Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers

Luke Marris [1 2] **Paul Muller** [1 3] **Marc Lanctot** [1] **Karl Tuyls** [1] **Thore Graepel** [1 2]

Abstract

Two-player, constant-sum games are well studied in the literature, but there has been limited progress outside of this setting. We propose Joint Policy-Space Response Oracles (JPSRO), an algorithm for training agents in n-player, general-sum extensive form games, which provably converges to an equilibrium. We further suggest correlated equilibria (CE) as promising meta-solvers, and propose a novel solution concept Maximum Gini Correlated Equilibrium (MGCE), a principled and computationally efficient family of solutions for solving the correlated equilibrium selection problem. We conduct several experiments using CE meta-solvers for JPSRO and demonstrate convergence on n-player, general-sum games.

1. Introduction

Recent success in tackling two-player, constant-sum games (Silver et al., 2016; Vinyals et al., 2019) has outpaced progress in n-player, general-sum games despite a lot of interest (Jaderberg et al., 2019; OpenAI et al., 2019; Brown & Sandholm, 2019; Lockhart et al., 2020; Gray et al., 2020; Anthony et al., 2020). One reason is because Nash equilibrium (NE) (Nash, 1951) is tractable and interchangeable in the two-player, constant-sum setting but becomes intractable (Daskalakis et al., 2009) and potentially non-interchangeable[1] in n-player and general-sum settings. The problem of selecting from multiple solutions is known as the equilibrium selection problem (Goldberg et al., 2013; Avis et al., 2010; Harsanyi & Selten, 1988).[2]

Outside of normal form (NF) games, this problem setting arises in multi-agent training when dealing with empirical games (also called meta-games), where a game payoff tensor is populated with expected outcomes between agents playing an extensive form (EF) game, for example the StarCraft League (Vinyals et al., 2019) and Policy-Space Response Oracles (PSRO) (Lanctot et al., 2017), a recent variant of which reached state-of-the-art results in Stratego Barrage (McAleer et al., 2020).

In this work we propose using correlated equilibrium (CE) (Aumann, 1974) and coarse correlated equilibrium (CCE) as a suitable target equilibrium space for n-player, general-sum games[3]. The (C)CE solution concept has two main benefits over NE; firstly, it provides a mechanism for players to correlate their actions to arrive at mutually higher payoffs and secondly, it is computationally tractable to compute solutions for n-player, general-sum games (Daskalakis et al., 2009). We provide a tractable approach to select from the space of (C)CEs (MG), and a novel training framework that converges to this solution (JPSRO). The result is a set of tools for theoretically solving any complete information[4] multi-agent problem. These tools are amenable to scaling approaches; including utilizing reinforcement learning, function approximation, and online solution solvers, however we leave this to future work.

In Section 2 we provide background on a) correlated equilibrium (CE), an important generalization of NE, b) coarse correlated equilibrium (CCE) (Moulin & Vial, 1978), a similar solution concept, and c) PSRO, a powerful multi-agent training algorithm. In Section 3 we propose novel solution concepts called Maximum Gini (Coarse) Correlated Equilibrium (MG(C)CE) and in Section 4 we thoroughly explore its properties including tractability, scalability, invariance, and

[1]DeepMind [2]University College London [3]Université Gustave Eiffel. Correspondence to: Luke Marris <marris@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] That is, there are no longer any guarantees on the expected utility when each player plays their part of some equilibrium; guarantees only hold when all players play *the same* equilibrium. Since players cannot guarantee what others choose, they cannot optimize independently, so the Nash equilibrium loses its appeal as a prescriptive solution concept.

[2] The equilibrium selection problem is subtle and can have various interpretations. We describe it fully in Section 4.1 based on the classical understanding from (Harsanyi & Selten, 1988).

[3] We mean games (also called environments) in a very general sense: extensive form games, multi-agent MDPs and POMDPs (stochastic games), imperfect information games, are all solvable with this approach.

[4] Payoffs for all players are required for the correlation device.

a parameterized family of solutions. In Section 5 we propose a novel training algorithm, Joint Policy-Space Response Oracles (JPSRO), to train policies on n-player, general-sum extensive form games. JPSRO requires the solution of a meta-game, and we propose using MG(C)CE as a meta-solver. We prove that the resulting algorithm converges to a normal form (C)CE in the extensive form game. In Section 6 we conduct an empirical study and show convergence rates and social welfare across a variety of games including n-player, general-sum, and common-payoff games.

An important area of related work is α-Rank (Omidshafiei et al., 2019) which also aims to provide a tractable alternative solution in normal form games. It gives similar solutions to NE in the two-player, constant-sum setting, however it is not directly related to NE or (C)CE. α-Rank has also been applied to ranking agents and as a meta-solver for PSRO (Muller et al., 2020). MG(C)CE is inspired by Maximum Entropy Correlated Equilibria (MECE) (Ortiz et al., 2007), an entropy maximizing CE based on Shannon's entropy that is harder to compute than Gini impurity.

Another important area of related work concerns optimization based approaches (von Stengel & Forges, 2008; Dudik & Gordon, 2012; Farina et al., 2019a) and no-regret approaches (Celli et al., 2019; 2020; Morrill et al., 2021). These approaches identify specific subsets or supersets of (C)CE in the extensive-form game by constructing constraint programs or by local regret minimization using the full representation of the information state space. In contrast, the oracle approach can iteratively identify meta-games with smaller support that summarize the strategic complexity of the game compactly.

2. Preliminaries

This section introduces correlated equilibrium and the multi-agent training algorithm PSRO.

2.1. Correlated Equilibrium

Each player, p, in a game has a set of actions $a_p \in \mathcal{A}_p$ (also known as pure strategies) available to it. Let n be the number of players in a game. Let $\mathcal{A} = \otimes_p \mathcal{A}_p$ be the joint action space and $a = (a_1, ..., a_n) \in \mathcal{A}$ be a joint action.

Let us index quantities relating to all players apart from player p as $-p = \{1, ..., p-1, p+1, ..., n\}$. Let $\sigma(a) = \sigma(a_p, a_{-p})$ be the probability that joint action $a \in \mathcal{A}$ is played in a game. Let $\sigma(a_p) = \sum_{a_{-p} \in \mathcal{A}_{-p}} \sigma(a_p, a_{-p})$ be the marginal probability of player p taking action $a_p \in \mathcal{A}_p$. Let $\sigma(a_{-p}|a_p)$ be the conditional probability that other players play $a_{-p} \in \mathcal{A}_{-p}$, when player p plays $a_p \in \mathcal{A}_p$. σ without arguments should be interpreted as a vector of size $[|\mathcal{A}|]$.

