determine whether the associated kickers and goalkeepers play a Nash equilibrium., and simplifying the above game to a 2-player, 2-action games by noting that shooting Left for a left-footed player is the same thing as shooting Right for a right-footed player. They therefore introduce the notion of a Natural Side action (the easiest action according to kicker footedness, which also includes the Center), and a Non-Natural Side action (the harder action).

Here we revisit the work of Palacios-Huerta [144], by first reproducing several of its key results with a substantially larger and more recent data set from the main professional football leagues in Europe, Americas, and Asia (for comparison, the data set used in the work of Palacios-Huerta [144] consists of 1417 penalty kicks from the 1995-2000 period, whereas ours contains 12399 kicks from the 2011-2017 period). While several results of this earlier work are corroborated, including the decomposition of the game into the Natural / Non-Natural side actions, we also find surprising new additional insights under our larger dataset. We then go further to extend this analysis by considering larger empirical games (involving more action choices for both kick-takers and goalkeepers). Finally, we develop a technique for illustrating substantive differences in various kickers' penalty styles, by combining empirical game-theoretic analysis with Player Vectors [51] illustrating the added value and novel insights research of the microcosm can bring to football analytics.

6.1 Palacios-Huerta [144] Reproduction

For our analysis we use a data set of 12399 penalty kicks based on Opta data [1]. In Figure 6.1 we show a heatmap of the shot distribution of the penalty kicks in our data set. Table 6.1 shows the

Table 6.1: Penalty kick distribution over leagues considered (12399 kicks in total).

League	# Kicks
Italian Serie A	607
US Major League Soccer	575
Engl. Npower Champ.	569
Spanish Segunda Division	568
Spanish La Liga	531
French Ligue 1	497
German DFB Pokal	441
Brazilian Série A	440
Engl. Barclays Premier League	436
German Bundesliga	409
Dutch Eredivisie	398
German Bundesliga Zwei	389
Portuguese Primiera Liga	352
Saudi Arabian Profess. League	337
Russian Premier League	329
Chinese Super League	324
Copa Libertadores	322
Belgian Jupiler League	287
Turkish Super Lig	284
French Ligue 2	270
Argentina Primera (Anual)	261
English Capital One Cup	234
Mexican Primera (Clausura)	234
Colombia Primera Apertura	221
Norwegian Tippeligaen	219
AFC Champions League	193
International Champions Cup	188
Australian A-League	172
Copa Chile	172
English FA Cup	153
Copa do Brasil	153

League	# Kicks
Chile Primera (Apertura)	151
Japanese J-League	149
English League 1	139
English League 2	130
Austrian Bundesliga	129
Danish Superligaen	115
European World Cup Qualifiers	108
Internationals	93
African Cup of Nations	90
United Soccer League	80
Europ. Championship Qualifiers	78
Swedish Allsvenskan	74
Coppa Italia	67
Copa America	51
FIFA Club World Cup	51
World Cup	48
Europ. Championship Finals	45
Champions League Qualifying	41
Confederations Cup	39
UEFA Europa League Qualifying	32
Coupe de France	29
Belgian UEFA Europa Play-offs	24
German 3rd Liga	23
Russian Relegation Play-offs	15
Dutch Relegation Play-offs	13
Copa Sudamericana	9
Friendly	4
German Bundesliga Playoff	3
German Bundesliga 2 Playoff	3
Swedish Relegation Play-off	1

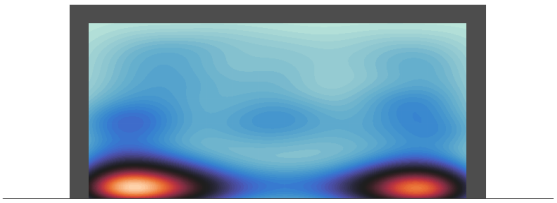


Figure 6.1: Visualization of shot distribution for the penalty kicks in the considered dataset.

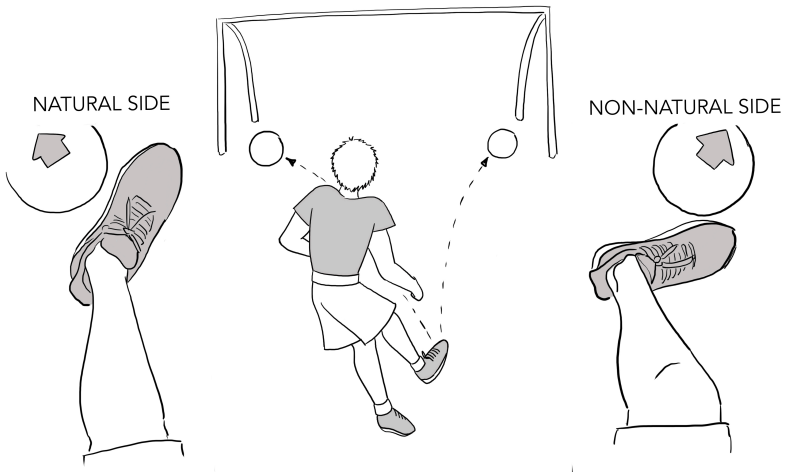


Figure 6.2: Illustration of natural vs. non-natural sides. All credit to Marta Garnelo for this beautiful artwork !

distribution of the penalty kicks over the various leagues we consider.

As in Palacios-Huerta [144]'s work, we first synthesize a 2-player 2-action empirical game based on our penalty kick data set. Table 6.2a illustrates the 2×2 normal form as presented by Palacios-Huerta [144]. The actions for the two players, the kicker and goalkeeper, are respectively visualized in the rows and columns of the corresponding payoff tables, and are detailed below. The respective payoffs in each cell of the payoff table indicate the win-rate or probability of success for the kicker (*i.e.* a score); for ease of comparison between various payoff tables, cells are color-graded in proportion to their associated values (the higher the scoring probability, the darker shade of green used).

The choice of player actions considered has an important bearing on the conclusions drawn via empirical game-theoretic analysis. The actions used by Palacios-Huerta [144] in Table 6.2a correspond to taking a shot to the natural (N) or non-natural (NN) side for the kicker, and analogously diving to the natural side or non-natural side for the goalkeeper. Figure 6.2 provides a visual definition of natural versus non-natural sides. Specifically, as players tend to kick with the inside of their feet, it is easier, for example, for a left-footed player to kick towards the right (from their perspective); thus, this is referred to as their natural side. Analogously, the natural side for a right-footed kicker is to kick towards their left. The natural side for a goalkeeper depends on the kicker in front of him. Specifically, when facing right-footed kickers, goalkeepers' natural side is designated to be their right; vice versa, when they face a left-footed kicker, their natural side is to their left. Importantly, shots to the center count as shots to the natural side of the kicker, because, as explained in Palacios-Huerta [144], kicking to the center is considered equally natural as kicking to the natural side by professional football players [144].

Figure 6.2b shows our reproduction of Figure 6.2a of Palacios-Huerta [144], computed using

Table 6.2: Natural (N) / Non-Natural (NN) payoff tables for Shots (S) and Goalkeepers (G). Here, Tables c and d compare Nash and empirical probabilities.

(a) Palacios-Huerta [144] payoff table.

	N-G	NN-G
N-S	0.670	0.929
NN-S	0.950	0.583

(b) Reproduced table.

	N-G	NN-G
N-S	0.704	0.907
NN-S	0.894	0.640

(c) Palacios-Huerta [144] Nash probabilities.

	NN-S	N-S	NN-G	N-G
Nash	0.393	0.607	0.432	0.568
Empirical	0.423	0.577	0.400	0.600

Jensen–Shannon divergence: 0.049%

(d) Reproduced table Nash probabilities.

	NN-S	N-S	NN-G	N-G
Nash	0.431	0.569	0.408	0.592
Empirical	0.475	0.525	0.385	0.615

Jensen–Shannon divergence: 0.087%

Table 6.3: Natural / Non-natural game restricted by footedness.

(a) Left-footed players payoff table

	N-G	NN-G
N-S	0.721	0.939
NN-S	0.903	0.591

(b) Right-footed players payoff table

	N-G	NN-G
N-S	0.700	0.898
NN-S	0.892	0.653

12399 penalty kicks spanning the aforementioned leagues in our Opta-based dataset; importantly, players (goalkeepers and kickers) appear at least 20 times each in this dataset, to ensure consistency with Palacios-Huerta [144]. The trends in these two tables are in agreement: when the goalkeeper and the kicker do not choose the same sides of the goal, shot success rate is high; otherwise, when the keeper goes to the same side as the kicker, success rate is higher for natural shots than for non-natural shots. We also include Nash and empirical probabilities for Palacios-Huerta's dataset and ours, respectively in Tables 6.2c and 6.2d, enabling us to conclude that payoffs, Nash probabilities, and empirical probabilities are all in agreement between Palacios-Huerta's results and our reproduction; more quantitatively, the Jensen-Shannon divergence between Palacios-Huerta's results and ours is 0.84% for the Nash distribution and 1.2% for the empirical frequencies. We also notice that players' empirical action selection frequencies are quite close to the Nash-recommended frequencies, as measured by their Jensen-Shannon Divergence, and are actually playing an ϵ-Nash equilibrium with a very low ϵ of 0.4%.

6.2 Natural Side Analysis

Having examined the similarity of payoff tables and distributions, we verify whether the Natural / Non-Natural game is statistically identical for left-footed and right-footed players (Table 6.3), as assumed in Palacios-Huerta [144]. To do so, we use a t-test to verify whether per-cell scoring rates

Table 6.4: Footedness equivalence p-value tables.

(a) Natural / Non-natural game p-values

	N-G	NN-G
N-S	0.924566	0.170504
NN-S	0.394900	0.407741

(b) Left / Center / Right game p-values

	R-G	C-G	L-G
R-S	0.000011	0.947369	6.931197e-01
C-S	0.592054	0.868407	1.305657e-01
L-S	0.017564	0.764020	7.791136e-07

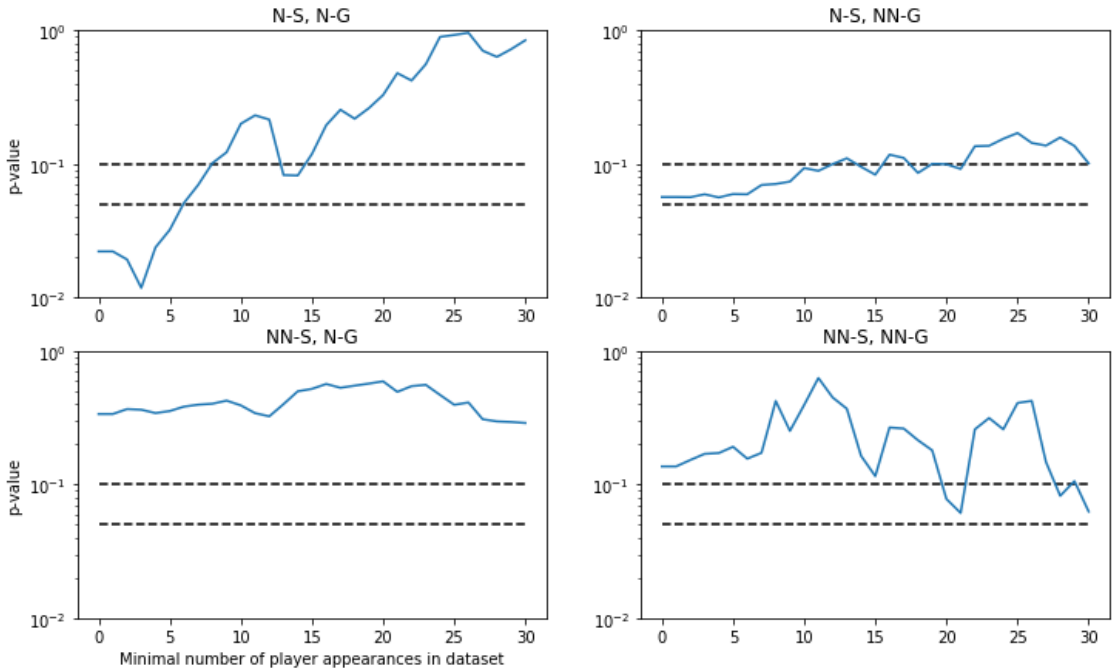


Figure 6.3: P-value table as a function of minimal experience.

are identical across footedness types. The t-tests' p-values are reported in Table 6.4a, and reveal that the games cannot be proven to be dissimilar across footedness with reasonable confidence and can, therefore, be assumed to be identical for left-footed and right-footed players. Figure 6.3 refines this result by representing the relationship between p-values of our t-test and minimal player appearance counts: when we modulate minimal appearance count of players in our test, the Natural Shot / Natural Goalkeeper cell goes from strongly dissimilar across footedness (low p-value) when including all players, to likely non-dissimilar (high p-value) when only including the players appearing the most in our dataset. This could be explained by low appearance counts, which we take here as a proxy for low experience, kickers being less able to control their kicks, resulting in different control effectiveness for different footedness preferences, and in goalkeepers being less proficient in stopping shots going to their less frequently-kicked side (left) than to the other, a preference that we infer has been trained away in professional goalkeepers. To remove potential side-effects of merging data from low- and high-experience players together, Figure 6.4 shows the relationship between p-values of our t-test and experience category where we allow for some overlap–between 1 and 7 shots, 5 and 12, etc.; the insight drawn from this figure is the same as that of Figure 6.3, supporting the conclusion that experience removes the difference between left- and right-footed penalty kicks.

We also analyzed the game defined by kicking to the left, center, or right, and confirmed Palacios-Huerta's intuition that it is fundamentally different across footedness preferences. Specifically,

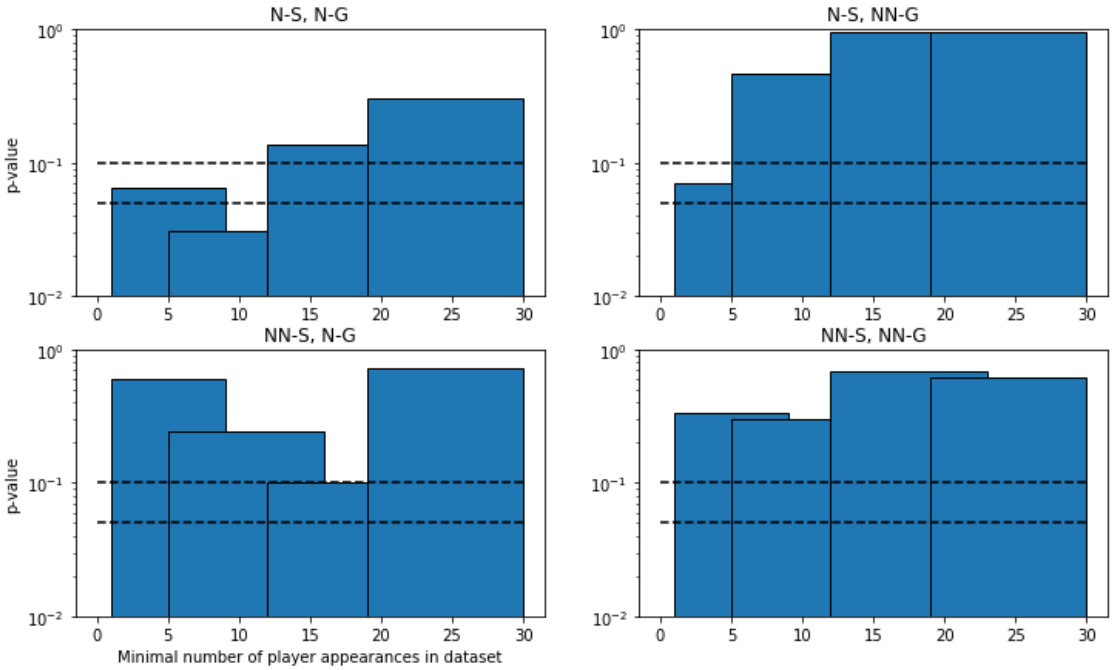


Figure 6.4: P-value table as a function of player-experience.

Table 6.5a synthesizes the empirical game corresponding to this new choice of actions, with aggregated scoring rates over both feet preferences. Note that in this case, left, center, and right are measured from the goalkeeper's perspective, such that the natural kick of a right-footed player would be considered a right kick. The per-cell t-tests' p-values for this game are reported in Table 6.4b. Interestingly, the game is different across footedness when the goalkeeper jumps to the same side as the ball, but is otherwise mostly similar across footedness preference. The empirical play frequencies for kickers, as reported in Table 6.5b, are also further away from Nash frequencies than observed in the Natural / Non-Natural game (Table 6.2d), as can be seen from the Jensen-Shannon divergence between empirical frequencies and Nash (0.75%, versus the the 0.087% of the Natural / Non-Natural game) These insights indeed confirm the intuition that such a game is neither correct across footedness, nor the one the players follow.

Overall, these results provide insights into the impacts that the choice of actions have on conclusions drawn from empirical payoff tables, and are illustrative of the practical usefulness of theoretically well-founded principles in application to real-world analytics, which is highlighted as a hallmark of useful theory in Szymanski [172].

However, behavior and shooting styles also vary wildly per-player given footedness. If one is willing to consider several payoff tables (e.g., one per footedness), it seems natural to also take into account kickers' playing styles, as considered in the next section.

6.3 Augmenting Game-Theoretic Analysis of Penalty Kicks with Embeddings

While the previous section undertook a descriptive view of the penalty kick scenario (*i.e.* providing a high-level understanding of kicker and goalkeeper play probabilities), here we investigate whether we can find the best strategy for a player given the knowledge of the kicker's play style. In game-theoretic terms, we conduct a prescriptive analysis of penalty kicks to enable informed decision-making for players and coaching staff in specific penalty kick situations. Ideally, one would iterate the earlier empirical payoff analysis for every possible combination of goalkeeper and

Table 6.5: Left (L) - Center (C) - Right (R) tables for Shots (S) and Goalkeepers (G), with the three directions of kick/movement defined from the goalkeeper's perspective.