Let $G_p : \mathcal{A} \rightarrow \mathbb{R}$ be the payoff function for player p when players play the joint action $a \in \mathcal{A}$. The full game payoff G can therefore be defined by a tensor of shape $[n, |\mathcal{A}_1|, ..., |\mathcal{A}_n|]$. A normal form game is defined by the tuple $\mathcal{G} = (G, \mathcal{A})$.

Define $A_p(a'_p, a_p, a_{-p}) = G(a'_p, a_{-p}) - G(a_p, a_{-p})$ as the advantage of player p switching action from a_p to a'_p, when other players play a_{-p}. This can be represented as a matrix, A_p, of shape $[|\mathcal{A}_p|(|\mathcal{A}_p| - 1), |\mathcal{A}|]$, since we do not need to compare an action with itself. The matrix is sparse and a fraction of $\frac{1}{\mathcal{A}_p}$ elements are non-zero. We use A, with shape $[\sum_p |\mathcal{A}_p|(|\mathcal{A}_p| - 1), |\mathcal{A}|]$, to denote the concatenation of A_p into a two-dimensional matrix.

A correlated equilibrium (CE), is a joint mixed strategy $P(a)$ such that no player p has payoff to gain from unilaterally choosing to play another action a'_p instead of a_p. An approximate correlated equilibrium (ϵ-CE)[5] is one where that gain from switching actions is no more than ϵ. When $\epsilon = 0$, the standard CE is recovered. This relationship is described mathematically in Equation 1, $\forall p \in \mathcal{P}, a'_p \neq a_p \in \mathcal{A}_p$. In matrix form, we can simply write $A\sigma \leq \epsilon$.

$$\sum_{a_{-p}} \sigma(a_{-p}, a_p) A_p(a'_p, a_p, a_{-p}) \leq \epsilon \tag{1}$$

Together, A_p and ϵ represent the CE linear inequality constraints. Mathematically these are equations of a plane, and separate the joint action space $\sigma(a)$ into half-spaces. Together these half-spaces intersect to form a convex polytope of valid CE solutions.

Of special interest are valid CEs that can factorize into their marginals $\sigma(a) = \prod_p \sigma(a_p)$, and correspond to NE solutions. All NEs are also CEs. Since an NE always exists (when there are finite players and actions) (Nash, 1951), a CE always exists. An NE is always on the boundary of the polytope for non-trivial games (Nau et al., 2004). Any convex combination of CEs is also a CE.

CEs provide a richer set of solutions than NEs. The maximum sum of social welfare in CEs is at least that of any NE. In particular, this allows more intuitive solutions to anti-coordination games such as chicken and traffic lights. Consider the traffic lights example; a symmetric, general-sum, two-player game consisting of two actions *go*, (G), and *wait*, (W). (G, G) results in a crash, in (W, W) no progress is made, and (G, W) and (W, G) result in progress for one of the players. Figure 1 shows the NE and CE solution space for the traffic lights game. The mixed NE solution $(G, W) = (\frac{1}{11}, \frac{10}{11})$ is clearly unsatisfactory ($\frac{1}{121}$ crashing and $\frac{100}{121}$ waiting). One could argue that the best solution is to have players flip a coin to decide who waits and who

[5] There are two competing definitions for approximate CE. We use the computationally convenient one (Section A.2).

goes. It turns out that this solution is a valid CE and is in fact the unique solution of $\min \epsilon$-MGCE, a novel solution concept that we introduce later in Section 4.3.

Correlation is achieved via a trusted external entity (correlation device) which samples a joint action from a public CE joint distribution. Each player is given their action in secret. The properties of the CE means that no individual player is motivated to deviate from the suggested action. If there are deviation actions with equal payoff available, the distribution is a weak equilibrium. If instead, the suggested actions are better than alternatives, the distribution is a strict equilibrium. Distributions that produce actions that are not better than all alternatives are called approximate equilibrium, and the maximum gain that can be obtained by an agent unilaterally deviating from any suggested action is described by $\epsilon > 0$. Weak and strict equilibrium have associated gain $\epsilon = 0$ and $\epsilon < 0$, respectively. Mathematically, the effect of reducing ϵ is shrinking the volume of the CE polytope. We can choose ϵ when solving for a CE and show in Section 4.3 how ϵ can be used to parameterise a family of MGCE solutions.

There are two important solution concepts in the space of CEs. The first is Maximum Welfare Correlated Equilibrium (MWCE) which is defined as the CE that maximises the sum of all player's payoffs. An MWCE can be obtained by solving a linear program, however the MWCE may not be unique and therefore does not fully solve the equilibrium selection problem (e.g. constant-sum game solutions all have equal payoff). The second such concept is Maximum Entropy Correlated Equilibrium (MECE) (Ortiz et al., 2007) which maximises Shannon's entropy (Shannon, 1948) as an objective. MECE also shares some interesting properties with MGCE such as computational scalability when the solution is full-support (positive probability mass everywhere). Drawbacks of this approach are that the literature does not provide algorithms when the solution is general-support (non-negative probability) and, maximising Shannon's entropy can be complex.

Finally, coarse correlated equilibrium (CCE) (Moulin & Vial, 1978) (Section A.3) is a simpler solution concept that contains CE as a subset: NE $\subseteq$ CE $\subseteq$ CCE. Intuitively, a game distribution is in CCE if no player wishes to deviate *before* receiving a recommended signal.

The solution concepts discussed so far apply to normal form (NF) games, and therefore are sometimes prefixed as such in the literature (NFCE and NFCCE) to disambiguate them from their extensive form (EF) counterparts (EFCE (von Stengel & Forges, 2008) and EFCCE (Farina et al., 2019a)). This distinction is important because although EF solutions are a natural choice in EF games; NF solutions can also be applied in EF games by using whole policies $\pi_p \in \Pi_p$ in place of actions $a_p \in \mathcal{A}_p$. These solutions are subsets of one another; NFCE $\subseteq$ EFCE $\subseteq$ EFCCE $\subseteq$ NFCCE (von Stengel & Forges, 2008), therefore NFCE is the most restrictive correlation device while NFCCE is the least restrictive and is therefore capable of achieving the highest welfare. The best correlation device to use is a matter of debate in the literature. However, we note that NF solutions are interesting in EF games because a) it permits the highest welfare, and b) only requires communicating recommendations once before the game starts (as opposed to EF(C)CEs which require communication at every timestep). (J)PSRO trains sets of policies and converges to an NF equilibrium. Therefore, all equilibria discussed in this work are NF and we do not use a prefix going forward. It is possible to extend PSRO to EF equilibria (McAleer et al., 2021).