(a) Payoff table.

	R-G	C-G	L-G
R-S	0.684	0.939	0.969
C-S	0.964	0.160	0.953
L-S	0.964	0.960	0.633

(b) Nash probabilities vs. Empirical frequencies corresponding to a.

	R-S	C-S	L-S	R-G	C-G	L-G
Nash	0.478	0.116	0.406	0.441	0.178	0.381
Empirical	0.454	0.061	0.485	0.475	0.089	0.436

Jensen–Shannon divergence: 0.75%

Table 6.6: Cluster statistics.

	# Players	# Goals	# Shots	Success %	Proportion left-foot goals (%)
Cluster 1	197	144	167	86.2	10.4
Cluster 2	216	494	612	80.7	21.9
Cluster 3	52	3	4	75.0	33.3
Cluster 4	82	58	73	79.4	51.7
Cluster 5	87	44	60	73.3	34.1
Cluster 6	1	0	0	-	0.0
Total	635	743	916	81.1	25.2

kicker in a given league, thus enabling decision-making at the most granular level; however, the inherent sparsity of penalty kick data makes such an approach infeasible. Instead, we introduce a meaningful compromise here by combining statistical learning with game theory, first quantifying individual playing styles, then using clustering techniques to aggregate players (*i.e.* both strikers and goalkeepers) based on said styles, and finally synthesizing empirical games for each identified cluster. We focus our analysis on penalties including all players who participated in Premier League matches from 2016 to 2019.

On a technical level, our approach consists of the three following steps. First, we characterize the playing style of a player in a manner that can be interpreted both by human experts and machine learning systems. In particular, we use Player Vectors [51] to summarize the playing styles of kickers using an 18-dimensional real-valued vector. These Player Vectors are extracted from historical playing trajectories in real matches, as done by [51]. Each dimension of the Player Vector corresponds to individual on-pitch player behaviors (e.g., styles of passes, take-ons, shots, etc.), and the value of each dimension is standardized and quantifies the weight of that particular action style for the considered player. We also filter experienced players with at least 50 appearances in the Premier League matches from 2016 to 2019. In total, we obtain 635 such vectors for the individual players in our dataset. Second, we cluster players in accordance to their Player Vectors, using K-means with the number of clusters chosen as the value causing the most significant drop in inertia (a standard heuristic). This process yields 6 clusters in total, with statistics summarized in Table 6.6. In particular, K-means clustering detects an outlier cluster with only one player (Cluster 6), and we also observe that there are very few shot samples in Cluster 3, as it consists of a cluster of goalkeepers (an interesting artifact illustrating the ability of Player Vectors and K-means clustering to discern player roles). Given the few samples associated with these two clusters, we

Table 6.7: Pair-wise comparison for the identified clusters. < indicates that data was missing and minimum true p-value may be lower than the reported minimum p-value in the table. The symbol * indicates we cannot reject the equality hypothesis at the 5% confidence level.

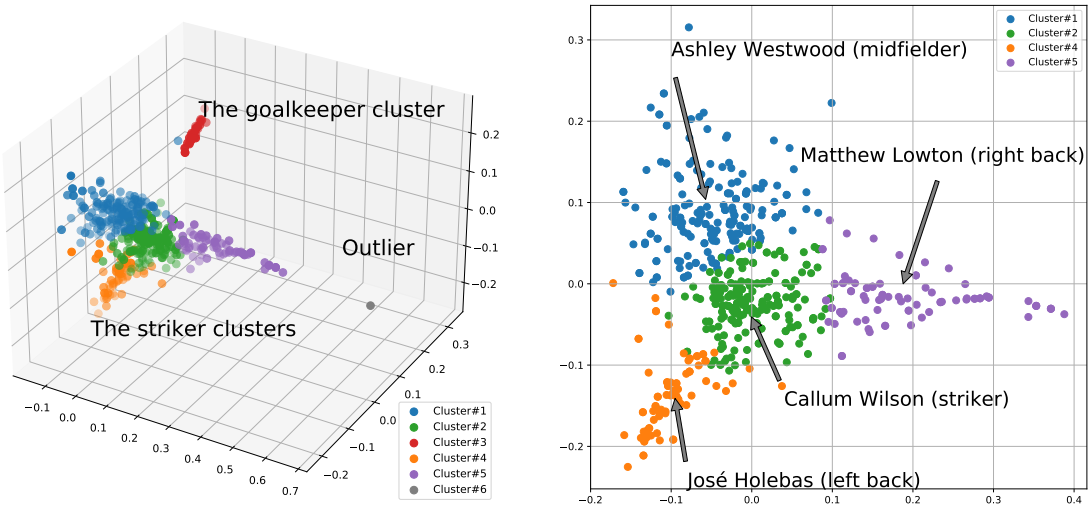
	1 vs. 2	1 vs. 4	1 vs. 5	2 vs. 4	2 vs. 5	4 vs. 5
Min. cell p-value of t-test over table equality	4.49e-2	< 9.56e-2*	< 1.09e-1*	4.49e-2	4.48e-2	< 3.39e-1*
Jensen-Shannon div. between Nash distr. (%)	0.03	0.57	0.09	0.35	0.02	0.21
Jensen-Shannon div. between empirical distr. (%)	0.06	0.01	0.06	0.08	0.24	0.04
Left footedness t-test p-value	3.43e-4	1.37e-7	3.18e-3	4.92e-5	1.07e-1	7.52e-2

Table 6.8: p-values for t-test that empirical action distributions are equal among different clusters. Minimum p-value (across kicker and goalkeeper roles) is indicated in bold for each row.

Kicker clusters compared	Kicker p-value	Goalkeeper p-value
1 vs. 2	0.52	**0.05**
1 vs. 4	**0.85**	0.95
1 vs. 5	0.42	**0.27**
2 vs. 4	0.52	**0.14**
2 vs. 5	0.51	**0.16**
4 vs. 5	0.4	**0.26**

henceforth exclude them from the game-theoretic analysis. We observe that cluster pairs (1, 2), (1, 4), (2, 4), and (2, 5) are significantly different, with the minimum cell-wise p-values for these cluster pairs smaller than 0.10 in Table 6.7. We therefore focus our game-theoretic analysis on these cluster pairs. Moreover, we also qualitatively illustrate differences between the clusters in Figures a,b, which visualize the results of reducing the Player Vectors dimensionality from 18 to, respectively, 3 and 2 via Principal Component Analysis. Here, we observe that the goalkeeper cluster is well-separated from the kicker clusters in Figure 6.5a, and in order to better visualize the kicker clusters, we project Figure a onto its x and y axis after removing the goalkeeper and outlier clusters in Figure b. We also identify therein the most representative kicker per-cluster (*i.e.* the player whose feature vector is closest to the mean of the corresponding cluster)

Finally, we conduct the aforementioned game-theoretic analysis for each cluster. In Table 6.6, we observe that the kickers in some clusters have different success rates in penalty kicks. Moreover, a closer behavioral analysis yields deeper insights. We first examine the Nash strategies played by each cluster, and then visualize the actual play behavior with respect to empirical probabilities in Figure 6.6. Table 6.9a summarizes the overall Nash distributions for all players considered, its subtables showing cluster-specific distributions. These tables illustrate that the kickers have the same empirical behavior, an assertion statistically confirmed in Table 6.8; yet their Nash-derived recommendations are different: although kickers in all clusters are recommended by the Nash to shoot more to their natural sides than to their non-natural sides, the recommended strategy for kickers in Cluster 1 is actually quite balanced between natural and non-natural shots. This greater imbalance is shown by comparing Jensen-Shannon divergence. As we see in Table 6.7, the Jensen-Shannon divergence of the Nash probabilities between Cluster 1 and 4 (0.57%) is 6-7 times greater than that between Cluster 1 and 5 (0.09%) and 19 times greater than that between



(a) (b)

Figure 6.5: Visualization of the identified player clusters. Subfigure a visualizes the goalkeeper cluster, the kicker clusters and an outlier automatically detected through K-means clustering. To show the separation of the kicker clusters clearly, we visualize them in Subfigure b after removing the goalkeeper and outlier clusters, and we also label each cluster with a Premier League player in it.

Cluster 1 and 2 (0.03%). We also notice that the clusters' players are all playing epsilon Nash equilibria with relatively low epsilon (Table 6.9). In other words, although their empirical strategies seem to deviate from corresponding Nash strategies action-wise, the expected payoffs of these two strategies are close, and they could still stand to gain in "stability" by switching to corresponding Nash strategy. Nevertheless, most of these Nash recommendations come from very low-sample empirical payoff tables, which entails potentially inaccurate Nash distributions. We nevertheless note that this low-data regime is induced by the restriction of our analysis to players having played in matches of Premier League only from 2016 to 2019. Obtaining Player Vector data for all players in our dataset would allow us to study cluster behavior with greater statistical precision. Nevertheless, the current study leaves no statistical doubt regarding the pertinence of clustering payoff tables using player embeddings–specifically Player Vectors.

Qualitatively, in addition to analyzing the strategies with respect to Nash probabilities, the patterns of positions of the ball of successful goals also vary from clusters to clusters, as visualized in Figure 6.6. For instance, kickers in Cluster 2 tend to score mostly to the bottom left corner of the goalmouth, while the scoring positions in other clusters are more balanced, though these could also be partly due to lower sample sizes for some clusters.

6.4 Limitations and Future Work

The current approach may not work on other, less data-straightforward situations such as corner kicks and freekicks, and it is yet unclear how to adapt game-theoretic methods to these cases. Generalizing this type of analysis to other situations still requires learning to automatically create game-theoretic-relevant abstractions. A promising research direction involves combining player vectors with unsupervised learning tools to derive interesting metrics, such as goal probability or shot probability.

Table 6.9: Nash probabilities and empirical (Empir.) frequencies tables for Shot (S) and Goalkeepers (G) with Natural (N) and Non-Natural (NN) actions. Note that Cluster 3 is omitted due to it consisting of very few shots (taken by goalkeepers).

(a) All players. 916 total shots.

	NN-S	N-S	NN-G	N-G
Nash	0.391	0.609	0.406	0.594
Empir.	0.503	0.497	0.413	0.587

ϵ-Nash equilibrium: $\epsilon = 2.71\%$

(b) Kickers in Cluster 1. 167 total shots.

	NN-S	N-S	NN-G	N-G
Nash	0.423	0.577	0.379	0.621
Empir.	0.485	0.515	0.371	0.629

ϵ-Nash equilibrium: $\epsilon = 0.08\%$

(c) Kickers in Cluster 2. 612 total shots.

	NN-S	N-S	NN-G	N-G
Nash	0.401	0.599	0.430	0.570
Empir.	0.520	0.480	0.418	0.582

ϵ-Nash equilibrium: $\epsilon = 2.89\%$

(d) Kickers in Cluster 4. 73 total shots.

	NN-S	N-S	NN-G	N-G
Nash	0.320	0.680	0.375	0.625
Empir.	0.479	0.521	0.438	0.562

ϵ-Nash equilibrium: $\epsilon = 5.17\%$

(e) Kickers in Cluster 5. 60 total shots.

	NN-S	N-S	NN-G	N-G
Nash	0.383	0.617	0.317	0.683
Empir.	0.450	0.550	0.400	0.600

ϵ-Nash equilibrium: $\epsilon = 4.86\%$

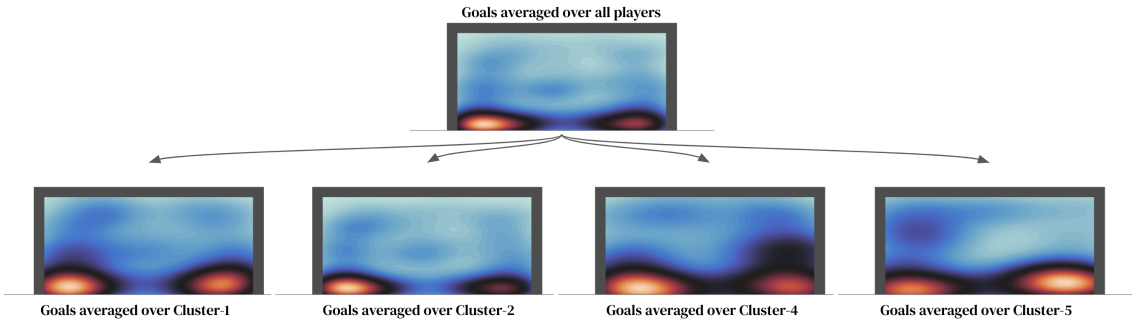


Figure 6.6: Heatmaps of goals by all kickers and kickers in individual clusters with respect to empirical probabilities. We exclude the goalkeeper cluster (Cluster 3) and the outlier cluster (Cluster 6) because of insufficient samples.

Chapter 7

Blissful renewal: Conclusion

At the beginning of this thesis, we laid out our research ambitions and research questions, to which we answered one-by-one in each subsequent chapter. This section will first explore how we answered each of these, then will summarize the contributions of our work in general, to finish - with great emotions - with the work's current limitations and future directions.

7.1 Answers to Research Questions

We reiterate here our research questions, and summarize our answers to them.

1. **Given a game theoretic equilibrium concept, how does one reach it in *any finite game*?**

We have provided, in Chapter 3, two examples of a modification of PSRO which allowed it to converge towards α-Rank, and towards correlated and coarse-correlated equilibria. These examples were generalized into a version of PSRO capable of converging towards *any equilibrium* of a specific form which fits the classical equilibria mentioned in this section, in *any finite game*.

2. **What are the Mean-Field equivalents of N-player equilibria? How can we use them to approximate N-player equilibria when N is very large?**

We answered both questions in Chapter 4. We first started by describing a few new Mean-Field equilibria. Nash equilibria had already been quite investigated in the Mean-Field case; we therefore analyzed this question in the case of correlated and coarse-correlated equilibria in Chapter 4. Starting from symmetric-anonymous games, we simplified the expression of correlated and coarse-correlated equilibria to only depend on the distribution of play of other players. This allowed for seamlessly passing to the Mean-Field limit. Once this limit had been passed, we provided an in-depth study of these new equilibria's properties. we then demonstrated that Mean-Field Nash, correlated and coarse-correlated equilibria can all be reused in N-player games, with an error of $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ when these equilibria are not continuous distributions (which is always the case for the algorithms we use). This shows that when we compute a Mean-Field equilibrium, we also compute an approximate N-player equilibrium, whose quality increases with N, which answers the second question.

3. **How can we compute equilibria in Mean-Field games?**

We answered this question in Chapter 5: We first define a notion of Mean-Field regret, then proved that two popular algorithms, Online Mirror Descent and Joint Fictitious Play, were

no-external-regret, a notion we proved meant they were converging towards Mean-Field coarse-correlated equilibria. We then introduced Mean-Field PSRO, which can converge to Mean-Field Nash, coarse-correlated and correlated equilibria depending on its parametrization. Combining this research question's answer with the former research question's answers, we now have a general method for computing equilibria in large-N N-player games: first compute the Mean-Field equilibrium of the game using any algorithm of choice, then reuse that equilibrium in the N-player version, yielding an acceptable approximation of the true equilibrium of the N-player game.

4. **How can we apply game-theoretic equilibria to optimize real-world scenarios?**

Chapter 6 provides an example of Game Theory used to analyze, and provide recommendations to, players' play styles in the case of set pieces. It provides a concrete, real-world example advocating for the use of Game Theory to improve real-world situations' outcomes.

7.2 Contributions of our Work

We now summarize the main contributions of our work:

On equilibrium computation: Chapter 3 provides a set of adaptable methods to compute equilibria for *any finite game*, settling once and for all the problem of finding equilibria. However, the problem of *quickly* finding equilibria remains open, as these methods remain computationally intense.

On PSRO: Chapter 3 provides PSRO-derived methods for reaching *all* equilibria in *all* player games, and analyzes two particular cases of equilibria, α-Rank and (coarse-)correlated equilibria, in detail. These developments, combined with the development of Chapter 4 regarding Mean-Field PSRO, have greatly contributed to the literature on PSRO methods, widening considerably our understanding of the extremely flexible capabilities of the algorithm.

On Mean-Field games: Chapter 4 introduces new equilibrium concepts to Mean-Field games, an extension of a classical N-player equilibrium: correlated equilibria. It provides an extensive study of this concept, exploring its properties of existence, relationship to other known equilibrium concepts, and algorithms to reach them. Through one of these algorithms, Mean-Field PSRO, it also provides the first Mean-Field algorithm able to compute Nash equilibria in *all Mean-Field games* of our framework.

On Regret-Minimization: Chapter 4 also provides, in its description of Mean-Field PSRO, a new method to compress the equilibria produced by no-regret learners, *bandit compression*. Empirically, this method significantly speeds up PSRO's equilibrium computation step, improved its computed equilibria's sparsity and its approximation quality.

On Sports Analytics: Chapter 6 shows an example of Game Theory providing clear insights into a common situation in Football, and how Game-Theory may help analyze and recommend new and better actions for players and coaches in the future.