2.2. Policy-Space Response Oracles (PSRO)

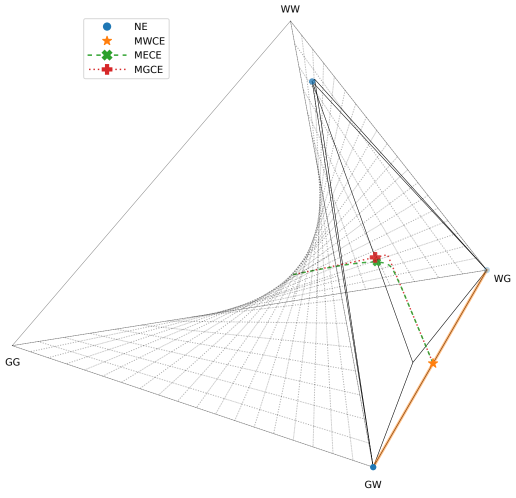
Policy-Space Response Oracles (PSRO) (Lanctot et al., 2017) (Algorithm 1) is an iterative population based training method for multi-agent learning that generalizes other well known algorithms such as fictitious play (FP) (Brown, 1951), fictitious self play (FSP) (Heinrich et al., 2015) and double oracle (DO) (McMahan et al., 2003).

PSRO finds a set of policies, $(\pi_p \in \Pi_p)_{p=1..n}$, and a distribution over this set for each player, $(\sigma_p)_{p=1..n}$. The distribution converges to an NE in two-player, zero-sum games, and has recently been extended to convergence to other types of equilibria (Muller et al., 2020; McAleer et al., 2021). This work is in line with these developments, studying convergence of a variant of PSRO with joint policy distributions and (C)CE meta-solvers in n-player, general-sum games.

PSRO consists of a response oracle that estimates the best response (BR) to a joint distribution of policies. Commonly the response oracle is either a reinforcement learning (RL) agent or a method that computes the exact BR. The component that determines the distribution of policies that the oracle responds to is called the meta-solver (MS). The MS operates on the meta-game (MG), which is a payoff tensor estimated by measuring the expected return (ER) of policies against one another. This is a NF game, but instead of strategies corresponding to actions, a, they correspond to policies, π. The set of deterministic policies can be huge and that of stochastic policies is infinite, therefore PSRO only considers a subset of game policies: the ones found by the BR over all iterations so far. Different MSs result in different algorithms: the uniform distribution results in FSP, and using the NE distribution results in an extension of DO.

3. MG(C)CE and its Computation

The set of (C)CEs forms a convex polytope, and therefore any strictly convex function could uniquely select amongst



(a) CE Convex Polytope

	G	W
G	$-10, -10$	$1, 0$
W	$0, 1$	$0, 0$

(b) Payoff Table

	G	W
G	0.033	0.334
W	0.334	0.299

(c) MECE
$(0, 0)$

	G	W
G	$\frac{1}{121}$	$\frac{10}{121}$
W	$\frac{10}{121}$	$\frac{100}{121}$

(d) Mixed NE
$(0, 0)$

	G	W
G	0	1
W	0	0

(e) Row NE
$(1, 0)$

	G	W
G	0	0
W	1	0

(f) Col NE
$(0, 1)$

	G	W
G	0.033	0.327
W	0.327	0.313

(g) MGCE
$(0, 0)$

	G	W
G	0	0.34
W	0.34	0.32

(h) F-MGCE
$(0.34, 0.34)$

	G	W
G	0	0.5
W	0.5	0

(i) MWCE
$(0.5, 0.5)$

Figure 1. The solution landscape for the traffic lights game. The solid polytope shows the space of CE joint strategies, and the dotted surface shows factorizable joint strategies. NEs are where the surface and polytope intersect. There are three unsatisfying NEs: mixed spends most of its time waiting and does not avoid crashing, the others favour only the row or column player. One MWCE provides a better solution (note that Row NE and Col NE, and any mixture of the two are also MWCE solutions). The center of the tetrahedron is the uniform distribution and the MECE and MGCE attempt to be near this point. The dashed lines correspond to the family of solutions permitted by MGCE and MECE when varying the approximation parameter ϵ. Both have $(GW, WG) = (0.5, 0.5)$ as the $\min \epsilon$ solution. Player payoffs are given in parenthesis.

this set. The literature only provides one such example: MECE (Ortiz et al., 2007) which has a number of appealing properties, but was found to be slow to solve large games. There is a gap in the literature for a more tractable approach, and propose to use the Gini impurity (GI) (Breiman et al., 1984; Bishop, 2006). GI is a member of Tsallis entropy family, a generalized entropy that is equivalent to GI under a certain parameterization. It is maximized when the probability mass function is uniform $\sigma = \frac{1}{|\mathcal{A}|}$ and minimized when all mass is on a single outcome. GI is popular in decision tree classification algorithms because it is easy to compute (Breiman et al., 1984). We call the resulting solution concept maximum Gini (coarse) correlated equilibrium (MG(C)CE). This approach has connections to maximum margin (Cortes & Vapnik, 1995) and maximum entropy (Jaynes, 1957). The derivations (Section C.2) follow standard optimization theory.

3.1. Quadratic Program

The Gini impurity is defined as $1 - \sigma^T \sigma$, and the MG(C)CE is denoted σ^*. We use an equivalent standard form objective $-\frac{1}{2}\sigma^T \sigma$. The most basic form of the problem can be expressed directly as a quadratic program (QP), consisting of a quadratic objective function (Equation 2) and linear constraints (Equations 3 and 4).

$$\text{Gini objective:} \quad \max_{\sigma} -\frac{1}{2}\sigma^T \sigma \quad \text{s.t.} \tag{2}$$

$$\text{(C)CE constraints:} \quad A_p \sigma \leq \epsilon \quad \forall p \tag{3}$$

$$\text{Probability constraints:} \quad \sigma \geq 0 \quad e^T \sigma = 1 \tag{4}$$

QPs are a well studied problem class and many techniques may be used to solve them, including convex and quadratic optimization software, such as CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018) and OSQP (Stellato et al., 2020).