7.3 Limitations of our Work and Future Directions

As with any work, our developments are a stepping stone to many potential improvements. We have categorized its limitations and interesting future directions in the same way as above:

PSRO: The main limitation of all PSRO-derived methods presented in this work is scalability. Indeed, despite the affordance for Deep Reinforcement Learning that PSRO naturally incorporates, the algorithm still potentially needs as many iterations as there are deterministic policies in the game before it terminates. One iteration also means one best-response computation step. The number of best responses of a game is typically exponential in its number of states - in very large games, it becomes almost impossible for PSRO to terminate in a reasonable timeframe. Future work could go in two different directions:

1. Speed-up PSRO by changing the algorithm's logic - switching from best-responses to mixed-strategies, and from a static objective (the metasolver's distribution) to a dynamic one. Some unpublished work of this thesis's author have looked at this direction to solve Stratego in the case of our effort [148], but unfortunately without unquestionable success; and the literature hasn't been able to get away from PSRO's worst-case convergence bounds either. This does not however mean that this is impossible, given the great results obtained by PSRO-derived methods on Starcraft [181] or Capture the Flag [88].

2. Adapt demonstrably scalable algorithms to converge to the wished equilibria. Algorithms such as FForel [146] converge to a Nash equilibrium, and have been able to scale to huge games such as Stratego [148]. Finding a generic way to regularize this algorithm, or a similar one, so that it converges towards a desired game-theoretic objective, could be a promising approach. However, how to do so is unclear, as handling joint strategies would intuitively make things much more difficult than the marginal strategy used in Nash equilibria.

Mean-Field: Correlated and coarse-correlated equilibria are inherently difficult to compute [121], and many other derived equilibrium concepts could be worthy of attention, such as every equilibrium introduced by Morrill et al. [121]. Note that doing so would hopefully be a straightforward adaptation of N-player concepts following our logic in extending N-player games correlated equilibria to Mean-Field via symmetric simplifications. Another huge limitation of all Mean-Field methods we are currently using resides in our current necessity to exactly compute policies' state distributions μ instead of estimating them. This means we have to do a full game tree pass to be able to compute the reward at a given state for a given player - in large games, this means that even computing rewards for new policies becomes unscalable. This is a vital area of improvement for our methods, which will allow them to - or prevent them from reaching - scale. Indeed, once this problem is solved, deep-learning-scaled methods such as Deep OMD, or Deep Fictitious Play, which we recently introduced [103], will be able to tackle highly complex environments.

Regret Minimization: Our new method, bandit compression, empirically improves convergence speed, reduces equilibrium complexity, and provides better equilibrium approximations. However, we have only managed to prove that the method in question would never make the equilibrium *worse*, without providing conditions for it to improve, or lower bounds for improvements. This, and the use of the method in traditional bandit settings, constitute fundamentally interesting areas of future work.

Sports Analytics: Our work has only focused on penalty kick situations; however, other set pieces exist in football: corner kicks and free kicks. These are however much more difficult to analyze, as team positioning presumably plays a very big role on set piece outcome - and one thus needs to be able to (1) have access to team formation data, and (2) find a clustering / compression / embedding system for formations that makes sense, in a presumably low data regime. The use of game theory is not restricted to set pieces, however, and could also be extended to *e.g.* broad game tactics in general, team formations, and overall strategy against a given club, given its and our players' strengths and weaknesses. We have already started taking a step in this direction by simulating possible player trajectories given a game state [140], and are eager to couple this type of generative approaches with Game Theory to estimate optimal strategies.

Bibliography

[1] Opta, 2020. URL https://www.optasports.com.

[2] Yves Achdou and Mathieu Laurière. Mean field games and applications: Numerical aspects. *Mean Field Games*, pages 249–307, 2020.

[3] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria, 2022. URL https://arxiv.org/abs/2202.05446.

[4] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*, 2020.

[5] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics, 2021. URL https://arxiv.org/abs/2106.13755.

[6] H. Soner Aplak and M. Ziya Sogut. Game theory approach in decisional process of energy management for industrial sector. *Energy Conversion and Management*, 74:70–80, 2013. ISSN 0196-8904. doi: https://doi.org/10.1016/j.enconman.2013.03.027. URL https://www.sciencedirect.com/science/article/pii/S0196890413001702.

[7] Aristophanes. The birds, 400BC.

[8] Aristotle. Politics, 350BC.

[9] Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

[10] Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

[11] Robert J. Aumann. Correlated equilibrium as an expression of bayesian rationality. 1987.

[12] Alexander Aurell, René Carmona, Gökçe Dayanikli, and Mathieu Laurière. Optimal incentives to mitigate epidemics: A stackelberg mean field game approach. *SIAM Journal on Control and Optimization*, 60(2):S294–S322, 2022. doi: 10.1137/20M1377862. URL https://doi.org/10.1137/20M1377862.

[13] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Pérolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning (ICML)*, 2019.

[14] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech M. Czarnecki, Julien Pérolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. *CoRR*, abs/1901.08106, 2019. URL http://arxiv.org/abs/1901.08106.

[15] Siddharth Barman and Katrina Ligett. Finding any nontrivial coarse correlated equilibrium is hard, 2015.

[16] Alain Bensoussan, Jens Frehse, and Sheung Chi Phillip Yam. *Mean Field Games and Mean Field Type Control Theory*. Springer Briefs in Mathematics. Springer, New York, 2013. ISBN 978-1-4614-8507-0; 978-1-4614-8508-7.

[17] Dirk Bergemann and Stephen Morris. *Robust Mechanism Design:The Role of Private Information and Higher Order Beliefs*. World Scientific Publishing Co. Pte. Ltd., 2012. URL `https://EconPapers.repec.org/RePEc:wsi:wsbook:8318`.

[18] Charles Bertucci. Optimal stopping in mean field games, an obstacle problem approach. *Journal de Mathématiques Pures et Appliquées*, 120:165–194, 2018.

[19] Patrick Billingsley. *Convergence of Probability Measures*. Wiley series in Probability and Statistics, 1999.

[20] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res. (JAIR)*, 53:659–697, 2015.

[21] A. Blum and Y. Mansour. Learning, regret minimization, and equilibria. In Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4, pages 79–102. Cambridge University Press, 2007.

[22] Avrim Blum and Yishay Mansour. From external to internal regret, 2005.

[23] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: On convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, pages 45–52, 2006.

[24] Géraldine Bouveret, Roxana Dumitrescu, and Peter Tankov. Mean-field games of optimal stopping: a relaxed solution approach. *SIAM Journal on Control and Optimization*, 58(4): 1795–1821, 2020.

[25] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up Limit Hold'em Poker is solved. *Science*, 347(6218):145–149, January 2015.

[26] Louis Bremaud and Denis Ullmo. A social structure description of epidemics propagation with the mean field game paradigm, 2022. URL `https://arxiv.org/abs/2206.11399`.

[27] George W Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951.

[28] G.W. Brown. Iterative solutions of games by fictitious play. *In Activity Analysis of Production and Allocation, T.C. Koopmans (Ed.), New York: Wiley.*, 1951.

[29] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 360(6385), December 2017.

[30] Noam Brown, Tuomas Sandholm, and Strategic Machine. Libratus: The superhuman AI for no-limit poker. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

[31] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. 2018. doi: 10.48550/ARXIV.1811.00164. URL `https://arxiv.org/abs/1811.00164`.

[32] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games, 2020. URL `https://arxiv.org/abs/2007.13544`.

[33] Edmund Burke. Reflections on the revolution in france, 1790.

[34] Luciano Campi and Markus Fischer. Correlated equilibria and mean field games: a simple model. *arXiv preprint arXiv:2004.06185*, 2020.

[35] Pierre Cardaliaguet. Notes on mean field games, 2013.

[36] Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

[37] Pierre Cardaliaguet and Catherine Rainer. On the (in) efficiency of mfg equilibria. *SIAM Journal on Control and Optimization*, 57(4):2292–2314, 2019.

[38] Rene Carmona. Applications of Mean Field Games in Financial Engineering and Economic Theory. Papers 2012.05237, arXiv.org, December 2020. URL `https://ideas.repec.org/p/arx/papers/2012.05237.html`.

[39] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. I*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018. ISBN 978-3-319-56437-1; 978-3-319-58920-6. Mean field FBSDEs, control, and games.

[40] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. II*, volume 84 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018. ISBN 978-3-319-56435-7; 978-3-319-56436-4. Mean field games with common noise and master equations.

[41] René Carmona, François Delarue, and Daniel Lacker. Mean field games with common noise. *The Annals of Probability*, 44(6):3740–3803, 2016.

[42] René Carmona and Mathieu Laurière. Deep learning for mean field games and mean field control with applications to finance, 2021. URL `https://arxiv.org/abs/2107.04568`.

[43] Anne Cheng. Histoire de la pensee chinoise, 1997.

[44] George Christodoulou, Martin Gairing, Yiannis Giannakopoulos, and Paul G. Spirakis. The price of stability of weighted congestion games, 2018. URL `https://arxiv.org/abs/1802.09952`.

[45] Antoine Cournot. *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. 1838.

[46] Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences USA*, 111:10810–10817, 2014.

[47] Wojciech M Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. Real world games look like spinning tops. *Advances in Neural Information Processing Systems*, 33, 2020.

[48] C. Daskalakis, R. Frongillo, C. Papadimitriou, G. Pierrakos, and G. Valiant. On learning algorithms for Nash equilibria. *Algorithmic Game Theory*, pages 114–125, 2010.

[49] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[50] Nicolas de Tocqueville. De la démocratie en amérique, 1835.

[51] Tom Decroos and Jesse Davis. Player vectors: Characterizing soccer players' playing style from match event streams. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.

[52] Laura Degl'Innocenti. *Correlated Equilibria in Static Mean-Field Games*. PhD thesis, Università degli Studi di Padova, 2018.

[53] Niccolo dei Machiavelli. Il principe, 1532.

[54] Francois Delarue. Restoring uniqueness to mean-field games by randomizing the equilibria. *Stochastics and Partial Differential Equations: Analysis and Computations*, 7(4):598–678, 2019.

[55] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[56] Boualem Djehiche, Alain Tcheukam Siwe, and Hamidou Tembine. Mean-field-type games in engineering. *AIMS Electronics and Electrical Engineering*, 1, 11 2017. doi: 10.3934/ElectrEng.2017.1.18.

[57] Josu Doncel, Nicolas Gast, and Bruno Gaujal. Mean-Field Games with Explicit Interactions. working paper or preprint, February 2016. URL `https://hal.inria.fr/hal-01277098`.

[58] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[59] Mohammed Elhenawy, Ahmed A. Elbery, Abdallah A. Hassan, and Hesham A. Rakha. An intersection game-theory-based traffic control algorithm in a connected vehicle environment. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 343–347, 2015. doi: 10.1109/ITSC.2015.65.

[60] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[61] Gabriele Farina, Tommaso Bianchi, and Tuomas Sandholm. Coarse correlation in extensive-form games, 2019. URL `https://arxiv.org/abs/1908.09893`.

[62] Gabriele Farina, Chun Kai Ling, Fei Fang, and Tuomas Sandholm. Correlation in extensive-form games: Saddle-point formulation and benchmarks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[63] Abdolvahhab Fetanat, Ehsan Khorasaninejad, and Gholamreza Shafipour. Energy security-based game theoretic approach for strategies selection in climate risk and energy resources management: a case study of iran. *International Journal of Energy and Environmental Engineering*, 12, 06 2021. doi: 10.1007/s40095-021-00400-5.

[64] Peter I. Frazier. A tutorial on bayesian optimization, 2018.

[65] A. Gaunersdorfer and J. Hofbauer. Fictitious play, shapley polygons, and the replicator equation. *Games and Economic Behavior*, 11:279–303, 1995.

[66] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *AAMAS 2022 (arXiv preprint arXiv:2106.03787)*, 2021.

[67] Alvin Goldman. *Theory of Mind*, pages 19–44. 09 2013. ISBN 9780199874187. doi: 10.1093/acprof:osobl/9780199874187.003.0002.

[68] Diogo Gomes, Levon Nurbekyan, and Edgard Pimentel. Economic models and mean-field games theory, 07 2015.

[69] Geoffrey J. Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 360–367, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390202. URL `https://doi.org/10.1145/1390156.1390202`.

[70] Jean-Bastien Grill, Florent Altché, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Rémi Munos. Monte-carlo tree search as regularized policy optimization, 2020. URL https://arxiv.org/abs/2007.12509.

[71] Audrūnas Gruslys, Marc Lanctot, Rémi Munos, Finbarr Timbers, Martin Schmid, Julien Perolat, Dustin Morrill, Vinicius Zambaldi, Jean-Baptiste Lespiau, John Schultz, Mohammad Gheshlaghi Azar, Michael Bowling, and Karl Tuyls. The advantage regret-matching actor-critic, 2020. URL https://arxiv.org/abs/2008.12234.

[72] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. *Mean Field Games and Applications*, pages 205–266. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-14660-2. doi: 10.1007/978-3-642-14660-2_3. URL https://doi.org/10.1007/978-3-642-14660-2_3.

[73] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Proceedings of NeurIPS*, 2019.

[74] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A General Framework for Learning Mean-Field Games. *arXiv e-prints*, art. arXiv:2003.06069, March 2020.

[75] Saeed Hadikhanloo. *Learning in Mean Field Games*. PhD thesis, University Paris-Dauphine, 2018.

[76] Nicholas Ham. Notions of anonymity, fairness and symmetry for finite strategic-form games, 2013. URL https://arxiv.org/abs/1311.4766.

[77] Kenza Hamidouche, Walid Saad, mérouane Debbah, and H. Vincent Poor. Mean-field games for distributed caching in ultra-dense small cell networks. pages 4699–4704, 07 2016. doi: 10.1109/ACC.2016.7526096.

[78] Nikolaus Hansen. The cma evolution strategy: A tutorial, 2016.

[79] Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.2000.2746. URL https://www.sciencedirect.com/science/article/pii/S0022053100927467.

[80] Sergiu Hart and Andreu Mas-Colell. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*, volume 4. World Scientific, 2013.

[81] Sergiu Hart and David Schmeidler. Existence of correlated equilibria. *Mathematics of Operations Research*, 14(1):18–25, 1989. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689835.

[82] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

[83] Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015.

[84] P. Jean-Jacques Herings and Ronald J. A. P. Peeters. A differentiable homotopy to compute nash equilibria of n-person games. *Economic Theory*, 18(1):159–185, 2001. ISSN 09382259, 14320479. URL http://www.jstor.org/stable/25055414.

[85] Bret Hoehn, Finnegan Southey, Robert C. Holte, and Valeriy Bulitko. Effective short-term opponent exploitation in simplified poker.

[86] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3), 2006. ISSN 1526-7555. URL http://projecteuclid.org/euclid.cis/1183728987.

[87] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/24550348.

[88] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

[89] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castañ eda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, may 2019. doi: 10.1126/science.aau6249. URL https://doi.org/10.1126%2Fscience.aau6249.

[90] J.S. Jordan. Three problems in learning mixed-strategy Nash equilibria. *Games and Economic Behavior*, 5(3):368 – 386, 1993. ISSN 0899-8256. doi: http://dx.doi.org/10.1006/game.1993.1022. URL http://www.sciencedirect.com/science/article/pii/S0899825683710225.

[91] Shizuo Kakutani. A generalization of brouwer's fixed point theorem. *Duke mathematical journal*, 8(3):457–459, 1941.

[92] R. Kleinberg, K. Ligett, G. Piliouras, and É. Tardos. Beyond the Nash equilibrium barrier. In *Symposium on Innovations in Computer Science (ICS)*, 2011.

[93] Lucy Klinger, Lei Zhang, and Zhennan Zhou. A mean field game analysis of consensus protocol design, 2021. URL https://arxiv.org/abs/2108.09999.

[94] H. W. Kuhn. A simplified two-person poker. 1:97–103, 1950.

[95] Daniel Lacker. Mean field games via controlled martingale problems: existence of markovian equilibria. *Stochastic Processes and their Applications*, 125(7):2856–2894, 2015.

[96] Daniel Lacker and Luc Le Flem. Closed-loop convergence for mean field games with common noise. *arXiv preprint arXiv:2107.03273*, 2021.

[97] Marc Lanctot. Further developments of extensive-form replicator dynamics using the sequence-form representation. volume 2, 05 2014.

[98] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2017.

[99] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[100] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, 2019.

[101] Marc Lanctot et al. Openspiel: A framework for reinforcement learning in games, 2020.

[102] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[103] Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Élie, Olivier Pietquin, and Matthieu Geist. Scalable deep reinforcement learning algorithms for mean field games, 2022. URL `https://arxiv.org/abs/2203.11973`.

[104] Wonjun Lee, Siting Liu, Hamidou Tembine, Wuchen Li, and Stanley Osher. Controlling propagation of epidemics via mean-field control. *SIAM Journal on Applied Mathematics*, 81 (1):190–207, 2021. doi: 10.1137/20M1342690. URL `https://doi.org/10.1137/20M1342690`.

[105] Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. Emergent coordination through competition. In *International Conference on Learning Representations (ICLR)*, 2019.

[106] Siqi Liu, Guy Lever, Zhe Wang, Josh Merel, S. M. Ali Eslami, Daniel Hennes, Wojciech M. Czarnecki, Yuval Tassa, Shayegan Omidshafiei, Abbas Abdolmaleki, Noah Y. Siegel, Leonard Hasenclever, Luke Marris, Saran Tunyasuvunakool, H. Francis Song, Markus Wulfmeier, Paul Muller, Tuomas Haarnoja, Brendan D. Tracey, Karl Tuyls, Thore Graepel, and Nicolas Heess. From motor control to team play in simulated humanoid football, 2021. URL `https://arxiv.org/abs/2105.12196`.

[107] Chris Lu, Timon Willi, Christian Schroeder de Witt, and Jakob Foerster. Model-free opponent shaping, 2022. URL `https://arxiv.org/abs/2205.01447`.

[108] A. MacIntyre. *After Virtue*. Bloomsbury Revelations. Bloomsbury Publishing, 2013. ISBN 9781623565251. URL `https://books.google.fr/books?id=00rsK2Y98gQC`.