3.2. Primal and Dual Forms

The primal objective that we wish to optimize is $\max_\sigma \min_{\alpha,\beta,\lambda} L(\sigma, \alpha, \beta, \lambda) = L^\sigma_{\alpha,\beta,\lambda}$, where $L^\sigma_{\alpha,\beta,\lambda}$ is the primal Lagrangian function, $\alpha_p \geq 0$ are the dual variable vectors corresponding to the $\epsilon-$(C)CE inequality constraints (Equation 3), $\beta \geq 0$ is the dual variable vector corresponding to the distribution inequality constraints (Equation 4), and λ is the dual variable corresponding to the distribution equality constraint (Equation 4). By augmenting the dual variables $\alpha = [\alpha_1, ..., \alpha_n]$ and constraints matrix $A = [A_1, ..., A_n]$, we can write the primal objective compactly as:

$$L^\sigma_{\alpha_p,\beta,\lambda} = -\frac{1}{2}\sigma^T \sigma - \alpha^T(A\sigma - \epsilon) - \beta^T \sigma + \lambda(e^T \sigma - 1), \tag{5}$$

where the constant vector of ones with appropriate size is denoted by e, and ϵ is a vector populated with the approximation parameter. We can also formulate a simplified dual version of the optimization as:

$$L_{\alpha,\beta} = -\frac{1}{2}\alpha^T ACA^T\alpha - b^T A^T\alpha + \epsilon^T\alpha - \frac{1}{2}\beta^T C\beta \\ - b^T\beta - \alpha^T AC\beta + \frac{1}{2}b^T b, \qquad (6)$$

where $C = I - eb^T$ normalizes by the mean, and $b = \frac{1}{|\mathcal{A}|}e$ is the uniform vector. The optimal primal solution σ^* can be recovered from the optimal dual variables α_p^* and β_p^* using

$$\sigma^* = CA^T\alpha^* + C\beta^* + b. \qquad (7)$$

The full-support assumption states that all joint probabilities have some positive mass, $\sigma > 0$. In this scenario, the dual variable vector corresponding to the non-negative probability constraint is zero, $\beta = 0$. Therefore we can define simplified primal and dual objectives.

$$L^\sigma_{\alpha,\lambda} = -\frac{1}{2}\sigma^T\sigma - \alpha^T(A\sigma - \epsilon) + \lambda(e^T\sigma - 1) \qquad (8)$$

$$L_\alpha = -\frac{1}{2}\alpha^T ACA^T\alpha - b^T A^T\alpha + \epsilon^T\alpha + \frac{1}{2}b^T b \quad (9)$$

$$\sigma^* = CA^T\alpha^* + b \qquad (10)$$

4. Properties of MG(C)CE

In this section we discuss some of the properties of ϵ-MG(C)CE[6]. Section C contains the proofs for this section.

4.1. Equilibrium Selection Problem

There are two levels of coordination; first is selecting an equilibrium before play commences, and second is selecting actions during play time. Both NEs and (C)CEs require agreement on what equilibrium is being played (Goldberg et al., 2013; Avis et al., 2010; Harsanyi & Selten, 1988): for (C)CEs this is a joint action probability distribution, and for NEs this is also a joint action probability distribution that can conveniently be factored into stochastic strategies for each player. Therefore, at this level of coordination, both NEs and (C)CEs are similar. We refer to this coordination problem as the *equilibrium selection problem* (Harsanyi & Selten, 1988). At action selection time only (C)CEs require further coordination. NEs are factorizable and therefore can sample independently without further coordination. (C)CEs rely on a central correlation device that will recommend actions from the equilibrium that was previously agreed upon.

This means that neither NEs nor (C)CEs can be directly used prescriptively in n-player, general-sum games. These solution concepts specify what subsets of joint strategies are in equilibrium, but does not specify how decentralized agents should select amongst these. Furthermore, the presence of a correlation device does not make (C)CEs prescriptive because the agents still need a mechanism to agree on the distribution the correlation device samples from[7]. This coordination problem can be cast as one that is more computational in nature: what rules allow an equilibrium to be uniquely (and perhaps de-centrally) selected?

This highlights the main drawback of MW(C)CE which does not select for unique solutions (for example, in constant-sum games all solutions have maximum welfare). One selection criterion for NEs is maximum entropy Nash equilibrium (MENE) (Balduzzi et al., 2018), however outside of the two-player constant-sum setting, these are generally not easy to compute (Daskalakis et al., 2009). CEs exist in a convex polytope, so any convex function can select among them. Maximum entropy correlated equilibrium (MECE) (Ortiz et al., 2007) is limited to full-support solutions, which may not exist when $\epsilon = 0$, and can be hard to solve in practice. Therefore, there is gap in the literature for a computationally tractable, unique, solution concept and this work proposes MG(C)CE fills this gap.

Theorem 1 (Uniqueness and Existence)**.** *MG(C)CE provides a unique solution to the equilibrium solution problem and always exists.*

4.2. Scalable Representation

MG(C)CE can provide solutions in general-support and, similar to MECE, MG(C)CE permits a scalable representation when the solution is full-support. Under this scenario, the distribution inequality constraint variables, β, are inactive, are equal to zero, can be dropped, and the α variables can fully parameterize the solution.

Theorem 2 (Scalable Representation)**.** *The MG(C)CE, σ^*, has the following forms:*

$$\textit{General Support:} \quad \sigma^* = CA^T\alpha^* + C\beta^* + b \qquad (11)$$

$$\textit{Full Support:} \quad \sigma^* = CA^T\alpha^* + b \qquad (12)$$

Where e is a vector of ones, $|\mathcal{A}| = \prod_p |\mathcal{A}_p|$, $C = I - e^T b$, and $b = \frac{1}{|\mathcal{A}|}e$ are constants. $\alpha^* \geq 0$ and $\beta^* \geq 0$ are the optimal dual variables of the solution, corresponding to the (C)CE and distribution inequality constraints respectively.

Let $|\mathcal{A}_p|$ correspond to the number of actions available to player p, and the total number of joint actions, σ, is $|\mathcal{A}| =$

[6] Some of the properties discussed here also apply to MECE (Ortiz et al., 2007).

[7] This is true if the correlation device is not considered as part of the game. If it was part of the game (for example traffic lights at a junction) the solution concept can appear prescriptive.

$\prod_p |\mathcal{A}_p|$. For each value of σ, there is a corresponding β dual variable. The number of α dual variables is no more than the number of pair permutations $\sum_p |\mathcal{A}_p|(|\mathcal{A}_p|-1)$ for CEs or actions $\sum_p |\mathcal{A}_p|$ for CCEs. Clearly, games with three or more players and many actions, $\sum_p |\mathcal{A}_p|(|\mathcal{A}_p|-1) \ll \prod_p |\mathcal{A}_p|$ for CEs and $\sum_p |\mathcal{A}_p| \ll \prod_p |\mathcal{A}_p|$ for CCEs, allow for a very scalable parameterization if the full-support assumption holds. Furthermore, optimal α^* are sparse so we can discard rows from A, in a similar spirit to SVMs (Cortes & Vapnik, 1995).