[109] Mohamed Maddouri, Amen Debbiche, Habib Elkhorchani, and Khaled Grayaa. Game theory and hybrid genetic algorithm for energy management and real time pricing in smart grid. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, pages 1–6, 2018. doi: 10.1109/CISTEM.2018.8613383.

[110] Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, and Thore Graepel. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers, 2021. URL `https://arxiv.org/abs/2106.09435`.

[111] Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, and Thore Graepel. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers, 2021. URL `https://arxiv.org/abs/2106.09435`.

[112] Luke Marris, Paul Muller, et al. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers, 2021.

[113] Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

[114] Anis Matoussi, Clémence Alasseur, and Imen Ben Taher. An Extended Mean Field Game for Storage in Smart Grids. 27 pages, 5 figures, March 2018. URL https://hal.archives-ouvertes.fr/hal-01740707.

[115] J. Maynard Smith and G. R. Price. The logic of animal conflicts. *Nature*, 246:15–18, 1973.

[116] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[117] H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *International Conference on Machine Learning (ICML)*, 2003.

[118] Barnabé Monnot and Georgios Piliouras. Limits and limitations of no-regret learning in games. *The Knowledge Engineering Review*, 32, 2017.

[119] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[120] Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R Wright, Amy Greenwald, and Michael Bowling. Hindsight and sequential rationality of correlated play. *arXiv preprint arXiv:2012.05874*, 2020.

[121] Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R Wright, Michael Bowling, and Amy R Greenwald. Efficient deviation types and learning for hindsight rationality in extensive-form games. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7818–7828. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/morrill21a.html.

[122] Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R. Wright, Amy Greenwald, and Michael Bowling. Hindsight and sequential rationality of correlated play. In *Proceedings of the The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[123] Hervé Moulin and J-P Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4): 201–221, 1978.

[124] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning, 2019. URL https://arxiv.org/abs/1909.12823.

[125] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2020.

[126] Paul Muller, Mark Rowland, Romuald Elie, Georgios Piliouras, Julien Perolat, Mathieu Lauriere, Raphael Marinier, Olivier Pietquin, and Karl Tuyls. Learning equilibria in mean-field games: Introducing mean-field psro, 2021.

[127] Paul Muller, Romuald Elie, Mark Rowland, Mathieu Lauriere, Julien Perolat, Sarah Perrin, Matthieu Geist, Georgios Piliouras, Olivier Pietquin, and Karl Tuyls. Learning correlated equilibria in mean-field games, 2022. URL https://arxiv.org/abs/2208.10138.

[128] François Mériaux, Vineeth Varma, and Samson Lasaulce. Mean field energy games in wireless networks. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 671–675, 2012. doi: 10.1109/ACSSC.2012.6489095.

[129] Anna Nagurney and Ding Zhang. *Projected dynamical systems and variational inequalities with applications*, volume 2. Springer Science & Business Media, 2012.

[130] Khac-Hoai Nam Bui and Jason J. Jung. Cooperative game-theoretic approach to traffic flow optimization for multiple intersections. *Computers & Electrical Engineering*, 71:1012–1024, 2018. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2017.10.016. URL `https://www.sciencedirect.com/science/article/pii/S0045790617318050`.

[131] Y. Narahari. *Game Theory and Mechanism Design*. World Scientific Publishing Company Pte. Limited, 2014. ISBN 9789814525046.

[132] J.F. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[133] John F Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[134] John F. Nash. Equilibrium points in ¡i¿n¡/i¿-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950. doi: 10.1073/pnas.36.1.48. URL `https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48`.

[135] Aqdas Naz, Nadeem Javaid, Muhammad Babar Rasheed, Abdul Haseeb, Musaed Alhussein, and Khursheed Aurangzeb. Game theoretical energy management with storage capacity optimization and photo-voltaic cell generated power forecasting in micro grid. *Sustainability*, 11(10), 2019. ISSN 2071-1050. doi: 10.3390/su11102763. URL `https://www.mdpi.com/2071-1050/11/10/2763`.

[136] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[137] Chris Nota and Philip S. Thomas. Is the policy gradient a gradient?, 2019. URL `https://arxiv.org/abs/1906.07073`.

[138] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Rémi Munos. α-Rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9, 2019.

[139] Shayegan Omidshafiei, Karl Tuyls, Wojciech M. Czarnecki, Francisco C. Santos, Mark Rowland, Jerome Connor, Daniel Hennes, Paul Muller, Julien Pérolat, Bart De Vylder, Audrunas Gruslys, and Rémi Munos. Navigating the landscape of multiplayer games. *Nature Communications*, 11(1), nov 2020. doi: 10.1038/s41467-020-19244-4. URL `https://doi.org/10.1038%2Fs41467-020-19244-4`.

[140] Shayegan Omidshafiei, Daniel Hennes, Marta Garnelo, Zhe Wang, Adria Recasens, Eugene Tarassov, Yi Yang, Romuald Elie, Jerome Connor, Paul Muller, Natalie Mackraz, Kris Cao, Pol Moreno, Pablo Sprechmann, Demis Hassabis, Ian Graham, William Spearman, Nicolas Heess, and Karl Tuyls. Multiagent off-screen behavior prediction in football. *Scientific Reports*, 12, 05 2022. doi: 10.1038/s41598-022-12547-0.

[141] Luis E. Ortiz, Robert E. Schapire, and Sham M. Kakade. Maximum entropy correlated equilibria. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 347–354, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[142] George Orwell. 1984.

[143] Georg Ostrovski and Sebastian van Strien. Payoff performance of fictitious play. *arXiv preprint arXiv:1308.4049*, 2013.

[144] Ignacio Palacios-Huerta. Professionals Play Minimax. *The Review of Economic Studies*, 70 (2):395–415, 04 2003. ISSN 0034-6527.

[145] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Neural Information Processing Systems (NIPS)*, 2017.

[146] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization, 2020. URL `https://arxiv.org/abs/2002.08456`.

[147] Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent, 2021.

[148] Julien Perolat, Bart de Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning, 2022. URL `https://arxiv.org/abs/2206.15378`.

[149] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Proc. of NeurIPS*, 2020.

[150] Jan Peters and J. Andrew Bagnell. *Policy Gradient Methods*, pages 774–776. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_640. URL `https://doi.org/10.1007/978-0-387-30164-8_640`.

[151] Viktoriya Petrakova and Olga Krivorotko. Mean field game for modeling of covid-19 spread. *Journal of Mathematical Analysis and Applications*, 514(1):126271, 2022. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2022.126271. URL `https://www.sciencedirect.com/science/article/pii/S0022247X22002852`.

[152] Georgios Piliouras, Mark Rowland, Shayegan Omidshafiei, Romuald Elie, Daniel Hennes, Jerome Connor, and Karl Tuyls. Evolutionary dynamics and ϕ-regret minimization in games, 2021. URL `https://arxiv.org/abs/2106.14668`.

[153] Plato. Euthyphro, 400BC.

[154] Plato. The republic, 400BC.

[155] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951.

[156] Julia Jean Robinson. An iterative method of solving a game. *Classics in Game Theory*, 2020.

[157] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015.

[158] Tim Roughgarden. *Twenty lectures on algorithmic game theory*. Cambridge University Press, 2016.

[159] Mark Rowland, Shayegan Omidshafiei, Daniel Hennes, Will Dabney, Andrew Jaegle, Paul Muller, Julien Pérolat, and Karl Tuyls. Temporal difference and return optimism in cooperative multi-agent reinforcement learning. 2021.

[160] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002. doi: 10.1073/pnas.032086299. URL http://www.pnas.org/content/99/7/4748.abstract.

[161] Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Player of games, 2021. URL https://arxiv.org/abs/2112.03178.

[162] Peter Schuster and Karl Sigmund. Replicator dynamics. *Journal of Theoretical Biology*, 100 (3):533–538, 1983.

[163] Aner Sela. Fictitious play in 'one-against-all'multi-player games. *Economic Theory*, 14(3): 635–651, 1999.

[164] Lloyd Shapley. Some topics in two-person games. *Advances in game theory*, 52:1–29, 1964.

[165] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

[166] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL https://arxiv.org/abs/1712.01815.

[167] Samuel Sokota, Edward Lockhart, Finbarr Timbers, Elnaz Davoodi, Ryan D'Orazio, Neil Burch, Martin Schmid, Michael Bowling, and Marc Lanctot. Solving common-payoff games with approximate policy iteration, 2021.

[168] Francisco J Solis and Roger J-B Wets. Minimization by random search techniques. *Mathematics of operations research*, 6(1):19–30, 1981.

[169] Finnegan Southey, B. Hoehn, and Robert Holte. Effective short-term opponent exploitation in simplified poker. *Machine Learning*, 74:159–189, 02 2009.

[170] Eric Steinberger. Single deep counterfactual regret minimization, 2019. URL https://arxiv.org/abs/1901.07621.

[171] Leo Strauss. The rebirth of classical political rationalism : an introduction to the thought of leo strauss : essays and lectures.