For CEs, full-support is not possible when an action is strictly dominated by another. This case can be easily mitigated by iterated elimination of strictly dominated strategies (IESDS) (Fudenberg & Tirole, 1991). This also has the desirable property of simplifying the optimization. In a similar argument, when actions are repeated (having the same payoffs), only one need be retained with appropriate modifications to the optimization.

Among the set of ϵ-MG(C)CE there always exists one with full-support. Note that any infinitesimal positive ϵ will permit a full-support (C)CE, but ϵ-MG(C)CE does not necessarily select these. An upper bound on ϵ which permits a full-support solution is given by Theorem 3.

Theorem 3 (Existence of Full-Support ϵ-MG(C)CE)**.** *For all games, there exists an $\epsilon \leq \max(Ab)$ such that a full-support, ϵ-MG(C)CE exists. A uniform solution, b, always exists when $\max(Ab) \leq \epsilon$. When $\epsilon < \max(Ab)$, the solution is non-uniform.*

4.3. Family of Solutions

ϵ-MG(C)CE provides an intuitive way to control the strictness of the equilibrium via the approximation parameter, ϵ, which parameterizes a family of unique solutions. Positive ϵ expands the solution set and results in a higher Gini impurity solution, at the expense of lower payoff, and approximate equilibrium. Negative ϵ shrinks the solution set to achieve a strict equilibrium and higher payoff at the expense of Gini impurity. This might also be a more robust solution (Wald, 1939; 1945; Ben-Tal et al., 2009) if the payoff is uncertain.

It is worth emphasizing a set of particularly interesting solutions within this family. Firstly the standard MG(C)CE, with $\epsilon = 0$, provides a weak equilibrium for non-trivial games (Theorem 4). Secondly, an edge case with positive ϵ is $\max(Ab)$-ϵ-MG(C)CE which guarantees a uniform distribution solution. Converging to uniform when increasing ϵ is a desirable property (principle of insufficient reason) (Leonard J. Savage, 1954; Sinn, 1980; Jaynes, 1957). Thirdly, note that all $\epsilon < \max(Ab)$ are guaranteed to have a non-uniform distribution (Theorem 3), therefore, a $\frac{1}{2}\max(Ab)$-ϵ-MG(C)CE could be an interesting way to regularise a MGCE towards a uniform distribution. Fourthly, because our algorithms are particularly scalable when full-support, working out the minimum ϵ such that a full-support solution exists, full-ϵ-MG(C)CE, would be useful. Finally, the solution with the smallest feasible ϵ is the $\min \epsilon$-MG(C)CE. This solution has the lowest entropy of the family, but the highest payoff, and constitutes the strictest equilibrium. Refer to Figure 1 for the family of solutions for the traffic lights game.

Theorem 4. *For non-trivial games (Nau et al., 2004), the MG(C)CE lies on the boundary of the polytope and hence is a weak equilibrium.*

Since the ϵ is deterministically known for the $\max(Ab)\epsilon$-MG(C)CE, $\frac{1}{2}\max(Ab)\epsilon$-MG(C)CE and MG(C)CE solutions, we can solve for these using the standard solvers discussed in Section 3. For the $\min \epsilon$-MG(C)CE we can tweak our optimization procedure to solve for this case directly by simply including a $c\epsilon$ term to minimize, where $c > 1$. We use bisection search to find full-ϵ-MG(C)CE.

4.4. Invariance

An important concept in decision theory, called cardinal utility (Mas-Colell et al., 1995), is that offset and positive scale of each player's payoff does not change the properties of the game. A notable solution concept that does not have this property is MW(C)CE.

Theorem 5 (Affine Payoff Transformation Invariance)**.** *If σ^* is the ϵ-MG(C)CE of a game, $\mathcal{G}$, then for each player p independently we can transform the payoff tensors $\tilde{G}_p = c_p G_p + d_p$ and approximation vector $\tilde{\epsilon}_p = a_p \epsilon_p$ for some positive c_p and real d_p scalars, without changing the solution. Furthermore rows of the advantage matrix A, and approximation vector, ϵ, can be scaled independently without changing the MG(C)CE.*

4.5. Computationally Tractable

In general, finding NEs is a hard problem (Daskalakis et al., 2009). While solving for any valid (C)CE is simple (basic feasible solution of a linear constraint problem) (Matouek & Gärtner, 2006), and finding a (C)CE with a linear objective is an LP, solving for a particular (C)CE can be hard. For example, MECE (Ortiz et al., 2007) requires optimizing a constrained nonlinear objective. α-Rank can be solved in cubic time in the number of pure joint strategies, $O(|\mathcal{A}|^3)$.

MG(C)CE, however, is the solution to a quadratic program, and therefore can be solved in polynomial time. Furthermore, if the assumption is made that the solution is full-support, the algorithm's variables scale better than the number of σ parameters.

Space requirements are dominated by the storage of the advantage matrix A, which requires a space of $O(n|\mathcal{A}_p||\mathcal{A}|)$ when exploiting sparsity. Computation is also on the order $O(n|\mathcal{A}_p||\mathcal{A}|)$ for gradient computation, exploiting sparsity.

The number of variables depends on whether we are solving the general-support, $|\mathcal{A}| + n|\mathcal{A}_p|^2$, or full-support, $n|\mathcal{A}_p|^2$ version. It is possible to make use of sparse matrix implementations and only efficient matrix-vector multiplications are required to compute the derivatives.

5. Joint PSRO

JPSRO (Algorithm 2) is a novel extension to Policy-Space Response Oracles (PSRO) (Lanctot et al., 2017) (Algorithm 1) with full mixed joint policies to enable coordination among policies. Although a conceptually straightforward extension, careful attention is needed to a) develop suitable best response (BR) operators, b) develop tractable joint distribution meta-solvers (MS), c) evaluate the set of policies found so far, and d) develop convergence proofs.

Using notation of Section 2.1, but policies instead of actions. Let $(\Pi^*_p)_{p=1..n}$ be the set of all policies of the extensive form game available for each player, and $\Pi^* = \otimes_p \Pi^*_p$ be the set of all joint policies. JPSRO is an iteration-based algorithm, let $\{{}^c\pi^t_p, ...\} = \Pi^t_p$ be the set of new policies found at iteration t for player p with $c \in \mathcal{C}$ indexing an individual policy within that set. The set of all policies found so far for player p is denoted $\Pi^{0:t}_p$ and the set of joint policies is denoted $\Pi^{0:t} = \otimes_p \Pi^{0:t}_p$. The expected return (ER), an NF game $(G^{0:t}_p)_{p=1..n}$, is tracked for each joint policy found so far such that $G^{0:t}_p(\pi)$ is the expected return to player p when playing joint policy π. We also define G^*_p to be the payoff over all possible joint policies.