[172] Stefan Szymanski. Sport analytics: Science or alchemy? *Kinesiology Review*, 9(1):57–63, 2020.

[173] Akash Talwariya, Pushpendra Singh, and Mohan Lal Kolhe. Stackelberg game theory based energy management systems in the presence of renewable energy sources. *IETE Journal of Research*, 67(5):611–619, 2021. doi: 10.1080/03772063.2020.1869109. URL https://doi.org/10.1080/03772063.2020.1869109.

[174] Oskari Tammelin. Solving large imperfect information games using cfr+, 2014.

[175] Takashi Tanaka, Ehsan Nekouei, and Karl Henrik Johansson. Linearly solvable mean-field road traffic games. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 283–289, 2018. doi: 10.1109/ALLERTON.2018.8636077.

[176] Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

[177] Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steele, Pauline Luc, Adria Recasens, Alexandre Galashov, Gregory Thornton, Romuald Elie, Pablo Sprechmann, Pol Moreno, Kris Cao, Marta Garnelo, Praneet Dutta, Michal Valko, Nicolas Heess, Alex Bridgland, Julien Perolat, Bart De Vylder, Ali Eslami, Mark Rowland, Andrew Jaegle, Remi Munos, Trevor Back, Razia Ahamed, Simon Bouton, Nathalie Beauguerlange, Jackson Broshear, Thore Graepel, and Demis Hassabis. Game plan: What ai can do for football, and what football can do for ai, 2020. URL `https://arxiv.org/abs/2011.09192`.

[178] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015. URL `https://arxiv.org/abs/1509.06461`.

[179] Oriol Vinyals, Igor Babuschkin, Wojciech Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John Agapiou, Max Jaderberg, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575, 11 2019.

[180] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, and et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019. doi: 10.1038/s41586-019-1724-z.

[181] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, pages 1–5, 2019.

[182] Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2012.07.003. URL `https://www.sciencedirect.com/science/article/pii/S0022053113000112`.

[183] Christian Graf Von Krockow. *Die Deutschen in ihrem Jahrhundert 1890-1990*. 1990.

[184] John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–300, 1928.

[185] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[186] Huisheng Wang, Yuejiang Li, and H. Vicky Zhao. Research on intelligent traffic control methods at intersections based on game theory, 2020. URL `https://arxiv.org/abs/2009.05216`.

[187] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

[188] Jinyi Xu, Chengchu Yan, Yizhe Xu, Jingfeng Shi, Kai Sheng, and Xiaoping Xu. A hierarchical game theory based demand optimization method for grid-interaction of energy flexible buildings. *Frontiers in Energy Research*, 9, 2021. ISSN 2296-598X. doi: 10.3389/fenrg.2021.736439. URL `https://www.frontiersin.org/articles/10.3389/fenrg.2021.736439`.

[189] H. Peyton Young. The evolution of conventions. *Econometrica*, 61:57–84, 1993.

[190] Qiao Zhang and Gang Li. A game theory energy management strategy for a fuel cell/battery hybrid energy storage system. *Mathematical Problems in Engineering*, 2019:1–12, 01 2019. doi: 10.1155/2019/7860214.

Thesis Summary

This thesis addresses the question of computing game-theoretic equilibria in N-player games, and focuses particularly on the question of computing equilibria in N-player games when N is tremendously large.

The thesis' body starts with methods to converge to three different types of equilibria in N-player games: correlated equilibria, coarse-correlated equilibria, and α-Rank. All three equilibria are converged-to using an alteration of Policy Space Response Oracle (PSRO), a popular population-based algorithm which computes a number of different policies and finds the optimal way to mix them in order to converge. More specifically, this alteration uses the target equilibrium and an innovative new-policy-computing algorithm to reach said equilibrium. We prove the convergence of our method to these equilibria of interest, and enlarge it to a broader class of equilibria which we define.

This answers the initial thesis question regarding converging towards any equilibrium in any *finite* N-player game. However, these PSRO-derived approaches are heavily dependent on the number of players in their game: the more players there are, the more difficult it becomes for them to find an equilibrium, and this difficulty quickly becomes prohibitive.

The second part of the thesis is therefore concerned with overcoming this difficulty when the number of agents is extremely large, by considering that their number is *infinite*. Paradoxically, this approximation simplifies equilibrium computation by eliminating combinatorial effects. We first analyze what becomes of correlated and coarse-correlated equilibria in Mean-Field games, derive their new expressions, properties, and their behavior when they are reused in N-player games. Under suitable conditions, reusing a Mean-Field (coarse-) correlated equilibrium in an N-player game yields an $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-approximate (coarse-) correlated equilibrium.

We then address the question of computing Mean-Field (coarse-) correlated equilibria. We show that two popular algorithms converge towards Mean-Field coarse-correlated equilibria, but in a spatially-complex way, via the notion of Mean-Field regret minimization. We introduce another variant of PSRO, Mean-Field PSRO, capable of converging towards correlated, coarse-correlated and Nash equilibria in all Mean-Field games of our framework. This is done by the use of black-box optimizers for Nash equilibria, and of no-adversarial-regret algorithms for (coarse-) correlated equilibria. These equilibria are also simplified by the introduction of a new compression method, bandit compression.

Finally, the thesis ends with an application of Game-Theoretical equilibria in a real-world situation: soccer penalty kicks. The game-theoretic analysis serves the purpose of analyzing how optimal the behavior of players is, characterizing each player's behavioral tendencies, and providing strategic suggestions to improve penalty kick outcomes.

Résumé de la Thèse

Cette thèse traite la question du calcul et de l'estimation d'équilibres de théorie des jeux dans des jeux à N-joueurs. Elle se concentre en particulier sur les jeux N-joueurs où N est extrêmement large

Le corps de cette thèse commence par décrire des méthodes permettant de converger vers trois types d'équilibres : corrélés, faiblement-corrélés (*coarse-correlated*), et α-Rank. Ces trois équilibres sont atteints via une altération de PSRO, un algorithme basé sur une population, c'est-à-dire qui calcule différentes stratégies et une manière optimale de les combiner. Plus spécifiquement, cette altération utilise l'équilibre recherché et un nouveau type d'algorithme calculant une nouvelle stratégie pour atteindre l'équilibre mentionné. Nous prouvons que notre méthode converge vers les équilibres que nous examinons, et élargissons ce résultat à une plus large classe d'équilibres que nous définissons.

Ces développements apportent une réponse à la question initiale de la thèse portant sur la convergence vers tout équilibre de théorie des jeux dans tout jeu *fini* à N-joueurs. Cependant, les méthodes dérivées de PSRO mentionnées plus haut peinent à converger rapidement lorsque N est élevé. Pour des valeurs de N très élevées, il devient presqu'impossible de trouver des équilibres en un temps raisonnable.

La seconde partie de cette thèse porte donc sur la question de contourner la complexité provenant du nombre d'agents, en considérant que leur nombre est en fait *infini*. Paradoxalement, cette approximation simplifie le calcul d'équilibres parce qu'elle élimine tout effet combinatoire provenant des N joueurs. Nous analysons d'abord ce que deviennent les équilibres (faiblement-) corrélés sous l'approximation des jeux à Champ Moyen (Jeux avec une infinité de joueurs), décrivons leur nouvelle expression, leurs propriétés, et leur comportement lorsqu'ils sont réutilisés dans un jeu à N-joueurs. Etant données des conditions raisonnables, réutiliser un équilibre à Champ Moyen dans un jeu à N-joueurs produit un équilibre (faiblement-) corrélé $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-approximatif.

La thèse aborde ensuite le sujet de calculer des équilibres (faiblement-) corrélés à Champ Moyen. Elle montre que deux algorithmes populaires convergent vers des équilibres faiblement-corrélés à Champ Moyen, d'une façon inefficace spatialement, via la notion de minimisation de regret à Champ Moyen. Nous définissons ensuite une nouvelle variante de PSRO, PSRO-à-Champ-Moyen, capable de converger vers des équilibres corrélés, faiblement corrélés et de Nash dans *tout jeu* à Champ Moyen conforme à notre formulation. Ce résultat est obtenu via l'utilisation d'optimiseurs boîte-noire pour le Nash; et d'algorithmes sans-regret-adversarial pour les équilibres corrélés et faiblement corrélés. Ces équilibres sont aussi simplifiés via l'utilisation d'un nouvel algorithme de compression, "compression de bandits".

Enfin, la thèse est conclue par une application d'équilibres de théorie des jeux dans une situation réelle : les tirs au but, lors de matchs de balle-aux-pieds[1]. L'analyse de théorie des jeux sert à analyser l'optimalité des stratégies adoptées par les joueurs, à caractériser les tendances comportementales de chaque joueur, et à leur faire des suggestions afin qu'ils puissent améliorer leurs comportements lors de tirs-aux-but.

[1]Nom qui devrait être popularisé pour ce sport populaire, arrivé en Angleterre grâce à la France lors du Camp du Drap d'Or.