The MS is a function taking in the ER and returning a joint distribution, σ^t, over $\Pi^{0:t}$, such that $\sigma^t(\pi)$ is the probability to play joint policy $\pi \in \Pi^{0:t}$ at iteration t. The BR operator finds a policy which maximizes the expected return over of opponent mixed joint policies, $\pi_{-p} \in \Pi^{0:t}_{-p}$. This mixture is defined in terms of the MS joint distribution, σ^t.

5.1. Best Response Operators

At iteration $t+1$ each set, $\Pi^{0:t}_p$, can be expanded using either using a CCE or CE best response (BR) operator. The type of BR operator used determines the type of equilibrium that JPSRO converges to (Section 5.4).

JPSRO(CCE)
There is a single BR objective for each player, which expands the player policy set, $\Pi^{0:t+1}_p = \Pi^{0:t}_p \cup \Pi^{t+1}_p$, where $\Pi^{t+1}_p = \{\text{BR}^{t+1}_p\}$, and $\sigma(\pi_{-p}) = \sum_{\pi_p \in \Pi^{0:t}_p} \sigma(\pi_p, \pi_{-p})$.

$$\text{BR}^{t+1}_p \in \underset{\pi^*_p \in \Pi^*_p}{\text{argmax}} \sum_{\pi_{-p} \in \Pi^{0:t}_{-p}} \sigma^t(\pi_{-p}) G^*_p(\pi^*_p, \pi_{-p})$$

The CCE BR attempts to exploit the joint distribution with the responder's own policy preferences marginalized out.

JPSRO(CE)
There is a BR for each possible recommendation a player can get, $\Pi^{t+1}_p = \Pi^{0:t}_p \cup \Pi^{t+1}_p$, where $\Pi^{t+1}_p = \{(\text{BR}^{t+1}_p(\pi^i_p))_{i=1..|\Pi^{0:t}_p|}\}$.

$$\text{BR}^{t+1}_p(\pi_p) \in \underset{\pi^*_p \in \Pi^*_p}{\text{argmax}} \sum_{\pi_{-p} \in \Pi^{0:t}_{-p}} \sigma^t(\pi_{-p}|\pi_p) G^*_p(\pi^*_p, \pi_{-p})$$

Therefore 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 (Balduzzi et al., 2019). Computing the argmax of the BRs can be achieved through RL or exactly traversing the game tree.

5.2. Meta-Solvers

We propose that (C)CEs are good candidates as meta-solvers (MSs). They are more tractable than NEs and can enable coordination to maximize payoff between cooperative agents. In particular we propose three flavours of equilibrium MSs. 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) attempt 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 MSs 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 MSs with JPSRO(CCE).

5.3. Evaluation

Measuring convergence to NE (NE Gap, Lanctot et al. (2017)) is suitable in two-player, constant-sum games. However, it is not rich enough in cooperative settings. We propose to measure convergence to (C)CE ((C)CE Gap in Section E.4) in the full extensive form game. A gap, Δ, 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. Both gap and value metrics need to be evaluated under a meta-distribution. Using the same distribution as the MS may be unsuitable because MSs 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 polices found over time. The goal of an MS is to expand policy space (by proposing a joint policy to best respond to). If it fails to find novel policies at an acceptable rate, this could be evidence it is not performing well. Not all novel policies are useful, so caution should be exercised when interpreting this metric. If using a (C)CE MS and the gap is positive, it is guaranteed to find a novel BR policy.

Algorithm 1 Two-Player PSRO

```
$\Pi_1^0, \Pi_2^0 \leftarrow \{\pi_1^0\}, \{\pi_2^0\}$
$G^0 \leftarrow \text{ER}(\Pi^0)$
$\sigma_1^0, \sigma_2^0 \leftarrow \text{MS}(G^0)$
for $t \leftarrow \{1, ...\}$ do
    $\pi_1^t, \Delta_1^t \leftarrow \text{BR}(\Pi_2^{t-1}, \sigma_2^{t-1})$
    $\pi_2^t, \Delta_2^t \leftarrow \text{BR}(\Pi_1^{t-1}, \sigma_1^{t-1})$
    $\Pi_1^t, \Pi_2^t \leftarrow \Pi_1^{t-1} \cup \{\pi_1^t\}, \Pi_2^{t-1} \cup \{\pi_2^t\}$
    $G^t \leftarrow \text{ER}(\Pi^t)$
    $\sigma_1^t, \sigma_2^t \leftarrow \text{MS}(G^t)$
    if $\Delta_1^t + \Delta_2^t = 0$ then
        break
return $(\Pi_1^{0:t}, \Pi_2^{0:t}), (\sigma_1^t, \sigma_2^t)$
```

Algorithm 2 JPSRO

```
$\Pi_1^0, ..., \Pi_n^0 \leftarrow \{\pi_1^0\}, ..., \{\pi_n^0\}$
$G^0 \leftarrow \text{ER}(\Pi^0)$
$\sigma^0 \leftarrow \text{MS}(G^0)$
for $t \leftarrow \{1, ...\}$ do
    for $p \leftarrow \{1, ..., n\}$ do
        $\{{}^1\pi_p^t, ...\}, \{{}^1\Delta_p^t, ...\} \leftarrow \text{BR}_p(\Pi^{0:t-1}, \sigma^{t-1})$
        $\Pi_p^{0:t} \leftarrow \Pi_p^{0:t-1} \cup \{{}^1\pi_p^t, ...\}$
    $G^{0:t} \leftarrow \text{ER}(\Pi^{0:t})$
    $\sigma^t \leftarrow \text{MS}(G^{0:t})$
    if $\sum_{p,c} {}^c\Delta_p^t = 0$ then
        break
return $\Pi^{0:t}, \sigma^t$
```

5.4. Convergence to Equilibria

JPSRO(CCE) converges[8] to a CCE and JPSRO(CE) converges to a CE. We provide a sketch of the proofs here, for full proofs see Section E.5.

Theorem 6 (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.*

Theorem 7 (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.*

Proof. A (C)CE MS provides a distribution that is in equilibrium over the set of joint policies found so far, $\Pi^{0:t}$. For the algorithm to have converged, it needs to also be in equilibrium over the set of all possible joint policies, Π^*. This is the case when the BR fails to find a novel policy with nonzero gap. Policies that have been found before, by definition of (C)CE, have zero gap. All behavioural policies can be defined in terms of a mixture of deterministic policies. Therefore, given that there are finite deterministic policies the algorithm will converge. □

6. CEs and CCEs as Joint Meta-Solvers

We evaluate a number of (C)CE MSs in JPSRO on pure competition, pure cooperation, and general-sum games (Section H). All games used are available in OpenSpiel (Lanctot et al., 2019). More thorough descriptions of the games used can be found in Section F. We use an exact BR oracle, and exactly evaluate policies in the meta-game by traversing the game tree to precisely isolate the MS's contribution to the algorithm.

We compare against common MS including uniform, α-Rank (Omidshafiei et al., 2019; Muller et al., 2020), Projected Replicator Dynamics (PRD) (Lanctot et al., 2017) 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 MSs 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 MS distribution and MW(C)CE to provide a consistent and value maximizing comparison. Experiments were ran for up to 6 hours, after which they were terminated.

Kuhn Poker (Kuhn, 1950; Southey et al., 2009; Lanctot, 2014) 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 2a show rapid convergence to equilibrium.

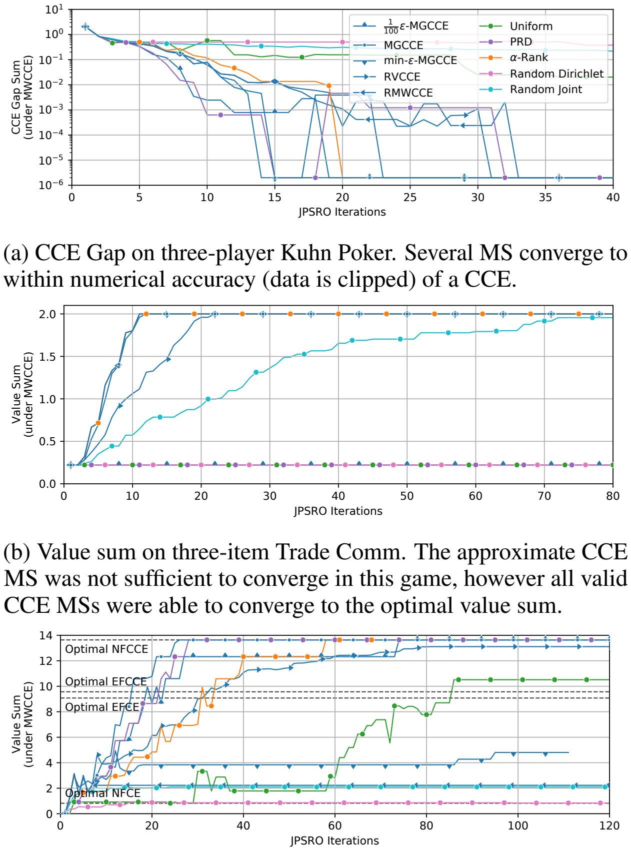
Trade Comm is a two-player, common-payoff trading game, where players attempt to coordinate on a compatible trade. This game is difficult because it requires searching over a large number of policies to find a compatible mapping, and can easily fall into a sub-optimal equilibrium. Figure 2b shows a remarkable dominance of CCE MSs. It is clear that traditional PSRO MSs cannot cope with this cooperative setting.

Sheriff (Farina et al., 2019b) 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 2c shows that JPSRO is capable of finding the optimal value.

7. Discussion

There has been significant recent interest in solving the equilibrium selection problem (Ortiz et al., 2007; Omid-

[8] In exponential time in the worst case, however in practice convergence is much faster.



(a) CCE Gap on three-player Kuhn Poker. Several MS converge to within numerical accuracy (data is clipped) of a CCE.

(b) Value sum on three-item Trade Comm. The approximate CCE MS was not sufficient to converge in this game, however all valid CCE MSs 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 2. JPSRO(CCE) on various games. Additional metrics can be found in Section H. MGCCE is consistently a good choice of MS over the games tested.

shafiei et al., 2019). This paper provides a novel approach which is computationally tractable, supports general-support solutions, and has favourable scaling properties when the solution is full-support.

The new solution concept MG(C)CE is rooted in the powerful principles of entropy and margin maximisation. Therefore it is a simple solution that makes limited assumptions, and is robust to many possible counter strategies (Jaynes, 1957). The MG(C)CE defines a family of unique solutions parameterized by ϵ, that can control for the properties of the distribution. We have compared it to other NE, CE, and α-Rank solutions, and have shown it has several advantages over these approaches, and performs very well across a variety of games.

PSRO has proved to be a formidable learning algorithm in two-player, constant-sum games, and JPSRO, with (C)CE MSs, 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 MSs, 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 MS 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 MSs (such as MG(C)CE) have better empirical value and convergence than the uniform MS. 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" (Vinyals et al., 2019)). Furthermore, one could also switch between MSs at each iteration to achieve the best mix of exploitation and exploration.

Another strength of (C)CE MSs is that they appear to perform well across many different games, with different numbers of players and payoff properties.

8. 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 MSs 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 MSs. Finally, we believe that both MG(C)CE and JPSRO can scale to large problems, by using stochastic online MSs for the former and exploiting function approximation and RL for the latter.

9. Acknowledgements

Special thanks to Shayegan Omidshafiei for help with α-Rank related discussions, Thomas Anthony for helpful comments and critiques of a draft of the paper, Gabriele Farina for providing maximum welfare values for the Sheriff game, and the anonymous ICML reviewers whose thoughtful feedback strengthened the paper considerably.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems.

Journal of Control and Decision, 5(1):42–60, 2018.

Anthony, T., Eccles, T., Tacchetti, A., Kramár, J., Gemp, I., Hudson, T. C., Porcel, N., Lanctot, M., Pérolat, J., Everett, R., Werpachowski, R., Singh, S., Graepel, T., and Bachrach, Y. Learning to play no-press diplomacy with best response policy iteration, 2020.

Aumann, R. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1): 67–96, 1974.

Avis, D., Rosenberg, G. D., Savani, R., and Von Stengel, B. Enumeration of Nash equilibria for two-player games. *Economic theory*, 42(1):9–37, 2010.

Balduzzi, D., Tuyls, K., Perolat, J., and Graepel, T. Re-evaluating evaluation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS, pp. 3272–3283, Red Hook, NY, USA, 2018. Curran Associates Inc.

Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W. M., Pérolat, J., Jaderberg, M., and Graepel, T. Open-ended learning in symmetric zero-sum games. *CoRR*, abs/1901.08106, 2019. URL http://arxiv.org/abs/1901.08106.

Ben-Tal, A., Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009. ISBN 9781400831050.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Brown, G. W. Iterative solutions of games by fictitious play. 1951.

Brown, N. and Sandholm, T. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019. ISSN 0036-8075. doi: 10.1126/science.aay2400.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16:1190–1208, September 1995. ISSN 1064-8275.

Celli, A., Marchesi, A., Bianchi, T., and Gatti, N. Learning to correlate in multi-player general-sum sequential games, 2019.

Celli, A., Marchesi, A., Farina, G., and Gatti, N. No-regret learning dynamics for extensive-form correlated equilibrium, 2020.

Cortes, C. and Vapnik, V. Support-vector networks. In *Machine Learning*, pp. 273–297, 1995.

Daskalakis, C., Goldberg, P., and Papadimitriou, C. The complexity of computing a Nash equilibrium. *SIAM J. Comput.*, 39:195–259, 02 2009.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dudik, M. and Gordon, G. A sampling-based approach to computing equilibria in succinct extensive-form games. 05 2012.

Farina, G., Bianchi, T., and Sandholm, T. Coarse correlation in extensive-form games, 2019a.

Farina, G., Ling, C. K., Fang, F., and Sandholm, T. Correlation in extensive-form games: Saddle-point formulation and benchmarks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019b.

Fudenberg, D. and Tirole, J. *Game Theory*. MIT Press, 1991.

Gerschgorin, S. *Uber die Abgrenzung der Eigenwerte einer Matrix*. 1931.

Goldberg, P. W., Papadimitriou, C. H., and Savani, R. The complexity of the homotopy method, equilibrium selection, and lemke-howson solutions. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–25, 2013.

Gray, J., Lerer, A., Bakhtin, A., and Brown, N. Human-level performance in no-press diplomacy via equilibrium search, 2020.

Harsanyi, J. and Selten, R. *A General Theory of Equilibrium Selection in Games*, volume 1. The MIT Press, 1 edition, 1988.

Havrda, J., Charvat, F., and Havrda, J. Quantification method of classification processes: Concept of structural a-entropy. *Kybernetika*, 1967.

Heinrich, J., Lanctot, M., and Silver, D. Fictitious self-play in extensive-form games. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015.

Hoehn, B., Southey, F., Holte, R. C., and Bulitko, V. Effective short-term opponent exploitation in simplified poker.

Jaderberg, M., Czarnecki, W., Dunning, I., Marris, L., Lever, G., Castañeda, A., Beattie, C., Rabinowitz, N., Morcos, A., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J., Silver, D., Hassabis, D., Kavukcuoglu, K., and Graepel, T. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364:859–865, 05 2019.

Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

Kaur, M. and Buttar, G. A brief review on different measures of entropy. 08 2019.

Kuhn, H. W. A simplified two-person poker. 1:97–103, 1950.

Lanctot, M. Further developments of extensive-form replicator dynamics using the sequence-form representation. volume 2, 05 2014.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. In *NIPS*. 2017.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B. D., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, 2019.

Leonard J. Savage, J. W. The foundations of statistics. 1954.

Lockhart, E., Burch, N., Bard, N., Borgeaud, S., Eccles, T., Smaira, L., and Smith, R. Human-agent cooperation in bridge bidding, 2020.

Mas-Colell, A., Whinston, M. D., and Green, J. R. *Microeconomic Theory*. Oxford University Press, 1995.

Matouek, J. and Gärtner, B. *Understanding and Using Linear Programming (Universitext)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3540306978.

McAleer, S., Lanier, J., Fox, R., and Baldi, P. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. In *Neural Information Processing Systems 33*, 2020.

McAleer, S., Lanier, J. B., Baldi, P., and Fox, R. XDO: A double oracle algorithm for extensive-form games. *CoRR*, abs/2103.06426, 2021. URL https://arxiv.org/abs/2103.06426.

McMahan, H. B., Gordon, G. J., and Blum, A. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 536–543, 2003.

Morrill, D., D'Orazio, R., Sarfati, R., Lanctot, M., Wright, J. R., Greenwald, A., and Bowling, M. Hindsight and sequential rationality of correlated play. In *Proceedings of the The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

Moulin, H. and Vial, J.-P. 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.

Muller, P., Omidshafiei, S., Rowland, M., Tuyls, K., Perolat, J., Liu, S., Hennes, D., Marris, L., Lanctot, M., Hughes, E., Wang, Z., Lever, G., Heess, N., Graepel, T., and Munos, R. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2020.

Nash, J. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Nau, R., Canovas, S. G., and Hansen, P. On the geometry of Nash equilibria and correlated equilibria. *International Journal of Game Theory*, 32(4):443–453, August 2004.

O'Leary, D. P. A generalized conjugate gradient algorithm for solving a class of quadratic programming problems. *Linear Algebra and its Applications*, 34:371–399, 1980/12// 1980.

Omidshafiei, S., Papadimitriou, C., Piliouras, G., Tuyls, K., Rowland, M., Lespiau, J.-B., Czarnecki, W. M., Lanctot, M., Perolat, J., and Munos, R. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9(1):9937, 2019.

OpenAI, :, Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., de Oliveira Pinto, H. P., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., and Zhang, S. Dota 2 with large scale deep reinforcement learning, 2019.

Ortiz, L. E., Schapire, R. E., and Kakade, S. M. Maximum entropy correlated equilibria. In Meila, M. and Shen, X. (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 347–354, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

Polyak, B. The conjugate gradient method in extreme problem. *USSR Computational Mathematics and Mathematical Physics*, 9:94–112, 12 1969.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sinn, H.-W. A Rehabilitation of the Principle of Insufficient Reason. *The Quarterly Journal of Economics*, 94(3): 493–506, 05 1980.

Sokota, S., Lockhart, E., Timbers, F., Davoodi, E., D'Orazio, R., Burch, N., Schmid, M., Bowling, M., and Lanctot, M. Solving common-payoff games with approximate policy iteration, 2021.

Southey, F., Hoehn, B., and Holte, R. Effective short-term opponent exploitation in simplified poker. *Machine Learning*, 74:159–189, 02 2009.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 2020.

Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Vinyals, O., Babuschkin, I., Czarnecki, W., Mathieu, M., Dudzik, A., Chung, J., Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J., Jaderberg, M., and Silver, D. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575, 11 2019.

von Stengel, B. and Forges, F. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008.

Wald, A. Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.*, 10(4):299–326, 12 1939.

Wald, A. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46:265–280, 1945.

Wang, Y. and Xia, S. Unifying attribute splitting criteria of decision trees by Tsallis entropy. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2507–2511, 2017.